Optimizing ACS NSQIP Modeling for Evaluation of Surgical Quality and Risk: Patient Risk Adjustment, Procedure Mix Adjustment, Shrinkage Adjustment, and Surgical Focus

Mark E Cohen, PhD, Clifford Y Ko, MD, MS, MSHS, FACS, Karl Y Bilimoria, MD, MS, Lynn Zhou, PhD, Kristopher Huffman, MS, Xue Wang, PhD, Yaoming Liu, PhD, Kari Kraemer, PhD, Xiangju Meng, MS, Ryan Merkow, MD, MS, Warren Chow, MD, MS, Brian Matel, MA, Karen Richards, BA, Amy J Hart, BS, Justin B Dimick, MD, MPH, Bruce L Hall, MD, PhD, MBA, FACS

The American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) collects detailed clinical data from participating hospitals using standardized data definitions, analyzes these data, and provides participating hospitals with reports that permit risk-adjusted comparisons with a surgical quality standard. Since its inception, the ACS NSQIP has worked to refine surgical outcomes measurements and enhance statistical methods to improve the reliability and validity of this hospital profiling. From an original focus on controlling for between-hospital differences in patient risk factors with logistic regression, ACS NSQIP has added a variable to better adjust for the complexity and risk profile of surgical procedures (procedure mix adjustment) and stabilized estimates derived from small samples by using a hierarchical model with shrinkage adjustment. New models have been developed focusing on specific surgical procedures (eg, "Procedure Targeted" models), which provide opportunities to incorporate indication and other procedure-specific variables and outcomes to improve risk adjustment. In addition, comparative benchmark reports given to participating hospitals have been expanded considerably to allow more detailed evaluations of performance. Finally, procedures have been developed to estimate surgical risk for individual patients. This article describes the development of, and justification for, these new statistical methods and reporting strategies in ACS NSQIP. (J Am Coll Surg 2013; 217:336–346. © 2013 by the American College of Surgeons)

Motivated by a need to evaluate risk-adjusted surgical quality in Department of Veterans Affairs (VA) hospitals, a set of risk predictors and adverse outcomes, occurring within 30 days of surgery, was defined and applied to VA hospital surgical cases in the later 1990s; surgical quality was expressed in terms of logistic model observed-to-expected (O/E) ratios and evaluation of these outcomes allowed hospitals to identify opportunities for quality improvement, which led to improved surgical care.[1,2] Subsequent to a successful pilot of VA program methods in 14 public-sector hospitals from 2001 to 2004 (funded by an Agency for Healthcare Research and Quality grant), the American College of Surgeons (ACS) expanded the program to other non-VA US hospitals in 2005, and eventually to Department of Defense and international hospitals, resulting in similar improvements in hospital quality.[3,4] Since approximately 2005, the ACS National Surgical Quality Improvement Program (NSQIP) has evolved independently of the VA Surgical Quality Improvement Program (VASQIP; as officially referred to by the VA) by making programmatic changes, changes to procedure and outcomes definitions, and by changing statistical and reporting methodology. The ACS NSQIP now includes >525 participating

Disclosure Information: Nothing to disclose.

Received February 18, 2013; Accepted February 26, 2013.

From the Division of Research and Optimal Patient Care, American College of Surgeons (Cohen, Ko, Bilimoria, Zhou, Huffman, Wang, Liu, Kraemer, Meng, Merkow, Chow, Matel, Richards, Hart, Hall), Surgical Outcomes and Quality Improvement Center, Department of Surgery, Feinberg School of Medicine, Northwestern University (Bilimoria), Chicago, IL, Department of Surgery, University of California Los Angeles, David Geffen School of Medicine (Ko), VA Greater Los Angeles Healthcare System (Ko), Los Angeles, CA, Department of Surgery, Michigan Surgical Collaborative for Outcomes Research and Evaluation, University of Michigan, Ann Arbor, MI (Dimick), Department of Surgery and BJC Healthcare (Hall), Center for Health Policy and the Olin Business School (Hall), Washington University in St Louis, and John Cochran Veterans Affairs Medical Center (Hall), St Louis, MO.

Correspondence address: Mark E Cohen, PhD, American College of Surgeons, 633 N Saint Clair St, 22nd Floor, Chicago, IL 60611-3211. email: markcohen@facs.org

ISSN 1072-7515/13/$36.00
http://dx.doi.org/10.1016/j.jamcollsurg.2013.02.027

Abbreviations and Acronyms

ACS = American College of Surgeons
CPT = Current Procedural Terminology
O/E = observed to expected ratio
OR = odds ratio
SAR = semi-annual report
VA = Veterans Affairs
VASQIP = Veterans Affairs Surgical Quality Improvement Program

hospitals. This article focuses on the application of new statistical methodologies and reporting strategies to the ACS NSQIP.

The ACS NSQIP provides participants with risk-adjusted reports every 6 months (semi-annual reports [SARs] in January/February and July) using data collected during an earlier, rolling, 12-month period (July 1 to June 30, and January 1 to December 31, respectively). The 6- to 7-month delay from the date of last eligible case to the release of the SAR is the result of data collectors (surgical clinical reviewers) having 90 days to complete a case once a 30-day postoperative observational period has ended, and 2 to 3 months required to process the dataset, complete modeling, and construct and post various detailed reports.

Each of these SARs contains results for hundreds of models defined by combinations of outcomes (eg, death, any morbidity, cardiac event, pneumonia, unplanned intubation, ventilator dependence, deep vein thrombosis/pulmonary embolism, renal failure, urinary tract infection, surgical site infection, various regroupings of these, length of postoperative surgical stay, return to operating room, and operation-specific outcomes); surgical specialties (eg, all operations, general surgery, vascular surgery, colorectal surgery); individual operations (eg, colectomy, total knee replacement) or other surgical groupings; available predictors accessed (approximately 45 standard "essential" predictors, limited predictor sets designed for National Quality Forum or Centers for Medicare and Medicaid Services defined "measures," or enhanced predictor sets for "targeted" procedures); and included patient population (eg, patients 65 years of age or older or pediatric patients). We will focus on statistical and reporting issues common to most ACS NSQIP statistical models.

Each of this article's main sections warrants an extended discussion of a larger scope than is presented here. However, it is our intent to provide a panoramic overview of ACS NSQIP modeling strategies and tactics rather than an in-depth and mathematically rigorous discussion of particular topics. This article provides a roadmap for navigating the numerous and inter-related statistical issues in risk determination and hospital profiling, which we think would be particularly helpful for investigators new to this area to understand. We have, wherever possible, provided appropriate references so that readers can investigate topics that are of particular interest to them.

RISK ADJUSTMENT AND LOGISTIC MODELING

It is certain that cases sampled for ACS NSQIP quality assessments will not be identical across hospitals. In fact, because hospitals not only perform different procedures, but also treat less or more complex and urgent patients, inequality is guaranteed. Hospitals will tend to have patients with different levels of endogenous risks for postsurgical complications and a fair comparison of surgical quality, indexed in terms of adverse outcomes (mortality and complications), requires adjustment for these differences in patient characteristics. One statistical approach for providing such adjustment is based on logistic regression. This was the approach used in ACS NSQIP at its inception, but was retired, with only a few exceptions, in 2011. Although current ACS NSQIP modeling has evolved beyond this approach, it will be described here for methodological context.

Logistic regression can be used to construct a prediction equation for binary outcomes (eg, death within 30 days of an operation vs survival). The ACS NSQIP often uses a forward selection process to choose a set of predictor variables for individual models. (Although reporting the presence or absence of adverse outcomes is required in ACS NSQIP, some, typically less important, predictor variables are allowed to be missing. To use all otherwise eligible cases, missing values are imputed using Buck's method.[5]) In forward selection, the strongest available predictor is entered into the model, correlations between remaining predictors and outcomes are adjusted for predictors already in the model, and the next strongest predictor is added until no predictor remains that makes a statistically significant independent contribution to the quality of the model, in the context of all previously entered predictors. Such a model allows for an estimated risk (for any particular result) to be assigned to each patient in the dataset based on each patient's risks factors. Although there is some controversy about using ordered (forward, backward, stepwise) selection, the approach has advantages, including analytic simplicity (with beneficial effects on widespread understanding and face validity) and an ability to easily automate the analytic steps required of the approach across a large number of model variations. However, there are controversies, which are beyond the scope of this article, about whether this

ordered selection of variables is an artificial constraint, whether different types of interactions among variables are neglected, and the extent to which there might be multicollinearity or overfitting.[6]

The strategy of forward selection of logistic regression variables has, as its primary intent, effective prediction. Variables are not selected based on clinical expectations and neither variables nor their resulting coefficients are required to have clinical face validity. Because of this, at times predictor variables will have effects contrary to what might be expected—sometimes this is the result of real artifacts (eg, American Society of Anesthesiologists class 5 might be protective for infection because many of these patients do not live long enough to get an infection), and sometimes this is because, with samples sizes exceeding many hundreds of thousands, there are simply many opportunities for weak predictors to enter a model with very small or even directionally strange effects. These outcomes can often be justified as the mathematically correct influence that a predictor has in the context of all the other predictors in a model (probably operating through a multicollinearity mechanism). Of course, despite the lack of a requirement for face validity about a predictor or its effect, face validity is still desired for its beneficial effect on confidence in any model.

Although there are advantages to building models that are clinically compelling, we have not found that such models routinely offer a predictive advantage. In addition, although clinically-driven models can be justified, even preferred, for purposes of focused research, the luxury of manually designing models variable by variable is often not a feasible option when models must be generated for hundreds of institutions and hundreds of combinations of populations and outcomes, and in a strict time frame. A similar reality holds for the evaluation of potential interactions between variables. The processing and manpower capacity necessary to investigate any number of variable interactions across hundreds of different models would be prohibitively costly in terms of the need for generating hundreds of reports for hundreds of hospitals according to a strict reporting schedule. Investigation of interactions, therefore, remains limited to extremely focused research efforts. It is also the case that, because ACS NSQIP has so many potential predictors and frequently has redundant explanatory power, we have not found that interactions dramatically improve prediction in many of our own investigations.

Applying the prediction equation to each patient modeled within a hospital, it is then necessary to sum the number of observed events and the predicted probabilities (expected events); these 2 sums can then be used to calculate an observed-to-expected (O/E) ratio. A 95% CI is then constructed for this risk-standardized mortality ratio using any of several methods that do, however, differ with respect to their coverage probabilities (ACS NSQIP uses Ulm's method), and a determination is made as to whether it overlaps 1.0.[7] An O/E ratio of 1.0 describes performance that exactly meets the expectation for the patients treated (sometimes called "as expected" or "average" performance). A CI entirely <1.0 indicates performance significantly better than expected (significantly fewer adverse events than expected, "better than average" or "exemplary"). A CI entirely >1.0 indicates performance significantly worse than expected (significantly more adverse events than expected, "worse than average" or "needs improvement").

Regardless of the various performance-level naming conventions, it should be understood that what is technically meant with respect to statistical significance is that if all providers represented in the dataset (the aggregate "average" provider) treated the same patients as the provider in question, and if all of those theoretical performances were compared, the performance of the provider in question would look better or worse (95% CI excludes one) than the aggregate performance of all the other institutions, given this set of patients treated. The interpretation is conditioned on the set of patients treated by the provider in question. The concept of "average" must be applied and understood carefully.

For most models, ACS NSQIP now uses 2 criteria for assigning hospitals to good or poor performance categories. In addition to the 95% CI criterion described, ACS NSQIP also examines where an institution sits in the distribution of hospitals as defined by decile (the distribution of ordered scores partitioned into 10 sequential groups of approximately equal size). Because of stability requirements that will be considered in the section on hierarchical modeling, the use of the decile criterion is limited to distributions of hierarchical model odds ratios (OR) (which, for the present expository purposes, can be interpreted as O/E ratios). Good performance would be defined as being either a low statistical outlier (95% CI < 1), or as being in the first (best) decile of ORs. "Needs improvement" would be defined as being a high statistical outlier (95% CI > 1) or in the 10th (worst) decile. However, the decile criterion is usually not used when a model has failed to detect a single outlier or when the number of hospitals represented in a model is very small.

The programmatic use of decile value in combination with the CI determinations is more of an operational policy issue than a statistical modeling issue. Decile consideration is added to emphasize to hospitals that there are multiple ways to extract information from these

models that can be valuable for driving quality improvement. At times, the incorporation of deciles increases the number of quality "flags" that a hospital might receive, or increases the number of hospitals that are provided quality flags to investigate. It is a mechanism to encourage activity around these assessments. The purpose of measurement is an important consideration though. Higher numbers of decile-based flags are probably more appropriate for internal quality assessments and less appropriate for higher-stakes public reporting.[8] The relative worth of an assessment based on decile vs an assessment based on statistical CI is a topic of some debate; that debate will not be resolved in this article. Suffice to say for now that the experience of ACS NSQIP is that hospitals find the incorporation of both criteria useful for driving quality improvement.

Logistic modeling applied to a common set of patient risk factors was the original analytic approach of both the VASQIP and ACS NSQIP. We provide code to implement a standard logistic regression with SAS PROC LOGISTIC (SAS Institute) in the Appendix (online only).

As mentioned earlier, the VASQIP and ACS NSQIP programs are now independent and have been since roughly 2005. Methods within the programs have evolved apart. The evolution of ACS NSQIP statistical methodology is described here, after a brief discussion of methods used for evaluating the quality of statistical modeling and prediction.

DISCRIMINATION, CALIBRATION, AND BEYOND

Although no model will perfectly reflect reality, there are several statistics that can be used to determine whether a model is useful in risk adjusting for patient-level factors and whether one particular model is better than another. These statistics have also been used to determine which predictors are most essential to particular models. In practical terms, program efficiency is enhanced (data collection and other resource requirements lessened), if models can be built with smaller predictor sets.[9]

The c-statistic (area under the receiver operating characteristic curve), or the sensitivity vs (1-specificity) plot, refers to the proportion of all possible pairings, from the dataset, of one case with an event and one case without an event, where the predicted probability of the former was greater than the predicted probability of the latter. This is referred to as discriminating a "case" from a "non-case," or simply "discrimination."[10] If the c-statistic is 1.0, discrimination is perfect, if is 0.5, discrimination is no better than chance. The c-statistic is not an ideal index of performance because it is based on rank, focuses on category comparisons, and does not directly evaluate the accuracy of prediction.

In comparison, the Hosmer-Lemeshow statistic is a measure of "calibration," or a reflection of bias in predicting risk across the range of risk.[11] It is constructed by ordering cases by predicted risk, creating 10 groups of approximately equal size sequentially from that list, and then constructing a chi-square statistic for the 10×2 table of group (10 levels) by observed events vs predicted events (2 levels). If there is a tendency to over- or underestimate risk for different risk groups, the chi-square statistic will become larger. One problem with the Hosmer-Lemeshow is that it is evaluated as chi-square statistic (it is asymptotically chi-square distributed), so that as the sample size gets larger, smaller and smaller deviations from perfect calibration will appear statistically significant.[11] For this reason, ACS NSQIP has focused on graphical representations of calibration. To improve ease of interpretation, we construct graphs based on sequential groups having equal numbers of observed events rather than equal numbers of patients (low-risk groups will therefore tend to have more patients than high-risk groups).

A third model statistic is the Brier score.[12] It is defined as the mean squared difference between patients' predicted probabilities and observed outcomes (1 or 0 depending on event or nonevent). Because the Brier score is computed from differences between actual events and predicted probabilities, it can be more informative than the rank-based c-statistic. As a model's predicted scores approach 0 and 1 for nonevents and events, respectively, the Brier score will approach 0.0 (perfect prediction). If a probability of 0.5 is assigned to every patient, the Brier score will equal 0.25. Another useful benchmark for the Brier score is its value when the observed overall event rate is assigned to each patient. The Brier score for estimates coming out of this "null model" allows one to evaluate the added predictive contribution from patient-level risk factors. It is also important to note that the Brier score reflects discrimination and calibration simultaneously.[13]

Other model metrics to consider include "net reclassification index," "integrated discrimination improvement," generalized R^2, and information indices, such as the Akaike information criterion, which impose a penalty related to the number of predictors.[14,15] All these measures have various strengths and weaknesses.[10,11,13,16-18]

An important point, which is sometimes not understood, is that these model attributes apply to the magnitude of model-based risk adjustment, but not to its necessity.[19,20] Although a c-statistic approaching 1, a Hosmer-Lemeshow statistic approaching 0, and a Brier score approaching 0.0, all indicate that a model was effective and necessary, poor model metrics can mean only that risk adjustment was

unnecessary. If metrics show poor model performance, this might be the result of patients having so much inherent similarity, that risk-adjustment models are ineffective simply because there are no differences in risk mandating adjustment.[19] Despite this possibility that the fixed patient-level effects (patient characteristics) are nonsignificant, the hospital-level random effects might well be; hospitals can clearly influence surgical outcomes even if patients have identical risk profiles. Evaluation of patient-level risk factors is useful, but not necessarily definitive with respect to the efficacy of hospital profiling.

EVOLUTION OF ACS NSQIP STANDARD APPROACHES OVER TIME

Enhancing adjustment for case and procedure (July 2010)

The mix of cases performed by hospitals can be different in at least 2 obvious ways. The first is on the basis of patient characteristics—how sick is the patient or how complicated is their medical history? This is the traditional main focus of risk adjustment. But when it comes to assessing the quality of surgical services, another no less important axis is the profile of procedures actually performed. No one would argue that an esophagectomy is more complicated and more prone to complications than an inguinal hernia repair, even if performed on the "same" 50-year-old patient. Now the adjective "same" should be considered because it is likely that the 50-year-old undergoing a hernia repair has a different medical history than the 50-year-old undergoing an esophagectomy. But capturing the differences in medical history can be challenging, and capturing aspects of the procedure performed dramatically facilitates risk adjustment overall. It is useful to conceptualize capturing elements of the surgical procedure (apart from patient characteristics) as "procedure mix adjustment."

Adjustment for procedure mix or risk/difficulty in ACS NSQIP had originally been based on the following variables: a 9-level Current Procedural Terminology (CPT) categorical variable based on organ system (added in 2007) and a continuous variable defined by work relative value units (present at program inception) published by the Centers for Medicare and Medicaid Services. Although relative value units are reasonably appropriate for this purpose (having been designed to capture a measure of the effort required to perform a surgery), it is clear in retrospect that organ system CPT ranges are likely to be ineffective in discriminating between small and large procedures. For example, appendectomy and a total proctocolectomy were in the same range. For this reason, an alternative strategy was needed to more effectively use the information encapsulated in CPT codes. The logic of this approach was first described in 2009 and 2010, but the methodology used in subsequent SAR reporting periods has been enhanced.[21,22]

Ideally, it would be desirable to enter every CPT code as a categorical variable in each regression model. This would, in one sense, standardize the evaluation of each surgical case to the common observed results for that case code. However, with about 3,000 eligible CPT codes, there would typically be many empty data cells (independent and dependent variables), which would lead to failure of model convergence. To reduce this problem, ACS NSQIP developed an approach that creates a continuous linear variable to represent the "endogenous" risk that each CPT code category carries for a particular outcomes model. It should be noted that various machine learning approaches (eg, support vector machine) could potentially accommodate analysis of 3,000 separate codes, but implementation of these methods and interpretation of their findings have their own challenges, and often results are not considerably different from those using logistic regression.[23,24]

The ACS NSQIP approach begins by classifying all CPT codes into approximately 300 clinically meaningful categories, using a 4-tiered hierarchical classification system. The hierarchical levels of the classification are roughly based on physiological/organ system (eg, gastrointestinal), operation (eg, resection), body part (eg, colon), and modifier (eg, emergent, laparoscopic, or surgical subtype). The ACS NSQIP outcomes are then predicted via regression using the 300-level CPT category variable and a small number of powerful predictors (eg, age, American Society of Anesthesiologists class), which are themselves often related to the surgery type. By including these other variables, the result is a less biased, purer estimate of risk attributable to CPT code alone. Logits (linearly scaled probability) associated with each CPT group are then used as a single continuous variable in follow-on models. This CPT risk variable is often the most powerful predictor in ACS NSQIP models; it is referred to using the shorthand "CPT category linear risk."

We have recently revisited this problem and arrived at a different solution by leveraging the information we have available in >1 million records collected by ACS NSQIP in the last 3 to 4 years. By using a hierarchical model (with shrinkage as described here) on these data to predict outcomes for individual CPT codes (rather than CPT category), we were able to get more accurate estimates of risk. We are currently investigating a full transition from CPT category risk to individual CPT code risk in hospital profiling and surgical risk estimation.

Hierarchical modeling (July 2011)

The main advantages of hierarchical models (incorporating a shrinkage adjustment) for hospital surgical quality profiling have been well documented and can generally be assigned to 4 attributes that are described here.[25-33] It should be noted that these hierarchical model advantages have to do, primarily, with estimation of the "hospital" effect. Methods to evaluate model quality in terms of discrimination, calibration, etc, which were described previously, generally focus on patient-level fixed effects. It is often the case that logistic and hierarchical models perform similarly with respect to estimating the fixed effects, even though there might be profound differences in estimated hospital effects (between O/E ratios constructed post-logistic modeling vs ORs estimated within the hierarchical model or hierarchical model O/E ratios). The point that logistic and hierarchical models can yield nearly identical estimates of the fixed effects, but that this is not germane to determining the value of multilevel modeling for purposes of hospital profiling, was made by Clark and colleagues in an appropriate criticism of an earlier article.[28,34]

1. Although logistic models assume independence among observations, this is not actually the case; patients are grouped (nested) within hospitals and this affects variance estimates. As a result, a logistic implementation would typically underestimate variability in comparison with a hierarchical model and overstate the certainty of results. Hierarchical models directly address the hierarchical structure of this clustered data, more appropriately accounting for within- and between-hospital sources of variability.
2. When logistic regression is used, results are reported in terms of a large number of postmodeling hospital O/E ratios. Because there are many tests, each conducted on different data subsets, there are many opportunities for false-positive findings. In contrast to the logistic O/E ratio, the hierarchical OR is based, in part, on data from all hospitals; this mitigates inflation in false positives due to multiple testing.
3. When a hospital provides a small number of cases to a model, resultant estimates can be very unreliable. It is quite possible that when a hospital submits 10 cases, for example, that the logistic model O/E ratio will be 0 (if there are no observed events) or very, very large (if the hospital was unlucky enough to have a patient who experienced a rare event). These estimates are intuitively unreasonable; an ACS NSQIP hospital's true O/E ratio for morbidity will not actually be 0.0 and its true O/E ratio for mortality rate will not be, for example, 20.0. Rather than using these unreliable estimates when sample size is small, it is more reasonable to combine the limited hospital-level information we have with the average performance across all hospitals to arrive at a more realistic prediction of hospital quality. This concept is often referred to as "shrinkage" (toward the grand mean).

Although this concept can be implemented in different ways, a compelling attribute of hierarchical or generalized linear mixed models (as, for example, executed with SAS PROC GLIMMIX) is the ability to construct estimates incorporating an empirical, Bayes-like, shrinkage adjustment. This adjustment is constructed to optimally combine information from the particular hospital with information from the sample of all hospitals, to arrive at a best prediction for each hospital's performance. Sometimes called "reliability adjustment," but alternately described as shrinkage, smoothing, or pooling, this adjustment tends to shrink predicted hospital performance toward the grand mean hospital value, with the magnitude of shrinkage greatest when the hospital's sample size is small or when the hospital's estimated performance is extreme compared with other hospitals.[35-38] Still, there are controversies around the principle of shrinkage, such as whether a low-volume provider should be credited with average performance.

4. Although not an issue directly related to the modeling approach, because of technical reasons, the CI, which can be constructed around the logistic or hierarchical O/E ratio estimate, is less accurate than one that can be constructed around the hierarchical model OR (the risk- and shrinkage-adjusted odds for the event at the hospital compared with the odds at the average ACS NSQIP hospital). For this reason, we now report results in terms of hospital ORs rather than O/E ratios. (As described in the Appendix [online only], the hierarchical analog of the logistic O/E ratio is the BLUP/NOBLUP ratio, which, like the OR, includes shrinkage.[39] BLUP stands for "best linear unbiased prediction" and includes both the hospital random effect and the patient-level fixed effects; NOBLUP includes only the fixed effects.) The OR CI can be interpreted in the same way as the CI about the O/E ratio. If the CI overlaps 1.0, the hospital is doing "as expected." If the CI is entirely <1.0, or entirely >1.0, then the hospital is a statistical outlier.

As indicated in item 3, a compelling attribute of hierarchical models is the ability to incorporate shrinkage adjustments.[35-38] Because hierarchical model ORs are more stable than logistic model O/E ratios, it is more appropriate for ORs to be used to construct informative

deciles for purposes of categorizing hospital surgical quality.

Unfortunately, acceptance of "shrunken" ORs and the underlying multilevel modeling methodology by ACS NSQIP participants has not always been enthusiastic, perhaps because of their complexity vs logistic model O/E ratios. Careful explanation is required when an institution with no observed events (which previously would have received an O/E ratio of 0.0) is assigned a shrunken OR of, say, 0.97. Also, because of asymmetry (positive skew) in binomial distributions when sample sizes and probabilities are both small, many more hospitals saw their ORs (with reference to O/E ratios) moving upward from 0.0 toward 1.0, than downward from very high O/E ratio values toward 1.0. There are several easily accessible references on the benefits of shrinkage adjustment that do not require an understanding of the associated topics of hierarchical modeling and Bayesian methods.[40,41]

Although this work has described hierarchical models generically, this term actually refers to a wide variety of hierarchical or multilevel modeling designs that can be implemented through many different statistical programs.[42,43] The ACS NSQIP uses a relatively simple hierarchical implementation where there is a single hierarchy, with patients nested within hospitals, with a random intercept (the hospital effect) and fixed slopes (we do not attempt to evaluate possible interactions between hospitals and patient-level fixed effects). Analyses are performed with SAS GLIMMIX (generalized linear mixed models). A generic example of the SAS code implementation is provided in the Appendix (online only).

As described in the section on logistic regression, ACS NSQIP models were traditionally built using a forward selection process. However, forward (and stepwise) selection is unavailable in GLIMMIX. Although there are macros to accomplish this, these would be suboptimal for ACS NSQIP models because running a hierarchical model on datasets as large as those in the ACS NSQIP sometimes requires in excess of 30 minutes per step. For this reason, NSQIP uses forward logistic regression to select predictors and then uses that predictor set in a subsequent hierarchical model.

Modeling targeted procedures with procedure-specific variables (January 2013)

Originally, VASQIP and ACS NSQIP models used variables and outcomes common to all procedures. Models focused on large groups of procedures (eg, general surgery or vascular surgery) and outcomes relevant to procedures ranging from colectomy to hip replacement (eg, surgical site infection, pneumonia, etc). This was a specific design decision to facilitate broad assessments of institutions. However, it has become clear over time that different operations can have unique risk factors and that some important adverse outcomes might also be associated with only specific operations. In addition, increasing granularity (eg, anastomotic leak after colon surgery) is a more actionable target for quality improvement.

To provide this procedure-specific information to hospitals, ACS NSQIP developed a "Procedure-Targeted" program that has been rolled out in phases starting in July 2011. Hospitals select the operations of interest to the institution from a list of 34 surgical groups offered by ACS NSQIP that covers all surgical subspecialties (eg, pancreatectomy, lower-extremity bypass, hip replacement, brain tumor resection). Under this program, hospitals collect special additional procedure-specific data. Data are analyzed using both variables common to all of ACS NSQIP (referred to as "standard" or "essential" variables) and special procedure-specific variables (risk indicators and outcomes) specific to the target procedure. In addition, by focusing on surgical procedure groups, it becomes possible to build informative variables reflecting the indication for the procedure using ICD-9 codes and risk adjust for this new variable (eg, colectomy for diverticulitis, inflammatory bowel disease, cancer, etc). This would not be possible for standard ACS NSQIP models, as the same indications are not common to a broad collection of procedure types (eg, general surgery models).

Real-time risk adjustment (July 2012)

Because the ACS NSQIP SARs are released for operations that have occurred from 6 to 18 months previously, there are limitations on inferring real-time surgical quality from the SAR. One effective solution to this problem is to monitor, as is possible on the ACS NSQIP data collection platform, unadjusted rates in real time. It is a reasonable first approximation to assume (for quality-monitoring purposes) that increases or decreases in raw rates will reflect both current quality and changes in quality over time. These first approximations based on raw rates can later be supported or modified in the face of risk-adjusted SAR information.

Risk-adjusted reports in real time would provide estimates that would be more beneficial to institutions. Conceptually, this could be accomplished by putting current data into recent earlier regression equations. However, statistical methods to accomplish this using hierarchical models and providing for sample-size-based shrinkage adjustments can be very complex.[44] The ACS NSQIP has approached this problem using a series of simplifying strategies and now provides

real-time, risk-adjusted quality estimates for the 6 ACS NSQIP measures models (Lower Extremity Bypass Death/Serious Morbidity; Colon Surgery Death/Serious Morbidity; Overall Death/Serious Morbidity; Deep and Organ Space Surgical Site Infection; Death/Serious Morbidity in Patients 65 Years of Age and Older; Urinary Tract Infection).

To obtain these reports, hospitals first choose (on the web-based platform), a time period to be evaluated. The data for this period are then processed using the results of regressions and transformations from the previous SAR cycle for each model. By applying these equations to the current data, expected event numbers can be estimated as well as the hospital's O/E ratio. This O/E ratio is then transformed (post analysis), with an extramodeling shrinkage adjustment, so that it more closely approximates the hierarchical model OR that would be reported for these data.[45] The transformation is applied so that, as much as possible, the correlation between the SAR data's O/E ratio and OR is maximized, and mean difference between them is minimized. Finally, the intercept and slope from a regression equation estimating the OR from the O/E ratio (applied to the earlier SAR data) are used as correction factors for the real-time data. For 6 recent SAR datasets, this approach achieved correlations between hierarchical model ORs and transformed O/E ratios of at least 0.985, with regression intercepts of 0.0 and slopes of 1.0. The ACS NSQIP will provide real-time risk adjustment using these methods for more models over time.

Targeted and universal surgical risk calculators (2009 and 2013)

Both logistic and hierarchical models yield, via prediction equations, estimated probabilities for events for each patient in the dataset. From this it is a simple matter to use these equations to produce estimates for patients not in the original dataset. The ACS NSQIP made its first surgical risk calculators available in 2009.[46] Risk calculators are available for colorectal surgery, cholecystectomy, bariatric surgery, and ventral hernia repair.

The approach of having operation-specific calculators was driven by the expectation that understanding the surgical indication, which is typically operation specific, was required to have effective models. This assumption was re-evaluated in 2012. The essential question was whether a universal model built with approximately 18 variables, including the previously described linear risk for CPT group (or individual CPT), but without any indication variable, using >1 million cases, would be as good as targeted models using target-specific indication on sample sizes of tens of thousands. Analyses revealed little difference in Brier scores for universal vs colorectal- and cholecystectomy-targeted models, both of which have important indication variables. As a result, a universal surgical risk calculator covering all ACS NSQIP-eligible CPT codes predicting 9 different adverse outcomes will be available in 2013.

QUALITY REPORTING: FROM A SINGLE SEMI-ANNUAL REPORT DOCUMENT WITH PSEUDO-IDENTIFICATIONS, TO HOSPITAL-SPECIFIC DOCUMENTS

The first SARs contained all information about all models for all hospitals. Hospital profiling results were presented in terms of "caterpillar" plots (Fig. 1) with hospitals identified by pseudo-identifications. As the number of models and participating hospitals increased, this reporting scheme became untenable, as staff at each hospital would have to search the x-axis of each model for their identification among hundreds. In addition, from a purely spatial display consideration, caterpillar plots are inefficient—each model requires up to 2 pages for legible display, and the shape of caterpillars are usually very similar (no unique information is provided). It is also apparent that caterpillars display between-hospital differences on a rank scale that exaggerates or diminishes real distinctions between hospitals (as indexed by differences in O/E ratios or ORs), as a function of where they are on the distribution.

Although retaining the SAR as a common and familiar introductory document, beginning in 2010 a transition was made to reporting site-specific information in site-specific documents. These include caterpillar plots in PowerPoint presentations, where the site is automatically identified; a tabular report giving the site details about all models for which it had data; and hospital-specific bar plots (Fig. 2), which condense caterpillars, retaining only the most critical information. Development of new reporting strategies remains a high priority for ACS NSQIP. In addition, the web-based data-collection platform has many reporting options for displaying non-risk-adjusted data in an on-demand fashion.

It should be noted that the y-axis on both caterpillar plots and hospital-specific bar plots is technically improper.[47] From an "expected" O/E ratio or OR of 1.0, values can range to 0.0 below, but to infinity above. However, in terms of absolute effect magnitude, an OR of 2.0 is equivalent to an OR of 0.5, an OR of 4.0 is equivalent to an OR of 0.25, 10 to 0.1, and so forth. So, ideally, the y-axis should be log scaled: it should present the log OR rather than the OR. This change in reporting display was considered but rejected based on concern that it would make the quality metric less intuitive.

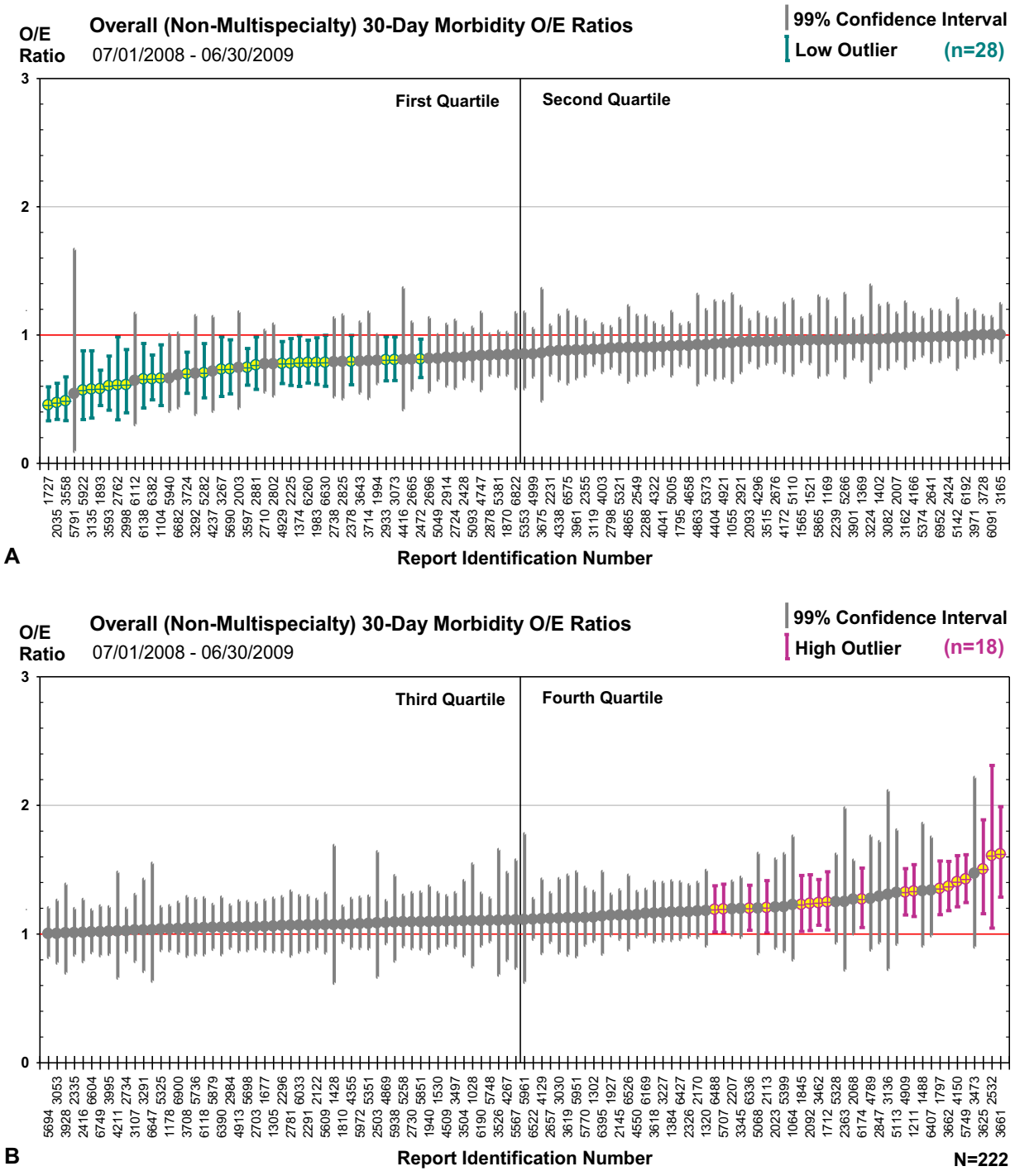

Figure 1. Caterpillar plot for 30-day morbidity following general or vascular surgery (during this semi-annual report time period referred to as "Overall"). Color coding is by outlier status. Although providing a graphical representation of the observed-to-effect (O/E) ratio and rank standing among hospitals, this plot has some shortcomings. It requires a great deal of space, the rank scale distorts the magnitude of real between-hospital differences in the O/E ratio, and it is difficult for hospitals to locate their pseudo-identification when that process is not automated. (A) First and second quartile. (B) Third and fourth quartile.

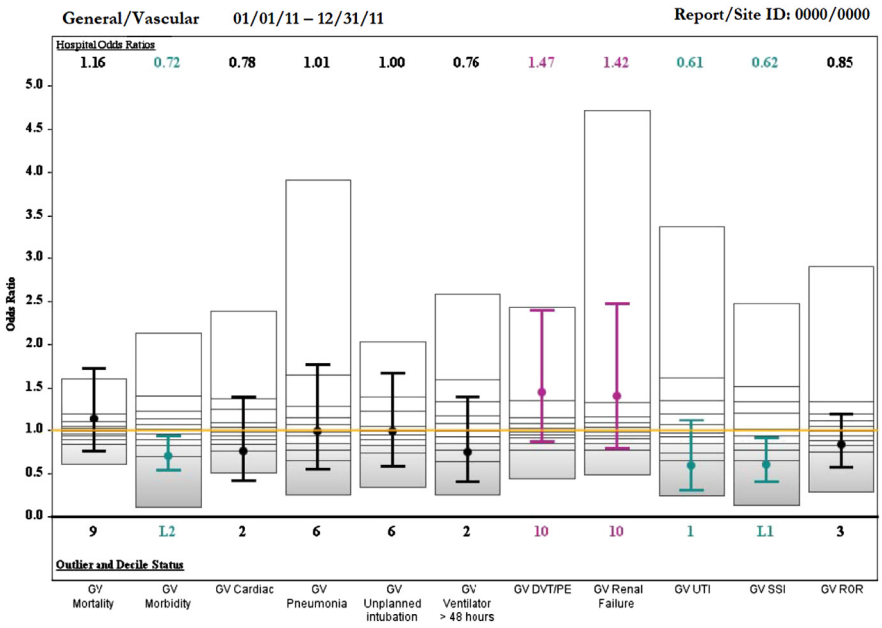

Figure 2. Hospital-specific bar plots. Each box represents the distribution of odds ratios for hospitals in the model; the bottom and top give the smallest and largest values, and the horizontal lines give decile demarcations. The point and vertical line within each box give the hospital's odds ratio and CI. Color coding is by outlier or decile status; these attributes are identified below the bar. Because information in the caterpillar has been condensed, up to 12 bars can reasonably be reported on a single page, in this case all general/vascular surgery (GV) outcomes models are reported. DVT, deep vein thrombosis; PE, pulmonary embolism; ROR, return to operating room; SSI, surgical site infection; UTI, urinary tract infection.

CONCLUSIONS

Modeling methodology has evolved rapidly in ACS NSQIP during the past 8 years. Although this has been challenging for staff and participants, hospital profiling results are of considerable consequence for allocation of quality-improvement resources, patient safety, and public reporting, and require that best methods be used. Still, best always represents a well-considered balance of statistical rigor, implementation practicality, and ease of understanding by users. Given the development and distribution of educational materials and increased speed in statistical data processing, rigor becomes more attainable. It is not anticipated that ACS NSQIP modeling will ever stop evolving—it will always be necessary to continuously evaluate how the program collects, analyzes, and presents data to provide state-of-the-art validity, reliability, and utility for quality improvement that is accessible to users and stakeholders.

Author Contributions

Study conception and design: Cohen, Ko, Bilimoria, Hall
Acquisition of data: Cohen, Zhou, Huffman, Wang, Liu, Kraemer, Meng, Matel, Richards, Hart
Analysis and interpretation of data: Cohen, Zhou, Huffman, Wang, Liu, Kraemer, Meng, Merkow, Chow, Dimick
Drafting of manuscript: Cohen, Hall
Critical revision: Cohen, Ko, Bilimoria, Zhou, Huffman, Wang, Liu, Kraemer, Meng, Merkow, Chow, Matel, Richards, Hart, Dimick, Hall

REFERENCES

1. 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.
2. 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.
3. Hall BL, Hamilton BH, Richards K, et al. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. Ann Surg 2009;250:363–376.
4. Khuri SF, Henderson WG, Daley J, et al. Successful implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the private sector: the Patient Safety in Surgery study. Ann Surg 2008;248:329–336.

5. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J R Stat Soc Ser B 1960;22:302–306.
6. Livingston E, Cao J, Dimick JB. Tread carefully with stepwise regression. Arch Surg 2010;145:1039–1040.
7. Ulm K. A simple method to calculate the confidence interval of a standardized mortality ratio (SMR). Am J Epidemiol 1990;131:373–375.
8. Bilimoria KY, Cohen ME, Merkow RP, et al. Comparison of outlier identification methods in hospital surgical quality improvement programs. J Gastrointest Surg 2010;14: 1600–1607.
9. Dimick JB, Osborne NH, Hall BL, et al. Risk adjustment for comparing hospital quality with surgery: how many variables are needed? J Am Coll Surg 2010;210:503–508.
10. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
11. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: the Hosmer-Lemeshow test revisited. Crit Care Med 2007;35:2052–2056.
12. Brier GW. Verification of forcasts expressed in terms of probability. Monthly Weather Rev 1950;78:1–3.
13. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–138.
14. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157–172; discussion 207–212.
15. Allison PD. Logistic Regression Using SAS: Theory and Application. 2nd ed. Cary, NC: SAS Publishing, Inc; 2012.
16. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biometric J Biometrische Zeitschrift 2008;50:457–479.
17. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–935.
18. Hilden J, Gerds TA. Evaluating the Impact of Novel Biomarkers: Do Not Rely on IDI and NRI. Copenhagen, Sweden: Department of Biostatics, University of Copenhagen; 2012.
19. Merkow RP, Hall BL, Cohen ME, et al. Relevance of the c-statistic when evaluating risk-adjustment models in surgery. J Am Coll Surg 2012;214:822–830.
20. Austin PC, Reeves MJ. The relationship between the C-statistic of a risk-adjustment model and the accuracy of hospital report cards: a Monte Carlo study. Med Care 2013;51:275–284.
21. Hall BL, Hsiao EY, Majercik S, et al. The impact of surgeon specialization on patient mortality: examination of a continuous Herfindahl-Hirschman index. Ann Surg 2009;249:708–716.
22. Raval MV, Cohen ME, Ingraham AM, et al. Improving American College of Surgeons National Surgical Quality Improvement Program risk adjustment: incorporation of a novel procedure risk score. J Am Coll Surg 2010;211:715–723.
23. Noble WW. What is a support vector machine? Nat Biotechnol 2006;24:1565–1567.
24. Verplancke T, Van Looy S, Benoit D, et al. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. BMC 2008;8:56.
25. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. Ann Intern Med 1997;127:764–768.
26. DeLong ER, Peterson ED, DeLong DM, et al. Comparing risk-adjustment methods for provider profiling. Stat Med 1997;16:2645–2664.
27. Normand SL, Shahian DM. Statistical and clinical aspects of hospital outcomes profiling. Stat Sci 2007;22:206–226.
28. Clark DE, Hannan EL, Wu C. Predicting risk-adjusted mortality for trauma patients: logistic versus multilevel logistic models. J Am Coll Surg 2010;211:224–231.
29. Moore L, Hanley JA, Turgeon AF, Lavoie A. Evaluating the performance of trauma centers: hierarchical modeling should be used. J Trauma 2010;69:1132–1137.
30. DeLong E. Hierarchical modeling: its time has come. Am Heart J 2003;145:16–18.
31. Shahian DM, Torchiana DF, Shemin RJ, et al. Massachusetts cardiac surgery report card: implications of statistical methodology. Ann Thorac Surg 2005;80:2106–2113.
32. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: comprehensive review and statistical critique. Ann Thorac Surg 2001;72:2155–2168.
33. Mukamel DB, Glance LG, Dick AW, Osler TM. Measuring quality for public reporting of health provider quality: making it meaningful to patients. Am J Public Health 2010;100:264–269.
34. Cohen ME, Dimick JB, Bilimoria KY, et al. Risk adjustment in the American College of Surgeons National Surgical Quality Improvement Program: a comparison of logistic versus hierarchical modeling. J Am Coll Surg 2009;209:687–693.
35. Dimick JB, Staiger DO, Birkmeyer JD. Ranking hospitals on surgical mortality: the importance of reliability adjustment. Health Serv Res;45:1614–1629.
36. Osborne NH, Ko CY, Upchurch GR Jr, Dimick JB. The impact of adjusting for reliability on hospital quality rankings in vascular surgery. J Vasc Surg 2011;53:1–5.
37. Dimick JB, Ghaferi AA, Osborne NH, et al. Reliability adjustment for reporting hospital outcomes with surgery. Ann Surg 2012;255:703–707.
38. Jackman S. Bayesian Analysis for the Social Sciences. Chichester, UK: John Wiley and Sons, Ltd; 2009.
39. Kipnis P, Escobar GJ, Draper D. Effect of choice of estimation method on inter-hospital mortality rate comparisons. Med Care 2010;48:458–465.
40. Greenland S. Principles of multilevel modelling. Int J Epidemiol 2000;29:158–167.
41. Efron B, Morris CN. Stein's paradox in statistics. Sci Am 1977;236:119–127.
42. Leyland AH, Goldstein HE. Multilevel Modelling of Health Statistics. Chichester: John Wiley & Sons, Ltd; 2001.
43. Brown H, Prescott R. Applied Mixed Models in Medicine. Chichester: John Wilery & Sons, Ltd; 1999.
44. Clark DE, Hannon EW, Raudenbush SW. Using a hierarchical model to estimate risk-adjusted mortality for hospitals not included in the reference sample. Health Serv Res 2010;45: 577–587.
45. Noren GN, Hopstadius J, Bate A. Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery. Stat Methods Med Res 2013;22:57–69.
46. Cohen ME, Bilimoria KY, Ko CY, Hall BL. Development of an American College of Surgeons National Surgery Quality Improvement Program: morbidity and mortality risk calculator for colorectal surgery. J Am Coll Surg 2009;208:1009–1016.
47. Levine MA, El-Nahas AI, Asa B. Relative risk and odds ratio data are still portrayed with inappropriate scales in the medical literature. J Clin Epidemiol 2010;63:1045–1047.

APPENDIX.

Generic SAS code for logistic (PROC LOGISTIC) and hierarchical (PROC GLIMMIX) models for generation of observed-to-expected (O/E) ratios and odds ratios. The input data, "dataset" has 5 variables: hospital (identified as 001 to 005), outcomes (with event coded as 1), patientvar1; patientvar2; and patientvar3. There are assumed to be multiple lines (patients) per hospital. All predictors are assumed categorical; continuous predictors should be removed from the class statement. Predictors not selected in the logistic model for O/E values should be removed from the other model statements.

These examples do not include data manipulation to construct the dataset or analysis of output files that are produced by the procedures. For example, construction of O/E ratios would require summing of O, E, and n, by hospital (coded directly or by using, for example, PROC SUMMARY). A macro for constructing the O/E confidence interval using Ulm's method is provided.

The output dataset, "logistic_oe," includes the predicted probability. Subsequent coding will need to sum the number of cases, cases where outcomes=1, and predevent to construct the O/E and CI for each hospital.

```
proc logistic data = dataset;
   class patientvar1 patientvar2 patientvar3 / param =
   ref;
   model outcome (event = '1') = patientvar1 patient-
   var2 patientvar3
   / selection = forward rule = single;
   output out = logistic_oe pred = predevent;
run;
```

Macro for computing the O/E confidence interval using Ulm's method.

```
/*************************************************/;
/*** Description: Generates confidence interval for the
  O/E ratio ***/;
/*** Parameters: indsn - name of the input data set ***/;
/*** alpha - (100 - Alpha) level ***/;
/*** parm1 - Variable for sum of observed events ***/;
/*** parm2 - Variable for sum of expected events ***/;
/*** parm3 - Variable for number of observations ***/;
/*** outdsn - name of the output data set ***/;
/*** Macro Call: %conintChi(ratio,95,obs,exp,nobs,
  oeratio) ***/;
/*************************************************/;

%macro
  conintChi(indsn,alpha,parm1,parm2,parm3,outdsn);
    data &outdsn;
    set &indsn(rename=(&parm1=o &parm2=e
    &parm3=n));
    alpha = (1 + (&alpha/100))/2;
    df2=2*(o+1);
    df1= (2*o);
    chil=CINV(1-alpha, df1);
    if chil='.' then chil=0;
    chiu=CINV(alpha, df2);
    rl=chil/(2*e);    /*** Lower Limit ***/
    ru=chiu/(2*e);    /*** Upper Limit ***/
    oe=o/e;
    drop df1 df2 chil chiu alpha;
    run;
%mend conintChi;
```

The output dataset, 'glimmix_oe,' includes the BLUP and NOBLUP predicted probabilities. Subsequent coding will need to sum BLUP, NOBLUP, and n for each hospital to construct the hierarchical analog of O/E ratio (BLUP/NOBLUP) and CI. The output dataset, 'glimmix_or,' includes the parameter estimates (SolutionR) and CI for hospitals. Subsequent coding will need to exponentiate these to get the OR and CI for each hospital.

```
proc glimmix data = dataset method=quad;
   class hospital patientvar1 patientvar2 patientvar3;
   model outcome (event='1')= patientvar1
   patientvar2 patientvar3
/link=logit dist = binary solution;
   random intercept /subject = hospital solution cl;
   output out = glimmix_oe
   pred(ilink blup) = obs_outcome
   pred(ilink noblup) = exp_outcome
   ods output SolutionR = glimmix_or;
run;
```