

Development and Validation of a Risk Quantification Index for 30-Day Postoperative Mortality and Morbidity in Noncardiac Surgical Patients

Jarrod E. Dalton, M.A.,* Andrea Kurz, M.D.,† Alparslan Turan, M.D.,‡ Edward J. Mascha, Ph.D.,§ Daniel I. Sessler, M.D.,|| Leif Saager, M.D.#

ABSTRACT

Background: Optimal risk adjustment is a requisite precondition for monitoring quality of care and interpreting public reports of hospital outcomes. Current risk-adjustment measures have been criticized for including baseline variables that are difficult to obtain and inadequately adjusting for high-risk patients. The authors sought to develop highly predictive risk-adjustment models for 30-day mortality and morbidity based only on a small number of preoperative baseline characteristics. They included the Current Procedural Terminology code corresponding to the patient's primary procedure (American Medical Association), American Society of Anesthesiologists Physical Status, and age (for mortality) or hospitalization (inpatient *vs.* outpatient, for morbidity).

Methods: Data from 635,265 noncardiac surgical patients participating in the American College of Surgeons National Surgical Quality Improvement Program between 2005 and 2008 were analyzed. The authors developed a novel algorithm to aggregate sparsely represented Current Procedural Terminology codes into logical groups and estimated univariable Procedural Severity Scores—one for mortality and

What We Already Know about This Topic

- Current perioperative risk-adjustment measures have been criticized for including baseline variables that are difficult to obtain and inadequately adjusting for high-risk patients

What This Article Tells Us That Is New

- Novel, highly predictive risk quantification models for 30-day mortality and composite major morbidity were successfully developed that use only three readily available patient and procedural characteristics

morbidity, respectively—for each aggregated group. These scores were then used as predictors in developing respective risk quantification models. Models were validated with *c*-statistics, and calibration was assessed using observed-to-expected ratios of event frequencies for clinically relevant strata of risk.

Results: The risk quantification models demonstrated excellent predictive accuracy for 30-day postoperative mortality (*c*-statistic [95% CI] 0.915 [0.906–0.924]) and morbidity (0.867 [0.858–0.876]). Even in high-risk patients, observed rates calibrated well with estimated probabilities for mortality (observed-to-expected ratio: 0.93 [0.81–1.06]) and morbidity (0.99 [0.93–1.05]).

Conclusion: The authors developed simple risk-adjustment models, each based on three easily obtained variables, that allow for objective quality-of-care monitoring among hospitals.

HETEROGENEITY among providers (*i.e.*, clinicians or hospitals) in the quality of health care delivered remains an issue of mounting concern in the United States.¹ A key issue in improving overall quality of the healthcare system is identifying this heterogeneity by studying outcomes-based performance measures. But valid outcome-based quality monitoring efforts must properly account for inherent variability in risk associated with differences in base-

* Senior Biostatistician, § Staff Biostatistician, Departments of Quantitative Health Sciences and Outcomes Research, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio. † Professor and Vice-Chair, ‡ Associate Professor, # Assistant Professor, Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic. || Professor and Chair, Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, and Visiting Scientist, Department of Anesthesiology, McMaster University, Hamilton, Ontario, Canada.

Received from the Departments of Outcomes Research and Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio. Submitted for publication February 4, 2010. Accepted for publication February 7, 2011. Support was provided solely from institutional and/or departmental sources. The American College of Surgeons National Surgical Quality Improvement Program and its participating hospitals are the source of the data used here; the College has not verified and is not responsible for the statistical validity of the data analysis or the conclusions derived by the authors.

Address correspondence to Dr. Saager: Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, 9500 Euclid Avenue, P-77, Cleveland, Ohio 44195. saagerl@ccf.org. Information on purchasing reprints may be found at www.anesthesiology.org or on the masthead page at the beginning of this issue. ANESTHESIOLOGY's articles are made freely accessible to all readers, for personal use only, 6 months from the cover date of the issue.

Copyright © 2011, the American Society of Anesthesiologists, Inc. Lippincott Williams & Wilkins. Anesthesiology 2011; 114:1336–44

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are available in both the HTML and PDF versions of this article. Links to the digital files are provided in the HTML text of this article on the Journal's Web site (www.anesthesiology.org).

line characteristics among patients as well as differences in procedural complexity.^{2,3}

The American Society of Anesthesiologists (ASA) Physical Status Classification System has become a routine method of preoperative risk assessment. Although the ASA Physical Status is a baseline risk assessment and thus operative risk is not included in the score, it has been associated with perioperative morbidity and mortality.^{4–7} Combined with other baseline factors, it may contribute to an accurate characterization of patient risk.

Currently available outcomes-based measures may not adequately adjust for risk, especially in sicker patients.⁸ To the extent that available indices inadequately account for risk in sicker patients, they encourage clinicians and health systems that must publically report outcomes to “cherry pick” relatively healthy patients.⁹

Other scores have been developed for specific risks or for specific patient populations, such as the Goldman Cardiac Risk Index for cardiac risk in noncardiac surgery patients and the European System for Cardiac Operative Risk Evaluation for estimating operative mortality in patients undergoing cardiac surgery.^{10,11} These scores, although well-established and implemented, were not developed for the purpose of risk-adjusted quality-of-care comparisons among providers with respect to noncardiac surgical patients. More importantly, they require information that usually is not available in administrative databases that most broadly represent the U.S. surgical population.

Recently, Sessler *et al.* introduced a novel Risk Stratification Index based on Medicare Provider Analysis and Review files.¹² This index improves greatly on previous models. The Medicare Provider Analysis and Review database contains information from claims for services and relies completely on administrative data, namely International Classification of Disease codes. In contrast, the procedure coding of the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) and several other administrative databases uses the Healthcare Common Procedure Coding System, of which the American Medical Association Current Procedural Terminology (CPT) codes are an integral part.¹³

The ACS-NSQIP, a national, prospectively implemented program for comparative assessment and enhancement of surgical outcomes among multiple institutions for several surgical subspecialties, incorporates risk-adjustment indices for 30-day mortality and 30-day morbidity.^{14,15} Although these models are highly predictive, they require collection of detailed information on a multitudinous array of patient risk factors (including laboratory values). Generalizable and practical risk-adjustment models would best rely on a limited number of risk factors that are routinely available for most patients. Thus, our goal was to develop practical, highly predictive risk-adjustment models for 30-day mortality and morbidity in the U.S. noncardiac surgical population based only on baseline patient characteristics and CPT codes.

Materials and Methods

Data Collection

The ACS-NSQIP is a prospective, outcomes-based registry for comparative assessment and enhancement of the quality of surgical care among multiple institutions for several surgical subspecialties.^{16,17} More than 250 academic and large community surgical institutions across the United States participate in the ACS-NSQIP. Participating institutions employ full-time surgical clinical reviewers to ensure the integrity of perioperative and 30-day outcomes data. Furthermore, routine audits are performed to monitor the accuracy of the data collection process. Data pertaining to patients undergoing certain low-morbidity, high-volume procedures are limited to avoid overwhelming the registry. Patients included in the ACS-NSQIP registry are enrolled in site-specific, 8-day cycles to avoid systematic bias introduced by weekly patterns in case loads.¹⁸ We used perioperative data on 635,265 noncardiac surgery patients treated between 2005 and 2008 at participating ACS-NSQIP surgical centers.

We withheld a randomly selected validation cohort of 50,000 patients to study the predictive accuracy and calibration of our risk models. The remaining patients' data were used for developing our risk models as described below. Although our validation cohort was not external to the set of ACS-NSQIP-participating hospitals, we thought the cohort was adequate to study whether we overfit our model to the training cohort.

The primary outcomes for our study were 30-day postoperative mortality and 30-day postoperative major morbidity, which was defined as an occurrence of one or more of the following postoperative complications: organ space infection, pneumonia, unplanned intubation, pulmonary embolism, ventilator dependence more than 48 h, acute renal failure, stroke/cerebral vascular accident with neurologic deficit, coma more than 24 h, cardiac arrest requiring cardiopulmonary resuscitation, myocardial infarction, transfusion of 5 or more units erythrocytes within 72 h, sepsis/septic shock, and mortality.

Statistical Analysis

SAS statistical software version 9.2 (SAS Institute, Cary, NC) was used for data management, whereas statistical modeling was performed using R software version 2.8.1 (The R Foundation for Statistical Computing, Vienna, Austria) and, specifically, its Design and Hmisc packages.¹⁹

Our overall modeling approach was implemented in two phases. First, for each of our two outcomes (mortality and morbidity) we developed a univariable score measuring procedure-associated risks (hereafter referred to as Procedural Severity Scores or PSS). Second, we developed multivariable logistic regression models for mortality and morbidity, each using the respective PSS and routinely available patient characteristics as predictor variables.

Univariable Risk Scores Associated with the Primary Procedure

Development of the univariable procedure-associated risk scores was as follows: The ACS-NSQIP database includes procedural data, encoded using the CPT codes.²⁰ Our main predictor variable was based on the primary CPT code associated with each patient's visit. Frequencies of individual (primary) CPT codes in the training data set varied widely, ranging from 1 patient (369 CPT codes) to 39,373 patients (laparoscopic cholecystectomy); among the 2,555 CPT codes present in the training data set, 1,721 codes were represented by 30 or fewer patients.

Aggregation of Procedures

The Clinical Classifications Software for Services and Procedures (CCS, U.S. Department of Health and Human Services Agency for Healthcare Research and Quality) aggregates each individual CPT code into 1 of 244 mutually exclusive, clinically appropriate categories. Based on these CCS procedure groups, we postulated that the aggregation of many sparsely represented CPT codes would provide increased precision for predicting the primary outcomes of our study. Furthermore, we considered the possibility that even some CCS procedure groups might have similarly low frequencies of cases. On the other hand, certain CPT codes (such as laparoscopic cholecystectomy) were sufficiently common not to require aggregation.

We therefore aggregated procedures represented by fewer than a certain number of patients—which we denote by N^* —into the associated CCS procedure group. Furthermore, we allowed for the possibility that even after this aggregation, certain CCS procedure groups would be sparsely represented. These sparsely represented CCS procedure groups were further aggregated into an all-purpose “other” group using the same rule (aggregate if the number of patients is less than N^*). For 30-day mortality and composite 30-day morbidity separately, we selected N^* objectively such that the aggregation routine yielded maximal predictive accuracy among the training cohort.

Determination of Minimum CCS Procedure Group Size for Aggregation

We implemented a cross-validation study (for each outcome) as follows: For candidate values of N^* ranging between 10 and 5,000, we randomly partitioned the training cohort into 10 subgroups of equal size. Then, for each subgroup, we used the other 90% of the data to (1) aggregate procedures using the methodology outlined above (with the candidate N^* in question); (2) estimate the incidence of the outcome within each (aggregated) group; and (3) evaluate the predictive accuracy of these incidences for the reserved 10% of the data using the c-statistic (the c-statistic is a quantity ranging between 0.5 and 1.0, for which 0.5 represents no discriminative ability beyond random guessing, whereas 1.0 represents perfect prediction¹⁹). This process was repeated over the 10

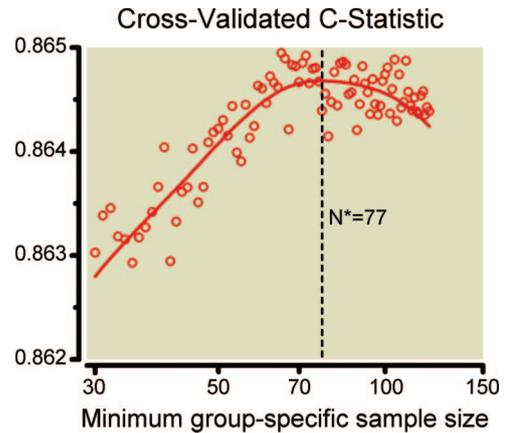


Fig. 1. Choice of N^* for mortality: Cross-validation was used to estimate the expected c-statistic for new data over minimum group sizes ranging from 10 to 5,000 (only sample sizes from 30 to 150 are shown). $N^* = 77$ was chosen because it maximized the smooth regression relationship between cross-validated c-statistic and log-sample size.

randomly partitioned subgroups, and the c-statistics were averaged to obtain an estimate of the expected c-statistic associated with the candidate N^* in question. Smoothing spline regression was then used to find N^* (*i.e.*, the value that maximized the cross-validated c-statistic), as shown in figure 1 for 30-day mortality.

Postaggregation Assignment of PSS

Once N^* was selected using the cross-validation routine described in the previous section (again, specific to each outcome), the incidence of the outcome for each aggregated group was estimated using the entire training cohort. Then the PSS (our measure of univariable procedure risk) was defined for 30-day mortality as the twelfth root of these estimated incidences (which we refer to as PSS-mortality). Because these incidence estimates were heavily skewed to the right (*i.e.*, most procedures are associated with low risk of the outcome, but a few procedures are associated with a larger risk), this twelfth-root transformation was warranted to model a linear relationship between PSS mortality and log-odds of mortality. With methodology akin to that used to obtain our PSS mortality score, we defined PSS morbidity using the cube-root transformation of the aggregated group-specific incidence observed among the training cohort.

Multivariable Modeling with PSS and Other Patient Characteristics

With the PSS now defined, we proceeded to build our overall risk models combining procedural and patient characteristics as follows: First, multivariable logistic regression was used with PSS and all baseline risk factors showing potential predictive ability as determined by relatively large values of test statistics for univariable association with the outcome (*e.g.*, univariable chi-square test statistic greater than 1,000 for

categorical predictors). In this “full” model, the relationship between PSS and the log-odds of the outcome was modeled linearly, whereas restricted cubic splines with four equally spaced knots were used for all other continuous risk factors to allow for potential nonlinearities in the estimated relationships. C-statistics for this model (and their associated 95% CI, estimated using normal approximation theory for proportions), applied to both the training and validation data sets, were then estimated. As mentioned, the utility of such a full model suffers from difficulties associated with collecting data on many baseline risk factors. Thus, we considered a “reduced” model, which incorporated a limited number of selected risk factors with the goal of minimizing the reduction in the model c-statistics.

Validation of Fitted Models

Our definition of the composite morbidity outcome differs slightly from that which was used to define the morbidity index currently available in the ACS-NSQIP database (for instance, we include mortality as one of the individual outcomes comprising composite morbidity, whereas the ACS-NSQIP morbidity index does not). Furthermore, the morbidity and mortality indices in the ACS-NSQIP database are estimated for only general and vascular surgery patients, and these indices supplement baseline and procedural risk factors with preoperative laboratory results. Nonetheless, we compared c-statistics arising from our models to those from the ACS-NSQIP indices for general and vascular surgery patients.

Calibration (the agreement between the predicted probabilities from a model and the observed outcome proportions) was assessed for our final models for mortality and morbidity within our validation cohort. First, we divided the validation cohort patients into groups based on their predicted probabilities of the outcome; then, we estimated group-specific ratios of observed-to-expected outcome counts (O/E ratios), using a Poisson log-linear model. Groups were selected to reflect clinically meaningful strata of risk. The expected outcome count within each group was the sum of the predicted probabilities of its members. A global Wald test of all O/E ratios equal to 1 was performed with a significance criterion of 0.05, and Bonferroni-adjusted 95% simultaneous CIs were estimated for each O/E ratio.

Results

Aggregation of Procedures

The optimal minimum group-specific sample size for the aggregation of procedures into groups was $N^* = 77$ for mortality (fig. 1) and $N^* = 23$ for morbidity. The resulting estimates of PSS mortality and PSS morbidity, as well as details about the aggregation of individual CPT codes into CCS procedure groups and of CCS procedure groups into an all-purpose other category, are given in Supplemental Digital

Content 1, <http://links.lww.com/ALN/A738>, for each individual CPT code.

Discrimination

In general, 30-day mortality was more accurately predicted than composite 30-day morbidity, as evidenced by the higher c-statistics among the randomly determined subset of 50,000 validation cases (table 1).

For mortality, PSS predicted the outcome better than did individual CPT codes (c-statistic [95% CI] 0.867 [0.859–0.876] vs. 0.847 [0.838–0.856]). Our reduced model for mortality included the following risk factors: PSS-mortality, ASA Physical Status, and age. This model resulted in a validation data c-statistic of 0.915 (0.906–0.924), whereas the ACS-NSQIP mortality score of Khuri *et al.*, as applied to our validation cohort, gave a c-statistic of 0.941 (0.931–0.950).¹⁴

Predictive ability for PSS-morbidity was essentially similar to that observed for CPT codes (table 1). For this outcome, our reduced model incorporated PSS-morbidity, ASA Physical Status, and hospitalization (*i.e.*, inpatient vs. outpatient); this model produced a validation data c-statistic of 0.867 (0.858–0.876). No appreciable increase in predictive accuracy was observed when using the comparator ACS-NSQIP morbidity score of Daley *et al.* ($C = 0.875$ [0.866–0.884]), whose analysis incorporated more risk factors and was based on a subset of the noncardiac surgical population (whereas our models were developed and validated using all patients in the registry).¹⁵

Before incorporation of patient risk factors, the procedural information alone predicted both outcomes quite accurately. The PSS developed for 30-day mortality resulted in a univariable c-statistic (95% CI 0.867 [0.859–0.876]; table 1), whereas the PSS developed for the composite morbidity outcome was 0.839 (0.830–0.848); these c-statistics were comparable or slightly higher (more discriminative) than the c-statistics achieved by using the individual CPT codes.

Combining patient demographic and morphometric information with the PSS increased the c-statistic for both outcomes. In the full model for 30-day mortality (which incorporated 24 total predictors), the c-statistic was 0.936 (0.927–0.945), which was comparable with the predictive ability achieved by the ACS-NSQIP mortality score of Khuri *et al.* (as applied to our validation cohort).¹⁴ Our reduced model for 30-day mortality (which included only PSS-mortality, ASA Physical Status, and age) predicted the outcome only marginally less accurately than did the models that incorporated many predictors. Similar results were obtained for the composite 30-day postoperative morbidity/mortality outcome; although overall this outcome was predicted less accurately than 30-day mortality, our reduced model performed as accurately as the full

Table 1. C-statistics and 95% Confidence Intervals for Selected Models Applied to a Validation Cohort of 50,000 Randomly Selected Cases

Outcome	Model	Validation Cohort C-statistic (95% CI)
30-day mortality	Univariable model (CPT code)	0.847 (0.838–0.856)
	Univariable model (PSS mortality)	0.867 (0.859–0.876)
	Full model*	0.936 (0.927–0.945)
	Reduced model for 30-day mortality†	0.915 (0.906–0.924)
	NSQIP probability of mortality‡	0.941 (0.931–0.950)
30-day composite major morbidity	Univariable model (CPT code)	0.836 (0.827–0.845)
	Univariable model (PSS morbidity)	0.839 (0.830–0.848)
	Full model§	0.876 (0.867–0.885)
	Reduced model for 30-day morbidity	0.867 (0.858–0.876)
	NSQIP probability of morbidity‡	0.875 (0.866–0.884)

* Includes the following preoperative variables as predictors: PSS mortality, age, body mass index (truncated at 75), American Society of Anesthesiologists (ASA) Physical Status, emergent status, functional health status, preoperative sepsis/systemic inflammatory response syndrome, ventilator dependence, impaired sensorium, dyspnea, transfer status (admitted directly from home vs. admitted from other facility), ascites, current pneumonia, renal failure, bleeding disorders, open wound/wound infection, currently on dialysis, hospitalization status (inpatient vs. outpatient), history of chronic obstructive pulmonary disease, disseminated cancer, weight loss (more than 10% of body weight in prior 6 months), steroid use for chronic condition, rest pain/gangrene, and diabetes mellitus. † Includes PSS mortality, ASA Physical Status, and age as predictors. ‡ These models incorporate all baseline risk factors (including laboratory measurements) and procedural data. § Includes the following preoperative variables as predictors: PSS morbidity, age, body mass index (truncated at 75), ASA Physical Status, emergent status, functional health status, preoperative sepsis/systemic inflammatory response syndrome, ventilator dependence, impaired sensorium, dyspnea, transfer status (admitted directly from home vs. admitted from other facility), ascites, current pneumonia, renal failure, bleeding disorders, open wound/wound infection, hospitalization status (inpatient vs. outpatient), history of chronic obstructive pulmonary disease, and transfusion (more than 4 units erythrocytes within 72 h of start of surgery). || Includes PSS morbidity, ASA Physical Status, and hospitalization as predictors. CPT = Current Procedural Terminology; NSQIP = National Surgical Quality Improvement Program; PSS = procedural severity score.

model and as accurately as the ACS-NSQIP morbidity score of Daley *et al.*¹⁵

Calibration

Calibration of the two final models is summarized in table 2; the *P* value of the global Wald calibration test was 0.64 for 30-day mortality and 0.40 for 30-day morbidity,

meaning that we did not find sufficient evidence indicating lack of agreement between model predictions and the observed outcomes. Specifically, the ratio of observed/expected number of deaths among patients with more than 10% predicted probability based on our reduced model for mortality (*i.e.*, high-risk patients) was 0.93 (0.81–1.06). Likewise, the ratio of observed/expected number of

Table 2. Calibration of Reduced Logistic Regression Models Applied to a Validation Cohort of 50,000 Randomly Selected Cases

Outcome	Predicted Probability of Outcome	Test Data Set N (%) of Patients	Number of Events		O/E Ratio (95% CI)†
			Observed	Expected*	
30-day mortality	≤0.5%	32,832 (66.0%)	41 (5%)	37 (5%)	1.12 (0.74–1.69)
	0.6–1.0%	4,352 (8.7%)	32 (4%)	31 (4%)	1.02 (0.64–1.63)
	1.1–2.0%	4,195 (8.4%)	58 (7%)	61 (7%)	0.96 (0.68–1.36)
	2.1–5.0%	4,486 (9.0%)	150 (19%)	142 (18%)	1.05 (0.85–1.31)
	5.0–10.0%	1,996 (4.0%)	149 (19%)	137 (17%)	1.09 (0.88–1.35)
	>10.0%	1,891 (3.8%)	370 (46%)	399 (49%)	0.93 (0.81–1.06)
30-day major morbidity	1.0% or lower	15,653 (31.4%)	50 (1%)	67 (2%)	0.75 (0.51–1.08)
	1.1–2.0%	6,376 (12.8%)	90 (2%)	95 (3%)	0.94 (0.72–1.25)
	2.1–5.0%	9,927 (19.9%)	333 (9%)	322 (9%)	1.03 (0.89–1.20)
	5.0–10.0%	6,827 (13.7%)	500 (13%)	493 (13%)	1.01 (0.90–1.14)
	10.1–20.0%	5,745 (11.5%)	873 (23%)	847 (23%)	1.03 (0.94–1.13)
	>20.0%	5,245 (10.5%)	1,874 (50%)	1,898 (51%)	0.99 (0.93–1.05)

Calibration is the relationship between observed and expected outcome counts as a function of the predicted probability of outcome.

* The expected number of events is the sum of the predicted probability of the event among all patients within a stratum. † Estimated using a Poisson log-linear model (incorporating the expected number of events as an offset variable).

O/E = observed-to-expected.

30-Day Mortality

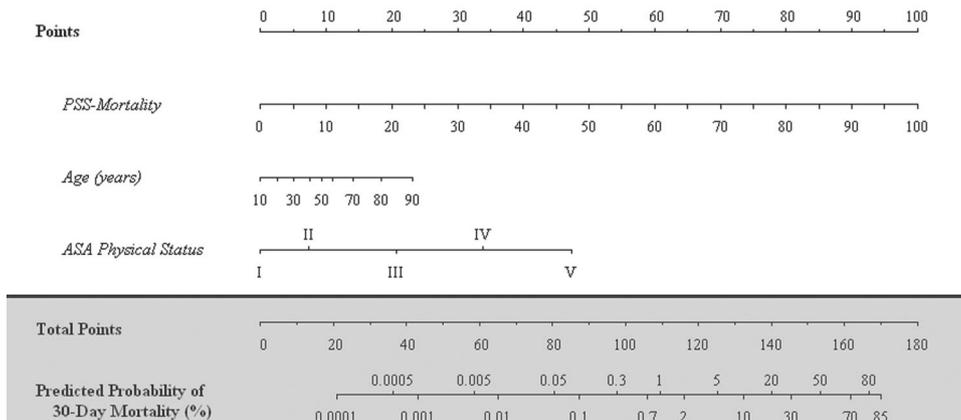


Fig. 2. Nomogram for predicting 30-day mortality based on Current Procedural Terminology (CPT) code, American Society of Anesthesiologists (ASA) Physical Status, and age. *Instructions:* Find the Procedural Severity Score for mortality (PSS mortality) in Supplemental Digital Content 1 for the patient’s primary CPT code. Locate this score on the axis labeled “PSS-Mortality” by interpolating between the displayed values, and draw a vertical line up to the “Points” axis to find the points toward 30-day mortality associated with the patient’s procedure. Repeat this process for age and ASA Physical Status and add the three resulting point values to obtain the patient’s total points. Locate this number on the “Total Points” axis (*shaded region*) and then draw a vertical line to the axis labeled “Predicted Probability of 30-Day Mortality (%)” This value represents the prediction for the patient.

patients experiencing composite major morbidity among patients with more than 20% predicted probability based on our reduced model for morbidity was 0.99 (0.93–1.05).

Obtaining Risk Estimates

Nomograms (graphic tools for obtaining regression model estimates) are provided for our reduced models (along with instructions) in figures 2 and 3. Within these nomo-

grams, the axis length corresponding to each risk factor represents the relative degree of contribution of that risk factor in the model; thus, PSS mortality and PSS morbidity were the strongest predictors among those chosen for their respective models. ASA Physical Status, the next strongest predictor in each of the reduced models, displayed a relatively larger influence on the prediction for morbidity than it did for mortality.

30-Day Major Morbidity/Mortality

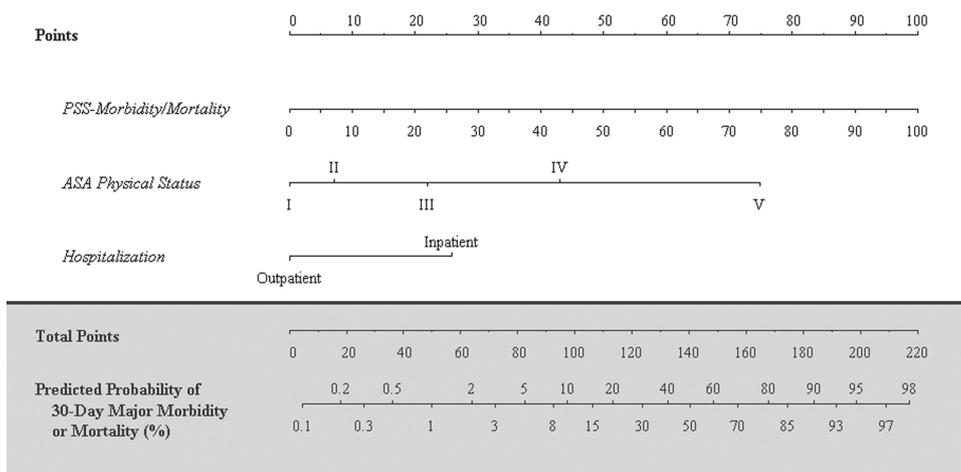


Fig. 3. Nomogram for predicting 30-day major morbidity based on Current Procedural Terminology (CPT) code, American Society of Anesthesiologists (ASA) Physical Status, and hospitalization. *Instructions:* Find the Procedural Severity Score for morbidity (PSS morbidity) in Supplemental Digital Content 1 for the patient’s primary CPT code. Locate this score on the axis labeled “PSS-Morbidity” by interpolating between the displayed values, and draw a vertical line to the “Points” axis to find the points toward 30-day morbidity associated with the patient’s procedure. Repeat this process for ASA Physical Status and hospitalization and add the three resulting point values to obtain the patient’s total points. Locate this number on the “Total Points” axis (*shaded region*) and draw a vertical line to the axis labeled “Predicted Probability of 30-Day Major Morbidity or Mortality (%)” This value represents the prediction for the patient.

We have developed an R package that contains functions for obtaining predictions for an entire data set; it is available on our Web site.**

Discussion

The models previously established for risk adjustment within the ACS-NSQIP database proved to be highly discriminative; consequently, the expected risk for participating general and vascular surgery patients is well-defined. Quality-of-care investigations that compare observed outcome rates to (model-based) expected rates among ACS-NSQIP providers appear to adequately account for differences in risk profiles of individual patient populations. However, external implementation of these models is severely hampered by their dependence on a multitude of patient and procedural risk factors, including many that are not generally available for each patient (such as preoperative laboratory measurements like serum creatinine). Our reduced models for 30-day mortality and morbidity each required only three pieces of data to obtain an estimate of risk and achieved excellent predictive accuracy for 30-day postoperative mortality (c-statistic 0.915 [0.906–0.924]) and morbidity (0.867 [0.858–0.876]). This level of discrimination was only marginally lower than the far-more-complicated ACS-NSQIP mortality¹⁴ and morbidity¹⁵ scores. Even in high-risk patients, observed rates calibrated well with estimated probabilities for mortality (O/E ratio: 0.93 [0.81–1.06]) and morbidity (0.99 [0.93–1.05]).

Quality improvement initiatives face several hurdles in clinical practice. Lack of confidence among physicians and hospitals in the validity of existing outcomes-based quality measures can impede successful clinical quality improvement.²¹ Specifically, there is a perception that currently available outcomes-based measures underrepresent risk associated with high-risk patients. In our models, the observed mortality rate among high-risk patients (as defined by more than 10% predicted probability of mortality) was between 19% less and 6% greater than that expected by our model, with 95% confidence. Similar results were observed for composite morbidity; using more than 20% predicted probability to characterize high-risk patients for this outcome, the observed morbidity rates were not more than 7% different from that expected by our model.

Our models suggest that ASA Physical Status, in conjunction with other variables, provides an accurate depiction of risk. Although ASA Physical Status previously has been shown to measure risk accurately, its implementation in quality-of-care monitoring has been limited. This is because of a perceived lack of availability and an ongoing concern that the scoring is subjective in nature.^{22–24} Although ASA Physical Status is not universally available in administrative datasets, it is a powerful parameter that condenses

relevant clinical measures of patient risk and acuity into a single variable.

The integration of clinical parameters with administrative variables allows for the greatest opportunity to predict baseline risk in surgical patients. Each source, clinical or administrative, has its benefits and disadvantages. The ACS-NSQIP data set as a clinical registry is highly regarded for its data quality. Critics often point out a lack of detail for intraoperative or anesthesia-related variables, but outcome definitions are highly detailed, universally applied, and rigorously audited. Using data from clinical registries such as the ACS-NSQIP brings a certain limitation compared with purely administrative data (mainly based on International Classification of Diseases, Ninth Revision codes) but also offers unique advantages, such as temporal relations. As Orkin states in a recent editorial, clinical data sets can be rich in detail but may lack uniform definitions and miss substantial amounts of data.²⁵ Because administrative data are frequently used for billing purposes, data sets are more complete and uniform. However, diagnostic and procedure codes tend to lack specificity to encode complex clinical scenarios and fail to distinguish temporal relationships.²⁵ We believe one of the strengths of the presented model is achieving an accurate prediction of outcomes with the combined use of clinical and administrative data while using only the most relevant data points.

In addition to use in studies evaluating relative quality of care among providers and health systems, these models may be used to provide specific risk assessments to individual patients. However, we note that physicians should not use our estimates as a basis for decisions between or among procedures in a given patient because our estimates do not represent the (immeasurable) differential risk for individuals among procedures. Instead, they represent the expected risk among large groups of patients with similar characteristics undergoing a given procedure. Patients undergoing a specific procedure will be more homogeneous than a global population of surgical patients. Applying broadly defined quality improvement models to narrowly defined populations likely will result in reduced discriminative ability. Thus, specific risk quantification models for a given disease should be used, when they exist, for predicting individual outcomes. Such models have been developed, for example, for colon resections, esophagectomies in resectable esophageal cancer, and gastric bypass surgery.^{26–28}

We used data on 585,265 patients available from the ACS-NSQIP database at the time of analysis to build our models. It is possible that incremental improvements in predictive accuracy may be achieved by including data from future years because more predictions specific to individual procedures will be possible. That is, more individual procedures will have enough patients to meet the minimum bucket size for PSS estimation, instead of being statistically grouped with other similar procedures. Furthermore, to the extent that providers vary in quality of care, our risk quanti-

** <http://www.clevelandclinic.org/rqi>. Accessed February 21, 2011.

fication models might be further improved by incorporating, say, the treating institution, as a predictor variable.

Our models displayed a high level of predictive accuracy on a randomly withheld validation cohort of 50,000 patients. But developing predictive models requires external validation as well as internal validation. To the extent that outcomes among participating hospitals differ from outcomes among nonparticipating hospitals at a given level of patient/surgical risk, our model may not calibrate well among the nonparticipating institutions. An external validation with complete 30-day outcomes on all hospitals would address this issue. However, identifying a suitable registry for such external validation will be challenging given the quality of the ACS-NSQIP data set and its specific definitions, outcomes, and scope.

Public reporting of provider outcome statistics is a central tenet of national health care quality improvement, and reported results will have considerable financial impact throughout the healthcare system. An obvious use of our prediction models is to risk-adjust estimates of overall performance indices (*i.e.*, observed-to-expected risk ratios) among institutions. Adjusted performance can then be compared across organizations to assess and improve quality of care.

In summary, we developed risk quantification models for 30-day mortality and composite major morbidity using novel severity scoring methodology for each primary CPT code. These models use just three readily available patient and procedural characteristics without appreciably sacrificing predictive accuracy compared with models that require far more information, including information that generally is not available for most patients. Furthermore, these models remain accurate even for high-risk patients. Thus, they can be used to establish accurately the expected risk in various surgical populations for the purpose of fairly comparing quality-of-care among American health-care institutions.

References

- Conway PH, Clancy C: Transformation of health care at the front line. *JAMA* 2009; 301:763-5
- Hibbard JH, Stockard J, Tusler M: Hospital performance reports: Impact on quality, market share, and reputation. *Health Aff (Millwood)* 2005; 24:1150-60
- Hibbard JH, Stockard J, Tusler M: Does publicizing hospital performance stimulate quality improvement efforts? *Health Aff (Millwood)* 2003; 22:84-94
- Lindenauer P: Public reporting and pay-for-performance programs in perioperative medicine: Are they meeting their goals? *Cleve Clin J Med* 2009; 76:S3-8
- Brattwall M, Warrén Stomberg M, Rawal N, Segerdahl M, Houltz E, Jakobsson J: Postoperative impact of regular tobacco use, smoking or snuffing, a prospective multi-center study. *Acta Anaesthesiol Scand* 2010; 54:321-7
- Vacanti CJ, VanHouten RJ, Hill RC: A statistical analysis of the relationship of physical status to postoperative mortality in 68,388 cases. *Anesth Analg* 1970; 49:564-6
- Wolters U, Wolf T, Stützer H, Schröder T: ASA classification and perioperative variables as predictors of postoperative outcome. *Br J Anaesth* 1996; 77:217-22
- Rosenthal MB: What works in market-oriented health policy? *N Engl J Med* 2009; 360:2157-60
- Lee TH, Torchiana DF, Lock JE: Is zero the ideal death rate? *N Engl J Med* 2007; 357:111-3
- Goldman L, Caldera DL, Nussbaum SR, Southwick FS, Krogstad D, Murray B, Burke DS, O'Malley TA, Goroll AH, Caplan CH, Nolan J, Carabello B, Slater EE: Multifactorial index of cardiac risk in noncardiac surgical procedures. *N Engl J Med* 1977; 297:845-50
- Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R: European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999; 16:9-13
- Sessler DI, Sigl JC, Manberg PJ, Kelley SD, Schubert A, Chamoun NG: Broadly applicable risk stratification system for predicting duration of hospitalization and mortality. *ANESTHESIOLOGY* 2010; 113:1026-37
- Cohen NA, Hannenberg AA: Risk stratification index: An important advance in comparing health care apples to oranges. *ANESTHESIOLOGY* 2010; 113:1004-6
- Khuri SF, Daley J, Henderson W, Hur K, Gibbs JO, Barbour G, Demakis J, Irvin G 3rd, Stremple JF, Grover F, McDonald G, Passaro E Jr, Fabri PJ, Spencer J, Hammermeister K, Aust JB: 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; 185:315-27
- Daley J, Khuri SF, Henderson W, Hur K, Gibbs JO, Barbour G, Demakis J, Irvin G 3rd, Stremple JF, Grover F, McDonald G, Passaro E Jr, Fabri PJ, Spencer J, Hammermeister K, Aust JB, Oprian C: 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; 185:328-40
- Khuri SF, Daley J, Henderson W, Hur K, Demakis J, Aust JB, Chong V, Fabri PJ, Gibbs JO, Grover F, Hammermeister K, Irvin G 3rd, McDonald G, Passaro E Jr, Phillips L, Scamman F, Spencer J, Stremple JF: 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
- Khuri SF: The NSQIP: A new frontier in surgery. *Surgery* 2005; 138:837-43
- Fink AS, Campbell DA Jr, Mentzer RM Jr, Henderson WG, Daley J, Bannister J, Hur K, Khuri SF: The National Surgical Quality Improvement Program in non-veterans administration hospitals: Initial demonstration of feasibility. *Ann Surg* 2002; 236:344-53
- Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA: Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; 3:143-52
- Beebe M, Dalton JA, Espronceda M, Evans DD, Glenn RL. *CPT 2007 Standard Edition: Current Procedural Terminology*, 1st edition. Chicago, American Medical Association Press, 2006
- Campbell DA Jr, Dellinger EP: Multihospital collaborations for surgical quality improvement. *JAMA* 2009; 302:1584-5
- Cohen ME, Bilimoria KY, Ko CY, Richards K, Hall BL: Effect of subjective preoperative variables on risk-adjusted assessment of hospital morbidity and mortality. *Ann Surg* 2009; 249:682-9
- Davenport DL, Bowe EA, Henderson WG, Khuri SF, Mentzer RM Jr: National Surgical Quality Improvement Program (NSQIP) risk factors can be used to validate American Society of Anesthesiologists Physical Status Classification (ASA PS) levels. *Ann Surg* 2006; 243:636-41
- Owens WD, Felts JA, Spitznagel EL Jr: ASA Physical Status

- classifications: A study of consistency of ratings. *ANESTHESIOLOGY* 1978; 49:239-43
25. Orkin FK: Risk stratification, risk adjustment, and other risks. *ANESTHESIOLOGY* 2010; 113:1001-3
 26. Bartels H, Stein HJ, Siewert JR: Preoperative risk analysis and postoperative mortality of oesophagectomy for resectable oesophageal cancer. *Br J Surg* 1998; 85:840-4
 27. DeMaria EJ, Portenier D, Wolfe L: Obesity surgery mortality risk score: Proposal for a clinically useful score to predict mortality risk in patients undergoing gastric bypass. *Surg Obes Relat Dis* 2007; 3:134-40
 28. Ondrula DP, Nelson RL, Prasad ML, Coyle BW, Abcarian H: Multifactorial index of preoperative risk factors in colon resections. *Dis Colon Rectum* 1992; 35:117-22

ANESTHESIOLOGY REFLECTIONS

Sir Humphry Davy, Bart., from Lady Davy



In December of 1820, Lady Davy commissioned the new President of the Royal Academy (P.R.A.), Sir Thomas Lawrence, to paint a portrait of her husband, Sir Humphry Davy, to celebrate his election as President of the Royal Society (P.R.S.). At one third that original's size, the smaller oil (*left*) was curatorially donated to the Wood Library-Museum in honor of Colin Davy Bause and Evan Blake Bause. The close up (*right*) reads "Sir Humphry Davy, Bart., from Lady Davy." Months after Davy posed for Lawrence, the two men would meet again. Representing the sciences and the arts, Davy, P.R.S., and Lawrence, P.R.A., would reunite in July of 1821 to walk down the coronation aisle together for the crowning of King George IV. (Copyright © the American Society of Anesthesiologists, Inc. This image also appears in the *Anesthesiology Reflections* online collection available at www.anesthesiology.org.)

George S. Bause, M.D., M.P.H., Honorary Curator, ASA's Wood Library-Museum of Anesthesiology, Park Ridge, Illinois, and Clinical Associate Professor, Case Western Reserve University, Cleveland, Ohio. UJYC@aol.com.