# The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2—Statistical Methods and Results



Sean M. O'Brien, PhD, Liqi Feng, MS, Xia He, MS, Ying Xian, MD, PhD, 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, Fred H. Edwards, MD, and David M. Shahian, MD

Duke Clinical Research Institute, Duke University Medical Center, Durham, North Carolina (SMO, LF, XH, YX); 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); Department of Surgery, University of Florida, Gainesville, Florida (FHE); and Department of Surgery and Center for Quality and Safety, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts (DMS)

*Background.* **The Society of Thoracic Surgeons (STS) uses statistical models to create risk-adjusted performance metrics for Adult Cardiac Surgery Database (ACSD) participants. Because of temporal changes in patient characteristics and outcomes, evolution of surgical practice, and additional risk factors available in recent ACSD versions, completely new risk models have been developed.**

*Methods.* **Using July 2011 to June 2014 ACSD data, risk models were developed for operative mortality, stroke, renal failure, prolonged ventilation, mediastinitis/deep sternal wound infection, reoperation, major morbidity or mortality composite, prolonged postoperative length of stay, and short postoperative length of stay among patients who underwent isolated coronary artery bypass grafting surgery (n = 439,092), aortic or mitral valve surgery (n = 150,150), or combined valve plus coronary artery bypass grafting surgery (n = 81,588). Separate models were developed for each procedure and endpoint except mediastinitis/deep sternal wound infection, which was analyzed in a combined model because of its infrequency. A surgeon panel selected predictors by assessing model performance and clinical face validity of full and progressively more parsimonious models. The ACSD data (July 2014 to December 2016) were used to assess model calibration and to compare discrimination with previous STS risk models.**

*Results.* **Calibration in the validation sample was excellent for all models except mediastinitis/deep sternal wound infection, which slightly underestimated risk and will be recalibrated in feedback reports. The c-indices of new models exceeded those of the last published STS models for all populations and endpoints except stroke in valve patients.**

*Conclusions.* **New STS ACSD risk models have generally excellent calibration and discrimination and are well suited for risk adjustment of STS performance metrics.**

**(Ann Thorac Surg 2018;105:1419–28)**


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.

The Supplemental Material can be viewed in the online version of this article [https://doi.org/10.1016/j.athoracsur.2018.03.003] on http://www.annalsthoracicsurgery.org.



0003-4975/$36.00
https://doi.org/10.1016/j.athoracsur.2018.03.003

**Abbreviations and Acronyms**

ACSD = Adult Cardiac Surgery Database
AVR = aortic valve replacement
BMI = body mass index
BSA = body surface area
CABG = coronary artery bypass grafting surgery
DSWI = deep sternal wound infection/mediastinitis
MVR = mitral valve replacement
MVr = mitral valve repair
PLOS = postoperative length of stay
STS = The Society of Thoracic Surgeons

Risk models are used for multiple purposes in adult cardiac surgery including quality measurement, clinical practice improvement, voluntary public reporting, and research. These risk models are used by The Society of Thoracic Surgeons (STS) to benchmark participant outcomes in comparison with national aggregate data in STS feedback reports; to enable case-mix adjustment in the calculation of participant and individual surgeon composite performance measures; and to support the STS voluntary public reporting initiative. To maximize the validity of its performance metrics, the STS has developed a portfolio of risk models that are customized for specific procedure populations and that adjust for numerous patient preoperative factors.

The 2008 STS adult cardiac surgery risk models were based on data from 2002 to 2006 [1–3]. Since the publication of these models, a recalibration factor has been applied to each subsequent harvest period so that the ratio of observed to expected outcomes would equal 1. In the decade since publication of the 2008 risk models, successive Adult Cardiac Surgery Database (ACSD) versions have been introduced to account for temporal changes in procedures, patient populations, surgical practices, outcomes, and the identification of new risk factors. Using the most recent data version available at the time of this analysis, we sought to develop a completely new set of STS adult cardiac surgery risk models.

Part 1 of this report [4] provides a detailed background and conceptual framework for the risk model update and provides a high-level methodologic summary of the update process. In Part 2, we provide the detailed statistical methods and results.

## Patients and Methods

### *Endpoints*

Risk models were developed for the following nine endpoints chosen for consistency with prior STS risk models and current performance metrics (eg, STS composite scores): (1) operative mortality, defined in all STS databases as all deaths, regardless of cause, occurring during the hospitalization in which the operation was performed even if after 30 days (includes patients transferred to other acute care facilities), and all deaths, regardless of cause, occurring after discharge from the hospital but before the end of the 30th postoperative day; (2) stroke—an acute episode of focal or global neurologic dysfunction caused by brain, spinal cord, or retinal vascular injury as a result of hemorrhage or infarction in which the neurologic dysfunction lasts for more than 24 hours; (3) renal failure—a new requirement for dialysis or meeting the RIFLE (Risk, Injury, Failure, Loss of kidney function, and End-stage kidney disease) criteria based on creatinine levels or glomerular filtration rate [5]; (4) prolonged ventilation or reintubation—more than 24 hours; (5) mediastinitis/deep sternal wound infection (DSWI) occurring during the index hospitalization or within 30 days of operation; (6) reoperation for bleeding, tamponade, or any cardiac reason; (7) major morbidity or mortality—a composite defined as the occurrence of any one or more of the above endpoints; (8) prolonged postoperative length of stay (PLOS)—PLOS more than 14 days (alive or dead); and (9) short PLOS, defined as PLOS less than 6 days and patient alive at discharge. The follow-up period for endpoint definitions was from operation until the latter of hospital discharge or 30 days for mortality and DSWI and until hospital discharge for all other endpoints.

Endpoints with notable definition changes compared with the STS 2008 models included stroke (changed duration of symptoms from more than 72 hours to more than 24 hours), reoperation (changed from any reason to any cardiac reason), DSWI (added mediastinitis and included both inhospital and 30-day timeframe), and renal failure (definition changed to more closely align with RIFLE criteria [5]).

### *Study Cohort*

Models were developed and evaluated using data from July 1, 2011, to December 31, 2016, and were limited to the three major procedure populations that have been designated for outcomes reporting in the STS participant feedback report: (1) isolated CABG; (2) isolated valve; and (3) valve plus CABG. Data collected under STS version 2.73 (July 1, 2011, to June 30, 2014) were used to develop the models and perform a preliminary internal assessment of discrimination and calibration. Data collected under STS version 2.81 (July 1, 2014, to December 31, 2016) were used to assess model performance in a separate patient sample.

The valve cohort includes operations for aortic valve replacement (AVR), mitral valve replacement (MVR), and mitral valve repair (MVr). The valve plus CABG population includes AVR plus CABG, MVR plus CABG, and MVr plus CABG. Definitions of these populations are provided in the Supplemental Material. Briefly, each operation type includes patients undergoing a standalone operation and excludes planned major concomitant operations with a few exceptions, most notably, that concomitant tricuspid valve repair, surgical ablation for atrial fibrillation, or repair of atrial septal defect are allowed concomitantly with MVR or MVr in the valve and

valve plus CABG populations. Patients on dialysis preoperatively were excluded from models predicting new-onset postoperative renal failure.

Among 1,556,593 records for patients aged 18 to 110 years undergoing a cardiac operation at a participating site in the United States or Canada during the study period, 1,250,165 (80%) records met criteria for one of the three major procedure populations (Table 1) and were included in the development cohort (n = 670,830) or validation cohort (579,335).

### *Risk Models*

For each endpoint except DSWI, separate risk models were developed for each major procedure population (8 endpoints × 3 populations = 24 risk models). For DSWI, the low number of endpoint events in the valve (n = 244) and valve plus CABG (n = 285) populations prompted concern that models in these populations may prove to be inaccurate because of overfitting the data. To mitigate overfitting, we developed a single DSWI model combining all three procedure populations. The DSWI models used indicator variables to adjust for operation type (eg, AVR, MVr, MVR, and so forth) and included interaction terms to account for the importance of selected risk factors that differ across these operation types.

### *Selection of Candidate Predictor Variables*

The 2018 STS risk models were developed using data from version 2.73, but these models will be applied to patients entered into the STS ACSD using versions 2.81 and later. Accordingly, to be an acceptable candidate variable, it was necessary to assure that the variable was present in version 2.73 and in version 2.81 (or a similar, mappable analogue in the latter). Because the main goal of the models is to adjust for case mix, only preprocedural patient variables were considered for inclusion.

To begin the selection process, each surgeon member of the working group (n = 10) independently reviewed a list of 187 potentially relevant preprocedure factors from the v2.73 data collection form and used an online questionnaire to rate his or her a priori assessment of each variable's prognostic potential. Variables identified as potential risk factors by at least four of the 10 surgeons were retained for further consideration and were discussed in detail in a series of conference calls. To facilitate this discussion, each variable's frequency distribution and percentage of missing data were tabulated overall and across operation types.

Missing data frequency was less than 1% for the majority of preprocedural variables. Those few variables with missing data rates greater than 5%, or variables associated with a test or study that had not been performed in more than 5% of the relevant study population, were also excluded. Specific examples of excluded variables are described in Part 1 of this report [4].

Considerations regarding adjustment of outcomes measures for socioeconomic status or sociodemographic factors (eg, race, ethnicity, education, income, payer [eg, Medicare-Medicaid dual eligible status]) are discussed in detail in Part 1 of this report [4]. In general, we based our modeling decisions on principles from epidemiology and causal inference, evaluating those available socioeconomic status or sociodemographic risk factors potentially having an empirical association with outcomes and relevant to case-mix adjustment, while avoiding more philosophical considerations.

Surgery date was included as a candidate predictor to adjust for temporal trends in endpoint occurrence rates and detection rates across the 3-year development period. Risk calculators implementing these models will account for time trends by predicting risk standardized to a January 1, 2014, surgery date.

### *Simulations to Assess Statistical Precision and Overfitting*

When the number of predictors in a model is too large in relation to the available sample size, the estimated numerical coefficients are likely to be inaccurate because of overfitting the current study data [6]. Using a data-driven variable selection procedure can reduce the number of predictors in a model but may not mitigate overfitting because each predictor tested for inclusion in the model has the potential to be selected because of overfitting [7].

To assess the potential statistical accuracy of risk models based on this project's available sample size and candidate risk factors, a simulation study was conducted. This involved creating 200 bootstrap samples by sampling records with replacement from the overall development sample and using each bootstrap sample to estimate two models for each combination of model population and endpoint, based on the surgeon panel's proposed list of candidate predictors. The first model included the entire set of proposed candidate predictor variables, a so-called full model. The second model started with the same set of predictors but applied backward selection with a significance threshold of 0.05. Regression coefficients from each bootstrap sample were then used to calculate predicted risk estimates for each patient in the overall development sample. Ideally, in a setting of high statistical precision,

*Table 1. Sample Sizes for Model Development and Evaluation*

| Procedure | Development[a] | Validation[b] |
|---|---|---|
| Overall | 670,830 | 579,335 |
| CABG | 439,092 | 385,179 |
| Valve | 150,150 | 129,511 |
| AVR | 87,629 | 72,719 |
| MVR | 26,850 | 25,888 |
| MVr | 35,671 | 30,904 |
| Valve+CABG | 81,588 | 64,645 |
| AVR+CABG | 55,064 | 43,822 |
| MVR+CABG | 9,227 | 8,737 |
| MVr+CABG | 17,297 | 12,086 |

[a] July 2011 to June 2014. [b] July 2014 to December 2016.

AVR = aortic valve replacement; CABG = coronary artery bypass grafting surgery; MVR = mitral valve replacement; MVr = mitral valve repair.

the predicted risk estimate for the same patient and endpoint should not vary depending on which bootstrap sample was used to estimate regression coefficients.

To quantify estimation error, we estimated the average Pearson correlation between risk estimates for the same patient across all possible pairs of bootstrap samples. This was done separately for each combination of population (CABG, valve, valve plus CABG), endpoint, and modeling strategy (all predictors, backward selection). Results indicated that predicted risk estimates were generally stable with Pearson correlation coefficients greater than 0.90 for most population and endpoints combinations whether retaining all predictors or using backward selection.

Models with low consistency across bootstrap samples were those with relatively few endpoint events, including stroke in the valve and valve plus CABG populations (correlations 0.58 to 0.69) and DSWI in the valve and valve plus CABG populations (correlations 0.25 to 0.41). These findings led the model committee to consider both predictive accuracy and parsimony in the final selection of candidate predictors and to estimate a single combined model for DSWI, instead of separate DSWI models for each procedure population, as mentioned previously.

### *Optimal Coding of Candidate Covariates*

We attempted to achieve the most computationally efficient and clinically relevant coding, or parameterization, of candidate variables. That typically involved collapsing or combining clinically related or collinear variables, and was particularly important in cases where multiple STS variables relate to a single underlying clinical concept, for example, insurance status (12 variables), previous cardiac interventions (31 STS variables), and preoperative arrhythmias (7 STS variables). In some instances, uncommon but important variables were combined with other related variables (eg, catheter-based assist devices and extracorporeal membrane oxygenation were combined with shock, their usual indication).

For some variables, informal exploratory analyses using STS data from an earlier period (2007 to 2011) helped to determine the optimal modeling strategy. For example, we used data from 2007 to 2011 to explore how best to model body mass index (BMI) and body surface area (BSA) given that both variables describe aspects of a patient's body habitus and are highly correlated. For these investigations, we initially estimated a multivariable model for mortality that did not adjust for BSA or BMI but included all other preoperative factors from the published STS 2008 mortality model. After fitting this model to data from 2007 to 2011, we then compared observed versus predicted mortality rates across subgroups defined by categorizations of BSA and BMI as well as sex. The observed pattern of residuals indicated that BSA and BMI were both independently associated with mortality and that inclusion of both variables was needed to capture variation in the residuals. When the model was reestimated after including BSA but not BMI, the pattern of residuals indicated a U-shape relationship between BMI and mortality, leading to the inclusion of both linear and quadratic terms for BMI. We investigated the following approaches for BSA and BMI: BSA linear; BSA quadratic; interaction between BSA and sex; BMI linear; BMI quadratic; and BMI alternatives. In some instances, extreme values were truncated (eg, BMI values greater than 50 were mapped to 50).

Similar analyses were conducted to explore modeling issues for insurance status, race, myocardial infarction history, and history of prior procedures, and to explore the functional form of various other continuous variables.

### *Selection of Final Covariates*

After choosing the list of candidate covariates, the final set of covariates for each model were selected. For each combination of population and endpoint, we estimated a full model that included all candidate covariates and a set of reduced models that were chosen by backward variable selection. To estimate the optimal significance level for backward selection, we repeated the backward selection process using five different significance levels (0.0001, 0.001, 0.01, 0.05, and 0.1) and estimated performance metrics for the resulting models. The goal of this analysis was to select the optimum significance level to use for each combination of population and endpoint. Because of overfitting the data, model performance is likely to be overestimated when models are developed and naively tested in the same sample of data. To obtain approximately unbiased performance estimates, each full model and the backward selection process was repeated in 200 bootstrap samples drawn with replacement from the original development sample.

To assess overfitting, we applied estimated regression coefficients from the bootstrap sample to patients in the overall development sample and then entered each patient's calculated risk score (log-odds) into a univariable logistic regression model predicting the endpoint. The slope coefficient for the risk score in this model was interpreted as a measure of overfitting, with a slope of 1.0 indicating perfect calibration and a slope less than 1 indicating possible overfitting [6]. Discrimination was assessed by calculating the c-statistic (the area under the receiver-operating characteristics curve) and using a bootstrap adjustment to correct for optimism [7].

To assess calibration, the backward selection process was subsequently repeated using ninefold cross validation. For each cross-validation replicate, models were developed in an 8/9 training sample and evaluated for calibration in a 1/9 testing sample. Calibration was assessed graphically by plotting observed versus expected endpoint event rates across deciles of predicted risk among patients in each testing sample. That was done for the full model and for each significance level when using backward selection. The main objective of this exercise was to determine whether there were compelling statistical differences between significance levels to support one particular choice.

In the absence of compelling statistical differences between the performance of various models, the final model was chosen by surgeon members of the working group, as

described in Part 1 of this report [4]. Beginning with the full model, surgeons carefully reviewed the predictors in each model (full, and using backward selection criteria $p = 0.1$, 0.05, 0.01, 0.001, and 0.0001). Each progressively more parsimonious model was evaluated to be certain that no variables had been eliminated that would jeopardize clinical face validity. Generally, the most statistically parsimonious model that did not compromise clinical face validity was chosen as the final model.

### *Missing Data and Imputation Strategies*

Covariate data were missing in fewer than 5% of cases in each procedure population for all but one candidate covariate (aortic root abscess in AVR and AVR plus CABG; missing = 13%). Overall, 15% of records had missing or unknown mortality data for at least one component of the operative mortality definition. Rates of missing or unknown data were 0.06% for discharge mortality status and 15.0% for 30-day mortality status. Previous linkage of the STS ACSD to the Social Security Death Master File [8] reveals that capture of 30-day deaths occurring before discharge is highly accurate, and that these inhospital deaths represent the majority (79%) of all 30-day deaths. Capture of the remaining 30-day deaths occurring after discharge was less complete and warranted improvement. Consequently, in 2016, the STS implemented more stringent requirements for all data fields related to operative mortality. As of January 1, 2016, participants were not included in the benchmark population for STS performance metrics, nor were these participants eligible to receive an STS star rating unless their rate of missing data for 30-day mortality and discharge mortality was less than 5% missing or unknown; in January 2017 this threshold was further decreased to 2%.

Missing data rates for endpoints other than mortality were less than 0.25%. For initial exploratory and variable selection analyses, missing covariate and endpoint values were handled using a simple single imputation strategy. Values were imputed to the most common category of binary or categorical variables and to the median or subgroup-specific median of continuous variables. This single imputation strategy was previously validated for the 2008 STS risk models by demonstrating that coefficients and predicted risk estimates obtained using single imputation were similar to the gold standard of multiple imputation [1].

After finalizing the selection of model covariates, as described above, regression coefficients were subsequently reestimated using a multiple imputation strategy for covariates with more than 5% missing data and for all endpoints. The principle motivation for using multiple imputation was to make efficient use of data from the discharge mortality status field when imputing operative mortality status among patients who were discharged alive. Multiple imputation was implemented using the method of chained equations as implemented in the SAS software (SAS Institute, Cary, NC) PROC MI procedure with the full conditional specification option [9, 10]. To avoid bias due to perfect prediction [11], separate imputation models were estimated for discharge deaths and discharge survivors. To speed computation and resolve convergence errors, covariates with less than 5% missing data were imputed by single imputation before estimating the multiple imputation model.

### *Final Model Assessment*

The validation sample was created by applying the study's inclusion criteria to STS data for the period July 1, 2014, to December 31, 2016, as the goal was to assess model performance in future data. Data from hospitals with more than 5% missing data for an endpoint within a procedure population were excluded from validation analyses for that population and endpoint. Discrimination was quantified

*Table 2. Percentage and Number of Endpoint Events by Model Population in Development Sample*

| Endpoint Events | All (n = 670,830) | CABG (n = 439,092) | Valve (n = 150,150) | Valve + CABG (n = 81,588) |
|---|---|---|---|---|
| Operative mortality | 2.9%<br>16,792/569,998 | 2.4%<br>8,852/373,683 | 3.2%<br>4,004/126,204 | 5.6%<br>3,936/70,111 |
| Stroke | 1.5%<br>9,866/669,561 | 1.3%<br>5,621/438,385 | 1.5%<br>2,237/149,800 | 2.5%<br>2,008/81,376 |
| Renal failure | 2.7%<br>17,202/648,808 | 2.2%<br>9,381/424,888 | 2.7%<br>3,868/145,454 | 5.0%<br>3,953/78,466 |
| Prolonged ventilation | 10.9%<br>72,984/670,830 | 9.3%<br>40,974/439,092 | 11.1%<br>16,604/150,150 | 18.9%<br>15,406/81,588 |
| Reoperation | 3.1%<br>20,872/670,778 | 2.4%<br>10,327/439,060 | 4.2%<br>6,371/150,137 | 5.1%<br>4,174/81,581 |
| Composite morbidity and mortality | 17.4%<br>101,180/581,976 | 15.0%<br>56,984/380,491 | 18.4%<br>23,724/129,140 | 28.3%<br>20,472/72,345 |
| Prolonged PLOS | 6.6%<br>44,533/670,428 | 5.0%<br>22,091/438,867 | 8.0%<br>11,941/150,024 | 12.9%<br>10,501/81,537 |
| Short PLOS | 42.7%<br>286,362/670,428 | 48.3%<br>211,820/438,867 | 37.4%<br>56,130/150,024 | 22.6%<br>18,412/81,537 |
| DSWI | 0.3%<br>1,875/669,392 | 0.3%<br>1,346/438,270 | 0.2%<br>244/149,778 | 0.4%<br>285/81,344 |

CABG = coronary artery bypass grafting surgery; DSWI = mediastinitis/deep sternal wound infection; PLOS = postoperative length of stay.

*Table 3. Candidate Predictors*

| | |
|---|---|
| Operation type | Illicit drug use |
| Age | Alcohol consumption (drinks per week) |
| Ejection fraction | Recent pneumonia |
| Body mass index | Mediastinal radiation |
| Body surface area | Cancer diagnosis within 5 years |
| Sex | Diabetes/diabetes control method |
| Renal function (dialysis/creatinine) | Number of diseased vessels |
| Hematocrit | Myocardial infarction history/timing |
| White blood cell count | Cardiac presentation on admission |
| Platelet count | Race/ethnicity |
| ADP receptor inhibitor usage/timing of discontinuation | Status |
| Hypertension | ACE/ARB inhibitor within 48 hours in nonelective operation |
| Immunosuppressive therapy within 30 days | Heart failure class and timing |
| Steroids within 24 hours | Recent smoker/timing |
| Glycoprotein IIb/IIIa inhibitor within 24 hours | Family history of CAD |
| Inotropes within 48 hours | Home oxygen |
| Preoperative IABP | Sleep apnea |
| Shock/ECMO/CBA | Liver disease |
| PAD | Unresponsive neurologic status |
| Left main disease | Syncope |
| Proximal LAD | Previous CABG |
| Aortic root abscess in AVR/AVR+CABG | Previous aortic valve procedure |
| Mitral stenosis | Previous mitral valve procedure |
| Aortic stenosis | Previous transcatheter valve replacement/percutaneous valve repair |
| Mitral insufficiency | Previous other valve procedure |
| Tricuspid insufficiency | Number of previous cardiovascular surgeries |
| Aortic insufficiency | Previous ICD |
| Arrhythmia and type | PCI history/timing |
| Endocarditis | Previous any other cardiac intervention |
| Chronic lung disease | Payer/insurance type |
| CVD/CVA/TIA | Tricuspid valve repair performed concomitantly |
| Carotid stenosis | Time trend (surgery date) |
| Previous carotid surgery | |

ACE = angiotensin-converting enzyme; ADP = adenosine diphosphate; ARB = angiotensin-receptor blocker; AVR = aortic valve replacement; CABG = coronary artery bypass grafting surgery; CAD = coronary artery disease; CBA = catheterization-based assist device; CVA = cerebrovascular accident; CVD = cardiovascular disease; ECMO = extracorporeal membrane oxygenation; IABP = intraaortic balloon pump; ICD = implantable cardioverter-defibrillator; LAD = left anterior descending artery; PAD = peripheral arterial disease; PCI = percutaneous coronary intervention; TIA = transient ischemic attack.

by the c-statistic. To provide context for interpreting discrimination results, c-statistics were calculated in the validation sample for both the current STS 2018 models and the prior STS 2008 models. Calibration was assessed by plotting observed versus expected event rates across deciles of predicted risk in the validation sample.

## Results

A total of 670,830 records met study inclusion criteria and were included in the development samples for CABG (n = 439,092), valve (n = 150,150), and valve plus CABG (n = 81,588). The number of endpoint events in the development sample ranged from 1,875 for DSWI to

*Table 4. C-statistics in Validation Sample for 2008 STS Risk Models and Current 2018 Risk Models*

| | CABG | | Valve | | Valve + CABG | |
|---|---|---|---|---|---|---|
| Endpoint | STS 2008 Models | STS 2018 Models | STS 2008 Models | STS 2018 Models | STS 2008 Models | STS 2018 Models |
| Operative mortality | 0.791 | 0.804 | 0.760 | 0.775 | 0.753 | 0.761 |
| Stroke | 0.682 | 0.697 | 0.669 | 0.656 | 0.631 | 0.632 |
| Renal failure | 0.801 | 0.826 | 0.770 | 0.787 | 0.746 | 0.759 |
| Prolonged ventilation | 0.756 | 0.772 | 0.761 | 0.777 | 0.731 | 0.744 |
| Reoperation | 0.600 | 0.621 | 0.604 | 0.616 | 0.577 | 0.588 |
| Composite morbidity and mortality | 0.725 | 0.738 | 0.712 | 0.723 | 0.702 | 0.712 |
| Prolonged PLOS | 0.761 | 0.777 | 0.778 | 0.796 | 0.726 | 0.739 |
| Short PLOS | 0.707 | 0.716 | 0.723 | 0.732 | 0.716 | 0.726 |
| DSWI | 0.665 | 0.681 | 0.592 | 0.665 | 0.648 | 0.659 |

CABG = coronary artery bypass grafting surgery; DSWI = mediastinitis/deep sternal wound infection; PLOS = postoperative length of stay; STS = The Society of Thoracic Surgeons.

286,362 for short PLOS (Table 2). As discussed above, the relatively small number of DSWI endpoints in valve (n = 244) and valve plus CABG (n = 285) populations raised concerns about potential overfitting in these populations, and that led to a decision to estimate a single combined model for DSWI. For the other eight endpoints, the number of occurrences ranged from 2,008 for stroke in valve plus CABG to 211,820 for short PLOS in CABG.

Table 3 summarizes the final list of candidate covariates. These 65 variables were included in the "full" model for each endpoint and population and were the starting point for variable selection by backward selection with bootstrapping and cross validation, and subsequent clinical assessment by the surgeon panel. Details of how each candidate variable was parameterized in the model are provided in the Supplemental Material. After inclusion of nonlinear, categorical, and interaction terms, the number of model parameters in the full model was 122 for CABG, 218 for valve, and 215 for valve plus CABG. The number of endpoint events per parameter in the full model ranged from 9 in the stroke model for valve plus CABG to 1,736 in the short PLOS model for CABG.

Supplemental Tables 1 to 4 in the Supplemental Material summarize risk factors in the final selected model for each population and endpoint. The number of risk factors in these models ranged from 25 in the model for stroke in valve plus CABG to 50 in the models for composite mortality or major morbidity and short PLOS in CABG. Full specifications for these models including formulas, coefficients, and intercept parameters will be publicly available from the STS website.

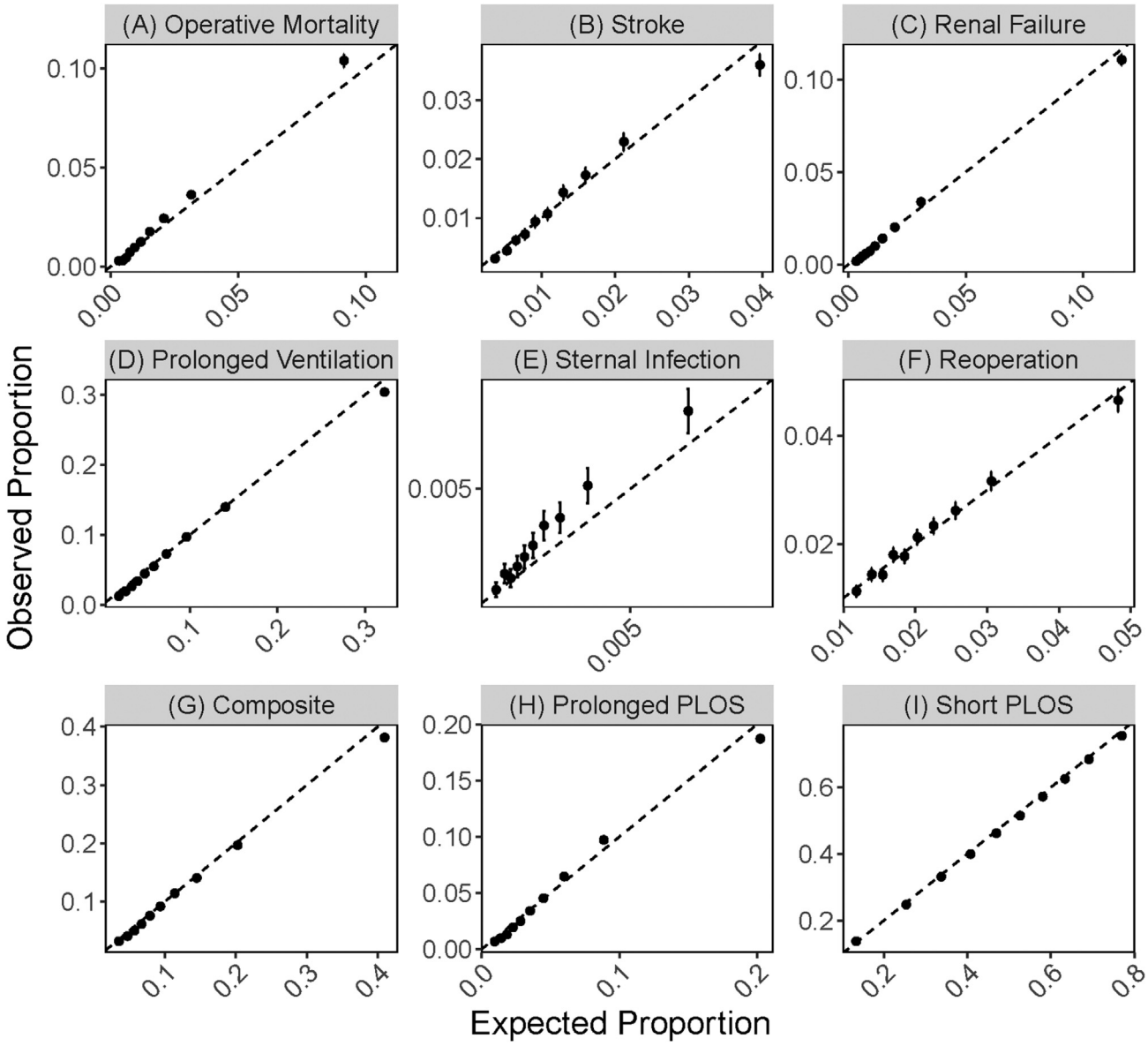


*Fig 1. Calibration for major endpoints in the validation sample: coronary artery bypass grafting surgery. (PLOS = postoperative length of stay.)*

Performance of the final models in the development sample was excellent for most of the population and endpoint combinations (Supplemental Material, Supplemental Tables 5, 6). Across the three model populations, the bootstrap-adjusted c-statistics were lowest for reoperation (range, 0.574 to 0.627) followed by stroke (range, 0.616 to 0.704) and were highest for renal failure (range, 0.749 to 0.810). Slopes to assess overfitting were generally close to the ideal value of 1.0 and were greater than 0.90 for all but three population-endpoint combinations. Models with slopes less than 0.90 were reoperation in valve (0.88), reoperation in valve plus CABG (0.78), and stroke in valve plus CABG (0.79). Calibration plots based on cross validation revealed acceptable calibration and no obvious violation of modeling assumptions.

After selecting the final set of models, regression coefficients were subsequently reestimated using multiple imputation to deal with missing endpoint data. After multiple imputation, the average predicted mortality risk across all populations increased from 2.50% to 2.58% (relative increase = 3%).

The c-statistics in the validation sample ranged from 0.588 for reoperation in valve plus CABG to 0.826 for renal failure in CABG. Table 4 presents c-statistics calculated in the validation sample for the final selected models and compares them with c-statistics calculated in the validation sample for the prior STS 2008 risk models. Although the DSWI model was estimated in a combined cohort that included all three procedure populations, its discrimination was assessed in each procedure population individually in Table 4. The c-statistics of the new STS models

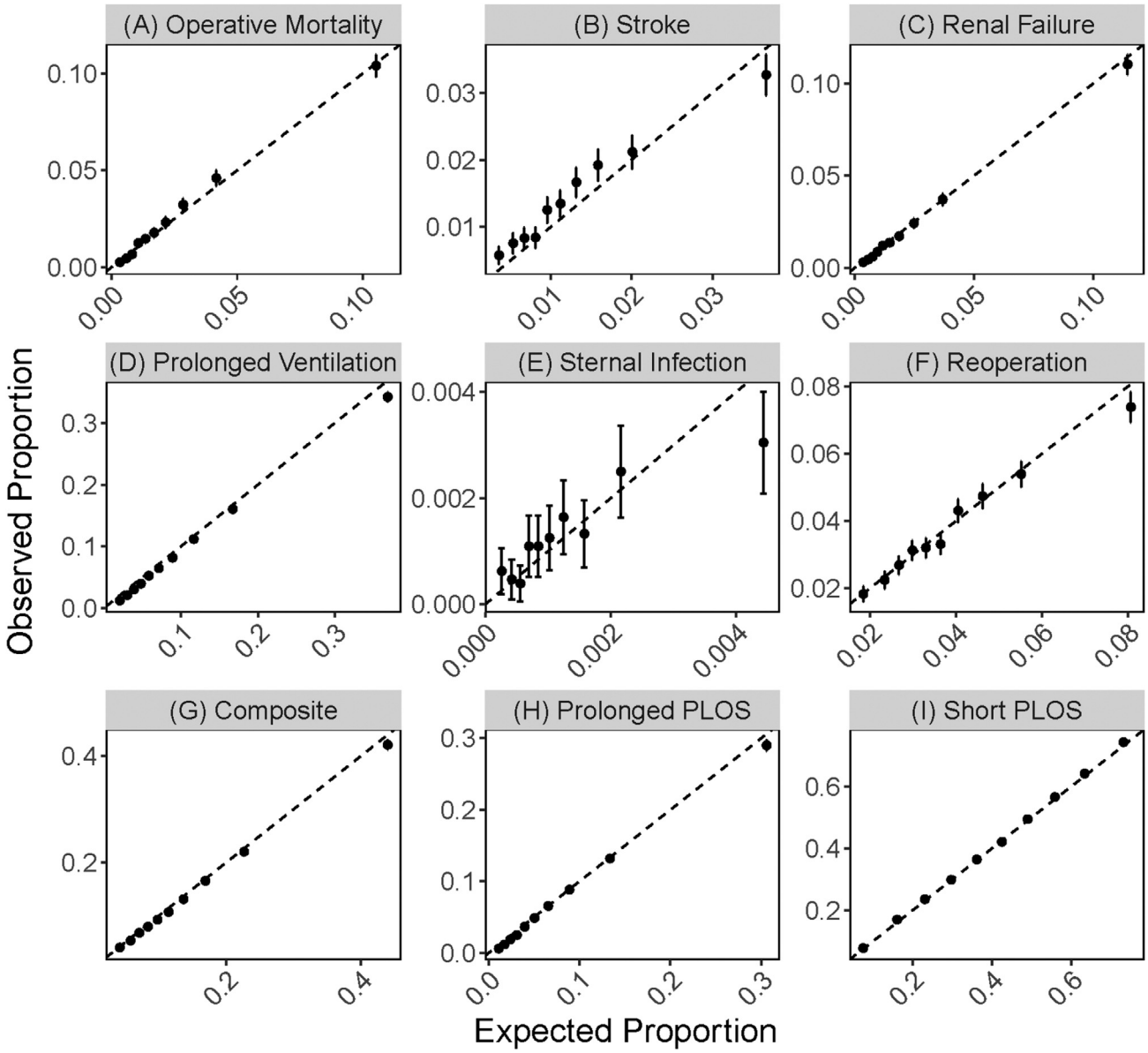


*Fig 2. Calibration for major endpoints in the validation sample: valve. (PLOS = postoperative length of stay.)*

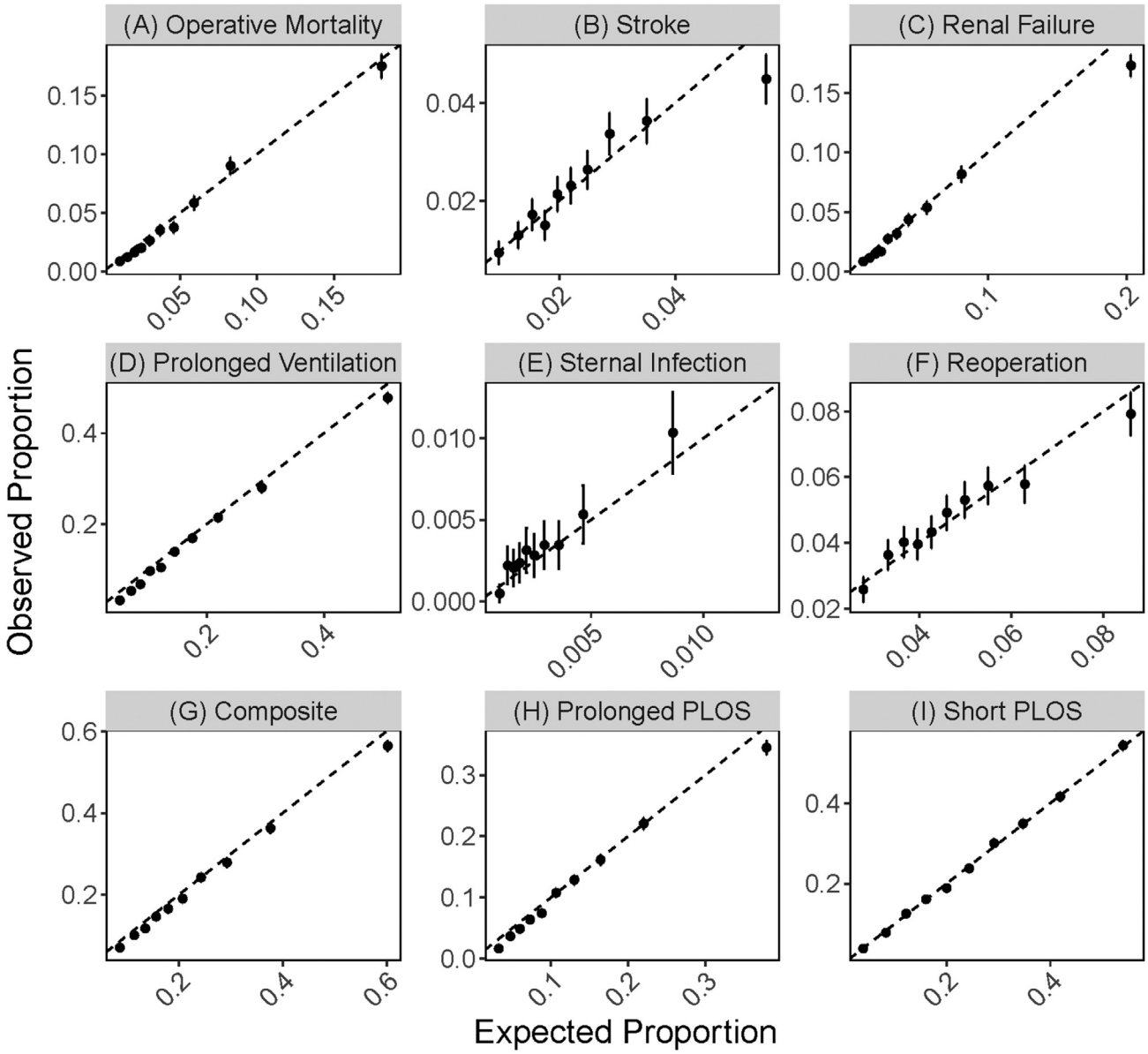


*Fig 3. Calibration for major endpoints in the validation sample: valve plus coronary artery bypass grafting surgery. (PLOS = postoperative length of stay.)*

exceeded those of the STS 2008 models for all populations and endpoints except for the valve model for stroke; all but two of the $p$ values were less than 0.05 (stroke and DSWI in valve plus CABG) and most were less than 0.0001 (Supplemental Table 7).

Calibration graphs in the validation sample are presented in Figures 1, 2, and 3. These reveal excellent calibration for the vast majority of populations and endpoints. The DSWI model appears to systematically underestimate infection risk in CABG by a factor of approximately 0.80, presumably because of a somewhat higher rate of this complication in more recent data. This underestimation of risk will be corrected when these models are used to calculate observed to expected ratios in the STS feedback reports; the report methodology applies a calibration factor that causes the expected rate to equal the observed rate within each calendar year of the reporting period. After applying the STS feedback report recalibration methodology to the validation sample, the calibration of the recalibrated DSWI model was excellent, as shown in Supplemental Figure 1. When these models are used to calculate STS composite scores, deteriorating calibration over time will be corrected automatically because model coefficients will be reestimated in the current STS data before composite scores are calculated.

## Comment

We have described the development and validation of a comprehensive set of new STS adult cardiac surgical risk

models that will be used to adjust for case mix in the STS participant feedback report and the STS voluntary public reporting program. Our approach to model development incorporated several novel features including the use of simulations to assess the feasible number of predictors in relation to sample size, and the combined use of bootstrapping and cross validation to estimate model operating characteristics as a function of the significance level for variable inclusion. Because the main intended use of these models was case-mix adjustment, we did not focus on parsimony (small number of covariates) as our primary goal but rather selected the optimal covariates for accurate risk prediction using a combination of statistical and clinical face validity approaches. The models showed good calibration, and 24 of 25 models had superior discrimination compared with the STS 2008 models when evaluated in the contemporary dataset used for model validation.

## References

1. 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.
2. 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.
3. 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.
4. Shahian DM, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery risk models: part 1—background, design considerations, and model development. Ann Thorac Surg 2018;105:1411–8.
5. Bellomo R, Ronco C, Kellum JA, Mehta RL, Palevsky P. Acute renal failure—definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Crit Care 2004;8:R204–12.
6. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
7. Harrell FE, Jr. Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. New York: Springer; 2015.
8. Jacobs JP, O'Brien SM, Shahian DM, et al. Successful linking of The Society of Thoracic Surgeons Database to Social Security data to examine the accuracy of Society of Thoracic Surgeons mortality data. J Thorac Cardiovasc Surg 2013;145:976–83.
9. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377–99.
10. SAS Institute Inc. SAS/Stat 13.2 User's Guide. Cary, NC: SAS Institute; 2014.
11. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. Comput Stat Data Anal 2010;54:2267–75.