The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1—Background, Design Considerations, and Model Development

David M. Shahian, MD, Jeffrey P. Jacobs, MD, Vinay Badhwar, MD, Paul A. Kurlansky, MD, Anthony P. Furnary, MD, Joseph C. Cleveland, Jr, MD, Kevin W. Lobdell, MD, Christina Vassileva, MD, Moritz C. Wyler von Ballmoos, MD, PhD, Vinod H. Thourani, MD, J. Scott Rankin, MD, James R. Edgerton, MD, Richard S. D'Agostino, MD, Nimesh D. Desai, MD, PhD, Liqi Feng, MS, Xia He, MS, and Sean M. O'Brien, PhD

Department of Surgery and Center for Quality and Safety, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (DMS); Division of Cardiac Surgery, Johns Hopkins University School of Medicine, Baltimore, Maryland (JPJ); Division of Cardiovascular Surgery, Johns Hopkins All Children's Heart Institute, St. Petersburg, Florida (JPJ); Department of Cardiovascular and Thoracic Surgery, West Virginia University, Morgantown, West Virginia (VB, JSR); Division of Cardiac Surgery, Columbia University, New York, New York (PAK); Starr-Wood Cardiothoracic Group, Portland, Oregon (APF); Division of Cardiothoracic Surgery, University of Colorado Anschutz School of Medicine, Aurora, Colorado (JCC); Atrium Health, Cardiovascular and Thoracic Surgery, Charlotte, North Carolina (KWL); Division of Cardiac Surgery, University of Massachusetts Medical School, Worcester, Massachusetts (CV); Houston Methodist DeBakey Heart and Vascular Center, Houston, Texas (MCWvB); Department of Cardiac Surgery, MedStar Heart and Vascular Institute, Georgetown University, Washington, DC (VHT); The Heart Hospital Baylor Plano, Plano, Texas (JRE); Division of Thoracic and Cardiovascular Surgery, Lahey Hospital and Medical Center, Burlington, Massachusetts (RSD); Division of Cardiothoracic Surgery, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania (NDD); and Duke Clinical Research Institute, Duke University Medical Center, Durham, North Carolina (LF, XH, SMO)

Background. **The last published version of The Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database (ACSD) risk models were developed in 2008 based on patient data from 2002 to 2006 and have been periodically recalibrated. In response to evolving changes in patient characteristics, risk profiles, surgical practice, and outcomes, the STS has now developed a set of entirely new risk models for adult cardiac surgery.**

Methods. **New models were estimated for isolated coronary artery bypass grafting surgery (CABG [n = 439,092]), isolated aortic or mitral valve surgery (n = 150,150), and combined valve plus CABG procedures (n = 81,588). The development set was based on July 2011 to June 2014 STS ACSD data; validation was performed using July 2014 to December 2016 data. Separate models were developed for operative mortality, stroke, renal failure, prolonged ventilation, reoperation, composite major morbidity or mortality, and prolonged or short postoperative length of stay. Because of its low occurrence rate, a combined model incorporating all operative types was developed for deep sternal wound infection/mediastinitis.**

Results. **Calibration was excellent except for the deep sternal wound infection/mediastinitis model, which slightly underestimated risk because of higher rates of this endpoint in the more recent validation data; this will be recalibrated in each feedback report. Discrimination (c-index) of all models was superior to that of 2008 models except for the stroke model for valve patients.**

Conclusions. **Completely new STS ACSD risk models have been developed based on contemporary patient data; their performance is superior to that of previous STS ACSD models.**

(Ann Thorac Surg 2018;105:1411–8)

Although recalibrated each harvest since their development in 2008, the last published version of The Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database (ACSD) risk models was based on patient data from 2002 to 2006. To incorporate evolving changes in patient characteristics, risk profiles, surgical practice, and outcomes, the STS has developed a set of entirely new risk models for adult cardiac surgery.

In this two-part report, we present the 2018 STS adult cardiac surgery risk models. Part 1 provides an introductory background regarding the history of STS risk modeling, general principles used to develop these

Accepted for publication March 9, 2018.

The STS Executive Committee approved this document.

Address correspondence to Dr Shahian, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114; email: dshahian@partners.org.

Published by Elsevier Inc.

0003-4975/$36.00
https://doi.org/10.1016/j.athoracsur.2018.03.002

Abbreviations and Acronyms

ACSD = Adult Cardiac Surgery Database
AVR = aortic valve replacement
CABG = coronary artery bypass grafting surgery
DSWI = deep sternal wound infection/mediastinitis
O/E = observed to expected ratio
PLOS = postoperative length of stay
SES = socioeconomic status
SLS = significance level to stay
STS = The Society of Thoracic Surgeons

models, and a systematic explanation of the model development process. Part 2 [1] provides more extensive technical detail regarding the statistical methodology and results.

Genesis of STS Risk Models—Federal Government "Death Lists"

The modern era of transparency and accountability in health care began not with the Affordable Care Act but more than 2 decades earlier, in March 1986, with the publication of essentially unadjusted hospital-level mortality data by the Health Care Financing Administration, the predecessor of the Centers for Medicare and Medicaid Services. One of the specific, high-profile procedures targeted was coronary artery bypass grafting surgery (CABG), leading hospitals and surgeons to complain that the inherent risk of their patients was not being considered in these so-called Medicare "death lists."

STS Ad Hoc Committee on Risk Factors

In response, STS formed an Ad Hoc Committee on Risk Factors chaired by Dr Nick Kouchoukos. This committee issued a Statement of Concern in October of the same year, followed by a formal report strongly advocating for risk adjustment when profiling provider performance [2]. The authors of the report correctly noted that unadjusted rates were misleading to the public as they did not account for preoperative patient severity and acuity. They suggested that this might lead providers to avoid treating high-risk patients, anticipating the phenomenon we now refer to as risk aversion [3, 4]. They advised that risk models and performance measures should focus on relatively homogeneous procedure categories, such as isolated CABG. Combined procedures (eg, CABG plus aortic valve replacement [AVR]) should be considered in separate models as they have higher inherent risk. Finally, they presciently noted that risk-adjusted mortality is an important but inadequate metric by which to assess performance; comprehensive performance measurement should also include risk-adjusted rates of other complications such as reoperation or stroke. This concept was ultimately realized in the development of STS composite performance measures [5–11]. Thirty years later, these principles remain fundamental tenets of all STS risk models and performance measures.

Risk Models Require Optimal Data—Origins of STS National Database

A prerequisite for the development of early STS risk adjustment models was the availability of clinically granular, standardized data to characterize preoperative patient comorbidities. Because such data were unavailable except for idiosyncratic institutional registries or research datasets for one-time studies, it was recognized that a national, clinical data registry for cardiothoracic surgery was needed. This was the proximate stimulus for development of the STS National Database in 1989 [12].

Early STS Risk Models

During the same period in the late 1980s, Dr Fred Edwards had begun to explore health care applications of Bayesian techniques, initially for diagnostic evaluation and soon thereafter for risk prediction in cardiac surgery. Initial studies were limited to single institution data, but in 1994, Edwards, Clark, and Schwartz published their landmark article on the application of these Bayesian approaches to cardiac surgery risk adjustment, using data from the nascent STS National Database [13]. Subsequently, logistic risk models were introduced and have been used in subsequent risk model iterations [13–21].

Expanding Family of STS Risk Models

This portfolio of risk models has expanded from its original focus on one procedure (CABG) and one outcome (mortality) to include other major cardiothoracic procedures—valve replacement, congenital heart surgery, and general thoracic surgery—as well as other endpoints such as complications and length of stay. These models have been used to provide risk adjustment for a growing family of STS performance metrics, including composite measures [5–11], many of which are publicly reported [22, 23].

Each adult cardiac surgery risk model is published in the peer-reviewed literature, and relevant intercepts and coefficients are publicly available either in peer-reviewed journals or on the STS or Duke Clinical Research Institute websites. As part of each data harvest, each model is recalibrated so that the expected number of events exactly matches the observed number during that harvest, resulting in an observed to expected (O/E) ratio of 1. Periodically these models are completely revised, as described in this report, to include new risk factors and to reflect changes in cardiothoracic practice and outcomes.

Risk Model Applications

Risk models have many potential applications. For example, they define a standard default set of covariates for STS research analyses, they provide risk scores for real-time patient counseling and shared decision making, and they can inform and document performance improvement initiatives. However, the primary motivation in developing the current risk models was to provide

robust case-mix adjustment for estimating risk-adjusted outcomes, which are subsequently used in STS feedback reports and voluntary public reporting. Although parsimony and ease of use were secondary considerations, predictive accuracy was the paramount goal that guided model development.

Appropriate Interpretation of Risk-Adjusted Outcomes

The proper interpretation of risk-adjusted outcomes rates or O/E ratios is crucial [24]. Because of technical considerations, including the large number of clinical covariates compared with relatively few strata of interest in most epidemiologic studies, health care risk models such as those used by the STS almost always use indirect rather than direct standardization. In direct standardization, rates from various strata (eg, age) within the study population of interest (eg, in the case of profiling, a specific hospital) are applied to a reference or standard population. That may allow direct comparisons of the standardized rates from different study populations, as they have all been applied to the same standard population.

In indirect standardization, the reverse process is carried out. Rates derived from the benchmark population for characteristics of interest are applied to the study population (in the case of profiling, a health care provider's patients) to obtain their so-called expected outcomes or rates. Each hospital's indirectly standardized rate of an endpoint or their O/E ratio are based only on the patients for whom it cares, and its case mix may be quite different than that of another specific hospital. There may be patients at one hospital for whom there is no analogue at another hospital.

Because of these differences, it would be inappropriate to compare indirectly standardized rates or ratios between specific providers. For example, consider one provider caring mainly for low-risk patients and another caring predominately for high-risk patients. The indirectly standardized outcomes at the former may not include any of the high risk-patients seen at the latter. An O/E ratio of 0.8 achieved by a hospital caring for low-risk patients gives no assurance that this hospital could achieve similar results if confronted with higher risk patients. These concepts have critical implications for proper interpretation, yet they are often misunderstood or ignored [24].

It is most appropriate to interpret indirectly standardized outcomes or O/E ratios, including those from STS risk models, as representing a provider's actual results for their specific patients compared with what would have been expected (technically referred to as the counterfactual outcomes) for the same patients based on the performance of all providers who contributed to the benchmark population. The STS and others often classify these results as worse than expected, as expected, or better than expected performance.

For the same reason, rating of providers (eg, better or worse than expected or as expected) compared with a national benchmark is appropriate, whereas rankings (no. 1, no. 2, and so forth, which implies that hospital 1 is better than hospital 2) are not.

Considerations in Risk Model Development

This 2018 complete update of STS adult cardiac surgery risk models by the STS Quality Measurement Task Force, spanning several years of work by surgeons and statisticians, was based on a number of important considerations.

Why Risk Adjust?

Although the need for risk models may seem axiomatic, it is worth remembering that some health care performance measures still lack adequate risk adjustment. Outcomes measures should be risk adjusted whenever there are important patient characteristics that significantly affect the outcome of interest, and when the prevalence of such factors varies across providers. Both these conditions are satisfied in cardiac surgery [25]. Robust risk adjustment also enhances provider acceptance of outcomes measures and helps to mitigate risk aversion.

Target Procedures

Selection of a target procedure or population for risk-adjusted outcomes involves a tradeoff between adequate sample size (which might argue for broader procedure categories) and clinically coherent, more homogeneous cohorts, which have greater face validity and specificity [25]. Because of the large numbers of all procedures available in the STS ACSD, the Quality Measurement Task Force created separate risk models for the most commonly performed adult cardiac surgical procedures: isolated CABG, isolated AVR, isolated mitral valve repair or replacement, AVR plus CABG, and mitral valve repair or replacement plus CABG.

Endpoint Selection

The nine outcomes we studied—operative mortality, stroke, renal failure, prolonged ventilation, reoperation, mediastinitis/deep sternal wound infection (DSWI), major morbidity or mortality composite, prolonged postoperative length of stay (PLOS), or short PLOS)—were chosen based on historical precedent, clinical impact, resource use, and inclusion in current performance metrics (eg, STS composite scores). We created separate risk models for each outcome and procedure except DSWI. Because its incidence is quite low (generally less than 0.5%), separate DSWI models for any one procedure are unreliable and subject to overfitting. Accordingly, as described in Part 2 of this report, we created a single model for DSWI that encompassed all the major procedures, with an indicator variable for the specific procedure. With the potential exception of CABG using bilateral internal thoracic arteries, the major DSWI risk factors (eg, diabetes mellitus, obesity, immunosuppression, severe chronic lung disease) were not expected to vary dramatically across procedures.

Socioeconomic Indicators

Whether outcomes measures, and the public reporting and reimbursement programs based on them, should consider socioeconomic status (SES) or sociodemographic

factors (eg, race, ethnicity, education, income, payer [eg, Medicare-Medicaid dual eligible status]) is a topic of intense health policy debate [26]. Some argue that in the absence of adjustment for these variables, the outcomes of hospitals that care for a disproportionate percentage of low SES patients will be unfairly disadvantaged, perhaps leading to financial or reputational penalties. Opponents argue that inclusion of SES factors in risk models may "adjust away" disparities in quality of care, and they advocate the use of stratified analyses instead. Also, readily available SES factors have often not demonstrated significant impact on outcomes, perhaps because they are not sufficiently granular or relevant. Finally, even SES proponents agree that these factors make more sense conceptually for some outcomes (eg, readmission) than for others (hospital mortality, complications). Notably, as part of an National Quality Forum pilot project, the STS specifically studied dual eligible status in the STS readmission measure [27] and found minimal impact.

In developing the new STS risk models, we avoided these more philosophical and health policy arguments regarding SES adjustment and based our modeling decisions on empiric findings and consideration of the model's primary intended purpose—optimal case mix adjustment. Conceptually, our goal was to adjust for all preoperative factors that are independently and significantly associated with outcomes and that vary across STS participants. For example, race will continue to be in our risk models as it has been previously, but not conceptually as a SES indicator. Race has an empiric association with outcomes and has the potential to confound the interpretation of a hospital's outcomes, although we do not know the underlying mechanism (eg, genetic factors, differential effectiveness of certain medications, rates of certain associated diseases such as diabetes and hypertension, and potentially SES for some outcomes such as readmission).

Interaction Terms

Consistent with previous STS risk models, to facilitate interpretability we elected to focus on main effects and did not include many interaction terms. In general, we allowed the effect of each patient factor to differ depending on the type of operation but not on other patient factors.

Parsimonious Versus Nonparsimonious Models

A fundamental decision in risk modeling is how many risk variables will be included. There are cogent arguments for both parsimonious and more expansive models. Most of the predictive power of risk models is contained in a relatively few important covariates [28, 29], and addition of more variables usually does not substantially change the c-index or area under the receiver-operating characteristics curve, a common (although not necessarily the best) measure of model performance. One often-used rule of thumb proposed by statistician Frank Harrell is that at least 10 endpoints are required for each variable in a model [30].

Parsimonious models are computationally simpler, faster, and more likely to converge. They are easier to run in production mode, easier for software vendors when changes or updates are implemented, and easier for practicing clinicians to use when providing patients with their estimated risk of surgery (ie, they only need to enter a small number of variables into the STS online risk calculator). An excessive number of predictors (over-parameterized model) for the number of available endpoints can lead to unstable, noisy estimates in which different random samples from the same population could produce markedly different results. They may overfit the results to the specific data used for model development but the model may not generalize well to other or subsequent populations. Finally, although our risk models are built using 3 years of data, these models will typically be utilized with smaller samples, often 1 year of data. A highly parameterized model might run adequately in the development set, but not in production mode with smaller numbers of patients.

However, there are also disadvantages to highly parsimonious models, including face validity. These models, which sometimes include only a few variables, may have reasonable overall performance at the population level. However, they necessarily exclude variables that, although uncommon, are highly predictive of adverse outcomes when present, such as severe liver disease. When these risk factors are present, the predictive accuracy of highly parsimonious models may be compromised, which would violate the guiding principle we followed in developing these models. Specifically, the risk of patients with high-impact features not adjusted for in the model will be underestimated. Failure to adjust for such high-risk characteristics could lead clinicians to avoid caring for patients with these risk factors, as their severity would not be accounted for in their risk-adjusted outcomes [3, 4].

We have opted for a middle ground: model building using backward selection from a full model, a process during which we attempted to optimize both clinical face validity and statistical performance. Using the process described below, surgeons and statisticians selected the most parsimonious models possible that did not exclude any clinically critical variables or significantly compromise predictive accuracy.

The 2018 STS Adult Cardiac Surgery Risk Models—General Principles

STS Database Version Mapping

The 2008 STS ACSD risk models were developed using data from STS versions 2.35, 2.41, and 2.52 and were designed for use with the concomitantly released version 2.61. The new 2018 risk models were developed using data from version 2.73 and tested using version 2.81. To be a viable candidate for the new 2018 models, a variable had to exist in version 2.73 and that same variable or a closely related, mappable analogue needed to be present in version 2.81. This was particularly challenging for

certain variables, such as preoperative atrial fibrillation, in which the categories or choices for specific variables (in technical terms, the parameterization) has changed between versions.

Exploratory Analyses of Candidate Variables

After defining the source data for the new risk models and performing any required mapping of variables across data versions, initial exploratory analyses of all potential candidate variables (those included in the earlier 2008 risk models plus new variables added in STS ACSD version 2.73) was performed. For each candidate variable, procedure, and outcome, we examined the percentage of patients in each group of a categorical variable; the mean and median value for continuous variables; the percentage missing data (or test not performed); the bivariate associations of the candidate variable with outcomes; and reliability estimates for full and reduced models for various outcomes.

Certain variables were retained as candidates for some scenarios but not others. For example, the use of angiotensin-converting enzyme inhibitors and angiotensin receptor-blocking agents may lead to postoperative renal dysfunction or vasoplegia, but continuing or discontinuing these agents preoperatively is based on surgeon judgment, making it a suboptimal variable for case-mix adjustment in elective cases. Conversely, surgeons may not have the option to consider discontinuing these agents for urgent or emergent cases, and they then become reasonable candidate variables for those scenarios.

Other risk factors known to effect individual patient outcomes, such as pulmonary hypertension, were, after extensive discussions, not included as candidate variables owing to temporal inconsistencies of parameter measurement (interventional laboratory versus operating room), potential acute alterations with intravenous fluid or drug therapy, and a high proportion of missing data.

Missing Data

The frequency of missing data was less than 1% for most preprocedural variables. We excluded from further consideration those few variables with missing data rates greater than 5%, or variables reflecting a test or study that had not been performed in more than 5% of the relevant study population. Examples of excess missing data precluding their use in modeling included bilirubin (missing 20%) and international normalized ratio (missing 8%), which prevented modeling of the Model for End-Stage Liver Disease score; hemoglobin A1c (missing 21%), an important marker of diabetes; etiology of valvular disease (missing in more than 10% in each valve population); and 5-m walk test (missing or not performed in 95% of patients). Previous studies have shown that the latter, when abnormal (greater than 6 seconds), increases risk twofold to threefold [31, 32]. This information regarding excluded variables is an important reminder to STS ACSD participants that complete data are essential to optimize risk model development.

Imputation strategies for the initial exploratory and variable selection analyses, and for reestimation of covariate regression coefficients in the final model, are discussed in Part 2 of this report.

Feasible Number of Candidate Predictors

Initial review identified more than 50 preoperative variables (and more than 100 potential parameters) for possible inclusion in the various models. To empirically assess the theoretical limitations posed by including such a large number of candidate variables (and their associated subcategories), we used bootstrap simulations to estimate a measure of signal-to-noise ratio reliability [33]. Here, signal refers to the amount of variation in risk across patients, whereas noise refers to the amount of error in the estimation of each patient's risk. We reasoned that the amount of acceptable noise depends in part on the magnitude of signal. For example, an average estimation error of ± 1 percentage point may be acceptable if true risk ranges from 0% to 50% across patients but unacceptable if true risk ranges from 0% to 2%.

A complete description of the methods and results of these calculations are presented in Part 2 of this report. Briefly, we found that the average estimation error was lower for CABG models than for valve or valve plus CABG models, mainly because of larger sample sizes. The most striking finding was that estimation error was consistently quite high for the sternal infection models for valve and valve plus CABG, as this is the least common endpoint (average prevalence 0.3%). That led to a decision to combine sternal infection results for all procedures into one model to increase the effective number of endpoints, as described above.

Surgeon Perspectives

Development of the new STS risk models was largely data driven and avoided forced, subjective inclusion or exclusion of variables. However, in addition to their active participation in the evaluation of the data analytics, surgeons did provide clinical insights at various stages of the development process. For example, as part of the initial exploratory analyses, surgeon members of the Quality Measurement Task Force assessed each variable's clinical importance for inclusion in the risk model, assigning it a rating of 1 to 10. Some variables (use of tobacco products other than cigarettes; dyslipidemia) were thought a priori to have little clinical relevance, had not been included in previous models, or demonstrated minimal association with outcomes in bivariate analyses. Given that the total number of variables we could include was limited, these variables were not considered further. Conversely, as described subsequently, some variables were considered so important for face validity that the p value for backward selection was largely based on the ability to retain these variables.

Optimal Coding Efficiency

Before final model selection, we determined the most efficient coding, or parameterization, of candidate variables, typically by collapsing or combining clinically

related or collinear variables, which often had similar magnitude associations with specific endpoints. Some variables had computationally excessive categories, or the categories had changed between versions, such as preoperative myocardial infarction, arrhythmias, and payer. In other instances, uncommon but important variables were combined with other related variables—for example, our decision to code catheter support devices and extracorporeal membrane oxygenation as shock, because that is their usual indication. Race and ethnicities have multiple categories and potential combinations—we chose to parameterize these as we had in the past, with six major categories. Similarly, although there are many payer options available for coding in the database, for modeling purposes we reduced these to a limited number of broad, coherent categories.

Other coding issues were similarly challenging. For example, body mass index and body surface area measure slightly different characteristics of body habitus and shape although both use the same height and weight inputs. After considerable discussion, both body mass index and body surface area were included as candidate variables, subject to further winnowing during the backward selection process.

For continuous variables, additional exploratory analyses helped determine how best to model the association of the variable with specific outcomes (eg, linear, quadratic, polynomial, splines).

Competing Risks

All nonfatal endpoints and complication analyses have the potential for bias due to the competing risk of death. If a patient dies on day 1 postoperatively, that patient does not have the opportunity to have other complications such as prolonged ventilation. Fortunately, the competing risk—mortality—occurs relatively uncommonly compared with most nonfatal endpoints. In most instances, risk models for complications of medical or surgical treatment have not considered the competing risk issue, other than to point out that the nonfatal endpoint (eg, readmission [34]) and mortality should both be examined or even combined in a composite, such as risk-adjusted mortality and morbidity in all STS adult cardiac surgery composite measures [5–11].

For several outcomes in the new risk models, the issue of competing risk was extensively discussed. For example, what patients should be included in the numerator and denominator of the short PLOS measure? In some instances, short PLOS may result from early death, so it could be argued those patients who die less than 6 days after surgery should be excluded from the denominator of this measure—in other words, they should be ineligible to receive credit for this measure as short PLOS is viewed as a favorable outcome. However, with a smaller denominator, the proportion of that hospital's short PLOS patients will appear larger (better), in effect "rewarding" hospitals for early mortality. Weighing the pros and cons of the various approaches, we decided that it was best to retain all patients in the denominator, irrespective of why they are discharged early, but not give numerator credit for a short PLOS to those patients who died at less than 6 days.

The prolonged PLOS (more than 14 days) measure presents similar issues, as it effectively gives credit for patients who die less than 14 days postoperatively. For example, a patient who dies on day 8 postoperatively will not be in the numerator, as their length of stay is not more than 14 days, which would appear to be favorable for the hospital. One option would be to change the measure numerator statement to both more than 14 days and discharged alive, but that would change the meaning of a longstanding STS measure and might be misleading for patients and other stakeholders. For example, a hospital might have a very low (favorable) prolonged PLOS because patients are efficiently managed and rarely require longer stays; alternatively, prolonged PLOS might be low because more of their patients die in hospital and therefore do not meet the numerator criteria. If, however, inhospital deaths are removed from the denominator, the apparent proportion of long hospital stays will increase. Therefore, a site that has better salvage rate from serious complications (lower failure to rescue) will erroneously appear to be worse performing (ie, prolonged PLOS numerator will be relatively larger in proportion to the reduced denominator).

After weighing all these possibilities, we adopted the following definitions. Short PLOS is discharge alive within 6 days of surgery—all patients are in the denominator but there is no numerator credit for patients who die less than 6 days after surgery. The prolonged PLOS measure denominator encompasses all patients, including those who died less than 14 days after surgery; all patients who are discharged at more than 14 days are included in the numerator, including nonsurvivors. These PLOS outcomes are primarily measures of resource use and efficiency and should always be viewed in combination with clinical balancing measures (in this case, mortality) or as part of a composite measure.

Final Model Selection

After candidate covariates had been chosen for each combination of population and endpoint, we systematically selected final model variables and estimated their associated coefficients. Separate, comprehensive analyses were considered for each major procedure group (CABG, valve, valve plus CABG). Backward selection was used for model development, as it is computationally efficient and provides the ability to compare the results of full and reduced (more parsimonious) models.

When selecting the final model, determining the optimal significance level to stay (SLS) in backward selection was the primary focus, as discussed in detail in Part 2 of this report. Historically, this has typically been decided a priori, choosing a "standard" level such as $p = 0.05$. Rather than using such a predetermined, identical value for each procedure and endpoint, we elected to make these decisions based on empirical data and face validity. Surgeons and statisticians first examined the empirical data (see Part 2) for each SLS: a variety of

statistical tests of model performance; internal and external (cross validated) calibration plots for the overall population cohort and selected subgroups; and the number of variables and parameters (including categories of variables) retained at each level.

To reduce the number of model comparisons, the SLS evaluation process was limited to six possible values: SLS = 0.0001, SLS = 0.001, SLS = 0.01, SLS = 0.05, SLS = 0.1, and SLS = 1, where SLS = 1 means that all candidate covariates are retained (ie, the full model, no backward selection).

For each population, endpoint, and SLS value, the following information was reviewed by the modeling committee, and will be discussed in more detail in Part 2 of this article:

- Optimism-adjusted c-statistic
- Optimism-adjusted slope
- Calibration plots of observed versus expected outcome by decile of predicted risk, overall and for clinically important subgroups, using ninefold cross validation
- List of variables selected and their associated parameter estimates
- Estimated odds ratios

The purpose of this information was to determine whether there were compelling statistical differences between SLS levels to support one specific choice, which in most cases we did not find. Then, surgeons examined both the full model and the variables retained at each successively more parsimonious (smaller p values) level of SLS. The goal was to choose the SLS that produced the most parsimonious model while not eliminating any variables that were thought to be important for face validity. In some instances, variables might be retained in several models with different SLS p values, but their specific coding was slightly different (eg, diabetes requiring insulin control as a separate variable, versus diabetes requiring either oral agents or insulin).

The following examples demonstrate the decision-making process used to determine the final models:

1. CABG—renal failure: SLS $p = 0.1$ selected because $p = 0.05$ resulted in loss of angiotensin-converting enzyme/angiotensin-receptor blocker in urgent/emergent patients, 2b/3a platelet inhibitors, and preoperative inotropic support variables
2. CABG—prolonged ventilation: SLS $p = 0.1$ selected because SLS $p = 0.05$ would have resulted in loss of recent cerebrovascular accident, history of mediastinal radiation, and coexisting severe mitral or aortic insufficiency variables
3. Valve procedures—stroke: SLS $p = 0.1$ selected because SLS $p = 0.05$ resulted in loss of hematocrit as a predictor variable
4. Valve procedures—prolonged ventilation: SLS $p = 0.1$ selected because SLS $p = 0.05$ resulted in omission of mitral stenosis variable
5. Valve procedures—renal failure in valve models: SLS $p = 0.05$ selected because SLS $p = 0.01$ resulted in loss of emergent/emergent salvage, moderate/severe aortic insufficiency, severe tricuspid insufficiency, and New York Heart Association class IV heart failure variables

In some instances, practical considerations overweighed what clinicians might have preferred. For example, for the combined DSWI risk model, the full model (SLS $p = 1$) contained 63 patient risk factors and 222 parameters. That was clearly too many for the number of available endpoints and would almost certainly lead to nonconvergence or "noisy" models. It was therefore necessary to select a model with SLS $p = 0.1$, which reduced the number of variables and parameters to 25 and 63, respectively. In doing so, however, some clinically relevant risk factors were lost from the model, including mediastinal radiation, steroid use (although immunosuppression was retained), home oxygen, and liver disease.

Further detailed descriptions of the statistical approaches used in developing these models, including discrimination and calibration, are provided in Part 2 of this report.

Conclusion

Comprehensive new STS risk models have been developed based on the most contemporary data available, with the primary goal of optimizing predictive accuracy for case-mix adjustment. A structured, standardized approach to model development considered previous STS models, missing data percentages, available sample size, number of endpoints, and association of variables with endpoints. Final model development used backward selection to estimate the most parsimonious model that retained clinically essential risk factors and achieved the desired statistical performance. This strategy provided the highest likelihood of model convergence in production mode and generalizability to future data. "Big data" and machine learning approaches may some day further advance the science of risk modeling by incorporating heretofore unrecognized patterns and associations.

References

1. O'Brien SM, Feng L, He X, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery risk models: part 2—statistical methods and results. Ann Thorac Surg 2018;105:1419–28.
2. Kouchoukos NT, Ebert PA, Grover FL, Lindesmith GG. Report of the ad hoc Committee on Risk Factors for Coronary Artery Bypass Surgery. Ann Thorac Surg 1988;45:348–9.
3. Shahian DM, Jacobs JP, Badhwar V, D'Agostino RS, Bavaria JE, Prager RL. Risk aversion and public reporting. Part 1: observations from cardiac surgery and interventional cardiology. Ann Thorac Surg 2017;104:2093–101.
4. Shahian DM, Jacobs JP, Badhwar V, D'Agostino RS, Bavaria JE, Prager RL. Risk aversion and public reporting. Part 2: mitigation strategies. Ann Thorac Surg 2017;104:2102–10.
5. Shahian DM, Edwards FH, Ferraris VA, et al. Quality measurement in adult cardiac surgery: part 1—conceptual framework and measure selection. Ann Thorac Surg 2007;83(4 Suppl):S3–12.
6. O'Brien SM, Shahian DM, DeLong ER, et al. Quality measurement in adult cardiac surgery: part 2—statistical

considerations in composite measure scoring and provider rating. Ann Thorac Surg 2007;83(4 Suppl):S13–26.
7. Rankin JS, Badhwar V, He X, et al. The Society of Thoracic Surgeons mitral valve repair/replacement plus coronary artery bypass grafting composite score: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg 2017;103:1475–81.
8. Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons composite measure of individual surgeon performance for adult cardiac surgery: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg 2015;100:1315–25.
9. Badhwar V, Rankin JS, He X, et al. The Society of Thoracic Surgeons mitral repair/replacement composite score: a report of The Society of Thoracic Surgeons Quality Measurement Task Force. Ann Thorac Surg 2016;101:2265–71.
10. Shahian DM, He X, Jacobs JP, et al. The STS AVR+CABG composite score: a report of the STS Quality Measurement Task Force. Ann Thorac Surg 2014;97:1604–9.
11. Shahian DM, He X, Jacobs JP, et al. The Society of Thoracic Surgeons isolated aortic valve replacement (AVR) composite score: a report of the STS Quality Measurement Task Force. Ann Thorac Surg 2012;94:2166–71.
12. Clark RE. It is time for a national cardiothoracic surgical data base. Ann Thorac Surg 1989;48:755–6.
13. Edwards FH, Clark RE, Schwartz M. Coronary artery bypass grafting: The Society of Thoracic Surgeons National Database experience. Ann Thorac Surg 1994;57:12–9.
14. Edwards FH, Grover FL, Shroyer AL, Schwartz M, Bero J. The Society of Thoracic Surgeons National Cardiac Surgery Database: current risk assessment. Ann Thorac Surg 1997;63:903–8.
15. Shroyer AL, Grover FL, Edwards FH. 1995 coronary artery bypass risk model: The Society of Thoracic Surgeons Adult Cardiac National Database. Ann Thorac Surg 1998;65:879–84.
16. Shroyer AL, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: The Society of Thoracic Surgeons Adult Cardiac National Database. Ann Thorac Surg 1999;67:1205–8.
17. Edwards FH, Peterson ED, Coombs LP, et al. Prediction of operative mortality after valve replacement surgery. J Am Coll Cardiol 2001;37:885–92.
18. Shroyer AL, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. Ann Thorac Surg 2003;75:1856–64.
19. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 1—coronary artery bypass grafting surgery. Ann Thorac Surg 2009;88(1 Suppl):S2–22.
20. O'Brien SM, Shahian DM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 2—isolated valve surgery. Ann Thorac Surg 2009;88(1 Suppl):S23–42.
21. Shahian DM, O'Brien SM, Filardo G, et al. The Society of Thoracic Surgeons 2008 cardiac surgery risk models: part 3—valve plus coronary artery bypass grafting surgery. Ann Thorac Surg 2009;88(1 Suppl):S43–62.
22. Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 1—history, rationale, consequences. Ann Thorac Surg 2011;92(3 Suppl):S2–11.
23. Shahian DM, Edwards FH, Jacobs JP, et al. Public reporting of cardiac surgery performance: part 2—implementation. Ann Thorac Surg 2011;92(3 Suppl):S12–23.
24. Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. Circulation 2008;117:1955–63.
25. Shahian DM, He X, Jacobs JP, et al. Issues in quality measurement: target population, risk adjustment, and ratings. Ann Thorac Surg 2013;96:718–26.
26. National Academies of Sciences, Engineering, and Medicine. Accounting for Social Risk Factors in Medicare Payment. Washington, DC: National Academies Press; 2017.
27. Shahian DM, He X, O'Brien SM, et al. Development of a clinical registry-based 30-day readmission measure for coronary artery bypass grafting surgery. Circulation 2014;130:399–409.
28. Jones RH, Hannan EL, Hammermeister KE, et al. Identification of preoperative variables needed for risk adjustment of short-term mortality after coronary artery bypass graft surgery. The Working Group Panel on the Cooperative CABG Database Project. J Am Coll Cardiol 1996;28:1478–87.
29. Tu JV, Sykora K, Naylor CD. Assessing the outcomes of coronary artery bypass graft surgery: how many risk factors are enough? Steering Committee of the Cardiac Care Network of Ontario. J Am Coll Cardiol 1997;30:1317–23.
30. Harrell FE, Jr. Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. New York: Springer; 2015.
31. Afilalo J, Mottillo S, Eisenberg MJ, et al. Addition of frailty and disability to cardiac surgery risk scores identifies elderly patients at high risk of mortality or major morbidity. Circ Cardiovasc Qual Outcomes 2012;5:222–8.
32. Afilalo J, Eisenberg MJ, Morin JF, et al. Gait speed as an incremental predictor of mortality and major morbidity in elderly patients undergoing cardiac surgery. J Am Coll Cardiol 2010;56:1668–76.
33. Adams JL. The Reliability of Provider Profiling: A Tutorial. RAND Health, prepared for the National Committee for Quality Assurance. Santa Monica, CA: RAND Corporation; 2009.
34. Gorodeski EZ, Starling RC, Blackstone EH. Are all readmissions bad readmissions? N Engl J Med 2010;363:297–8.