Journal of the American Heart Association

ORIGINAL RESEARCH

Supplementing Existing Societal Risk Models for Surgical Aortic Valve Replacement With Machine Learning for Improved Prediction

Arman Kilic, MD; Robert H. Habib, PhD; James K. Miller, PhD; David M. Shahian, MD; Joseph A. Dearani, MD; Artur W. Dubrawski, PhD

BACKGROUND: This study evaluated the role of supplementing Society of Thoracic Surgeons (STS) risk models for surgical aortic valve replacement with machine learning (ML).

METHODS AND RESULTS: Adults undergoing isolated surgical aortic valve replacement in the STS National Database between 2007 and 2017 were included. ML models for operative mortality and major morbidity were previously developed using extreme gradient boosting. Concordance and discordance in predicted risk between ML and STS models were defined using equal-size tertile-based thresholds of risk. Calibration metrics and discriminatory capability were compared between concordant and discordant patients. A total of 243 142 patients were included. Nearly all calibration metrics were improved in cases of concordance. Similarly, concordance indices improved substantially in cases of concordance for all models with the exception of deep sternal wound infection. The greatest improvements in concordant versus discordant cases were in renal failure: ML model (concordance index, 0.660 [95% CI, 0.632–0.687] discordant versus 0.808 [95% CI, 0.794–0.822] concordant) and STS model (concordance index, 0.573 [95% CI, 0.549–0.576] discordant versus 0.797 [95% CI, 0.782–0.811] concordant) (each $P<0.001$). Excluding deep sternal wound infection, the concordance indices ranged from 0.549 to 0.660 for discordant cases and 0.674 to 0.808 for concordant cases.

CONCLUSIONS: Supplementing ML models with existing STS models for surgical aortic valve replacement may have an important role in risk prediction and should be explored further. In particular, for the roughly 25% to 50% of patients demonstrating discordance in estimated risk between ML and STS, there appears to be a substantial decline in predictive performance suggesting vulnerability of the existing models in these patient subsets.

Key Words: aortic valve replacement ■ complications ■ machine learning ■ mortality ■ risk prediction

The Society of Thoracic Surgeons (STS) risk models derived from national STS registry data have long served as the gold standard for risk assessment and prognostication in adult cardiac surgery.[1] Although the risk models have traditionally been associated with excellent calibration, they exhibit only moderate discriminatory capability, findings that are consistent in other population-level risk models in clinical medicine.[2,3] Interest in machine learning (ML) has risen exponentially recently, and the role of ML has been evaluated in several series in cardiac surgery.[4–6] Although improvements in predictive capability of risk models has been demonstrated with ML, what has yet to be explored is the role that supplementing different approaches may have in better understanding risk model vulnerabilities and potential avenues for

Correspondence to: Arman Kilic, MD, Division of Cardiac Surgery, Medical University of South Carolina, 30 Courtenay Drive, MSC 295, Suite BM279, Charleston, SC 29425. E-mail: kilica2@upmc.edu

Supplementary Material for this article is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.120.019697

For Sources of Funding and Disclosures, see page 10.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- This is a study of the Society of Thoracic Surgeons National Database for isolated aortic valve replacement evaluating the role of supplementing existing risk models with machine learning for improved risk prediction.
- Model performance was substantially improved in cases where machine learning and existing societal risk models displayed concordant risk prediction.

What Are the Clinical Implications?

- Machine learning can be used independently and as an adjunct to existing societal risk models for improving risk prediction in cardiac surgery.
- Further research is needed to identify why discordance exists in these models and leads to vulnerability in risk prediction in these patient subsets.

Nonstandard Abbreviations and Acronyms

DSWI	deep sternal wound infection
LR	logistic regression
ML	machine learning
SAVR	surgical aortic valve replacement
STS	Society of Thoracic Surgeons

improvement. The aim of this study was to evaluate the potential role of supplementing ML and STS risk models in predicting outcomes after surgical aortic valve replacement (SAVR).

METHODS

Study Cohort

The authors declare that all supporting data are available within the article and its online supplementary files. Adults undergoing isolated SAVR in the STS Adult Cardiac Surgery Database between 2007 and 2017 were included. This corresponded to STS Adult Cardiac Surgery Database data versions 2.61, 2.73, and 2.81. Patients undergoing concomitant coronary artery bypass grafting, other valve surgery, or other major cardiac procedures such that they were excluded from the isolated SAVR category as defined by the STS were excluded from analysis. As this registry contains deidentified data with no direct patient identifiers and was originally collected for nonresearch purposes, the Duke University Health System Institutional Review Board deemed this research exempt from review, as it does not qualify as human subjects research.[7] The requirement for informed consent for this study for each individual subject was waived.

ML Models

The outcomes for which models were evaluated included operative mortality, each major morbidity (acute renal failure, prolonged ventilation, reoperation, stroke, and deep sternal wound infection [DSWI]), and the composite outcome of either operative mortality or major morbidity. The clinical definitions and criteria for these were defined by the STS.[8] The ML algorithm that was used was extreme gradient boosting, or XGBoost. The methodologic approach, which involved randomly dividing the study cohort into training (80%) and testing (20%) cohorts, was detailed previously.[9] The approach to ML model derivation and validation and evaluation of model performance were also detailed previously.[9] In addition, categories of risk (low, intermediate, high) were defined using equal-size tertiles based on STS predicted risk. Matrices were then created and the observed rates of outcomes in the testing set were evaluated on the basis of concordance and discordance in predicted risk between the STS and ML risk models.[9] Concordance was defined as similar categorization of risk (low, intermediate, or high) by STS and ML, whereas discordance was defined as dissimilar categorization.

Predictive Utility of Concordance and Discordance

For each outcome, patients were stratified according to having concordant versus discordant predicted risk in STS and ML models.[9] The predictive performance of each model was then evaluated and compared between concordant and discordant patients. Comparisons in performance for each outcome for concordant versus discordant patients was evaluated for each ML and STS model separately. In addition, the impact on performance of a combined model that represented the average predicted risk between the ML and STS models was evaluated as well. Performance measures were evaluated in the testing sets and included metrics related to calibration and discriminatory ability of the models. Calibration was assessed using observed-to-expected ratios of the outcome, with an optimal value equaling 1; calibration-in-the-large or y-intercept of the calibration plot, with an optimal value equaling 0; and slope of the calibration curve, with an optimal value equaling 1.[10] Discriminatory ability of the models was

Table 1. Significant Predictors of Concordance and Discordance in Predicted Risk Between the ML and STS Models

Variable	Concordance							Discordance						
	Mortality	Comp	Renal Failure	Prolonged Ventilation	Reoperation	Stroke	DSWI	Mortality	Comp	Renal Failure	Prolonged Ventilation	Reoperation	Stroke	DSWI
Age, y (increasing)									x	x	x	x		x
Female			x		x			x	x		x		x	x
White								x	x	x		x	x	
BMI (increasing)										x	x			
BSA (increasing)					x			x	x				x	x
Hyperlipidemia														
Diabetes mellitus	x	x	x	x	x	x								
Hypertension	x	x	x	x	x									
Chronic lung disease—mild									x	x	x	x	x	x
Chronic lung disease—moderate	x			x								x	x	x
Chronic lung disease—severe	x	x	x	x								x	x	x
Dialysis							x	x			x	x		
Creatinine (increasing)	x	x	x	x	x									
Immunosuppressed			x								x	x		
Infective endocarditis			x	x		x								
PAD		x			x			x		x	x			
CVD			x	x	x								x	
FHCAD									x	x		x		
Redo						x	x	x	x	x		x		
Prior MI												x		x
Shock								x	x	x	x	x	x	
IABP		x		x	x									
AV insufficiency									x	x	x	x	x	
EF (increasing)		x		x	x	x	x							
Urgent status		x		x	x								x	
Emergent status	x	x	x	x	x	x								
Intraoperative*														

(*Continued*)

Table 1. Continued

Variable	Concordance							Discordance						
	Mortality	Comp	Renal Failure	Prolonged Ventilation	Reoperation	Stroke	DSWI	Mortality	Comp	Renal Failure	Prolonged Ventilation	Reoperation	Stroke	DSWI
CPB time (increasing)														
Aortic XC time (increasing)														
Blood transfusion								x	x	x	x	x		
Mechanical valve		x			x	x								

Intraoperative variables were added subsequently after identifying significant preoperative predictors. AV indicates aortic valve; BMI, body mass index; BSA, body surface area; Comp, composite of mortality or morbidity; CPB, cardiopulmonary bypass; CVD, cerebrovascular disease; DSWI, deep sternal wound infection; EF, ejection fraction; FHCAD, family history of coronary artery disease; IABP, intra-aortic balloon pump; MI, myocardial infarction; ML, machine learning; PAD, peripheral arterial disease; STS, Society of Thoracic Surgeons; and XC, cross-clamp.

*Intraoperative variables were entered into the multivariable model only after fully executing the multivariable models using only preoperative variables.

measured and compared using the area under the receiver operating characteristic curve, or concordance index.

Multivariable logistic regression models were also generated to identify independent predictors of discordance in predicted risk between ML and STS models for each outcome. These models were generated by evaluating each preoperative variable in the STS registry in univariate logistic regression analysis. Those variables with a significant association (exploratory $P<0.05$) with the outcome of discordance in predicted risk were then entered into the multivariable model. Variables with >10% missing data were excluded from these models. In addition to models using only preoperative variables, the impact of including intraoperative data to evaluate discordance in predicted risk was examined. Intraoperative variables, which included cardiopulmonary bypass time, aortic cross-clamp time, intraoperative blood product usage, and use of a mechanical valve, were evaluated in univariate logistic regression analysis, and those variables predictive of discordance were entered into separate multivariable models. The ML models were developed and validated using Python programming software (Python Software Foundation, Wilmington, DE). The remaining statistical analyses were performed with STATA software version 14 (StataCorp, College Station, TX).

De Novo Logistic Regression Models

The STS risk models are updated every few years with adjusted regression coefficients to reflect temporal changes as well as ongoing accrual of data within the registry. Therefore, we conducted a subsequent analysis in which we developed de novo logistic regression (LR) multivariable models for each outcome and used these instead of the STS risk models to ensure that the findings persisted. For these LR models, we evaluated only the preoperative variables that were evaluated for inclusion in the STS and ML models, thus ensuring that no extraneous data were incorporated that were not evaluated for use in the other models. Further, the same derivation cohort of patients from which the ML models were constructed were used to derive these de novo LR models, and the same external validation set of patients was used for validation. Univariate logistic regression was initially conducted on each candidate variable, and those with an exploratory P value of <0.05 were entered into a multivariable model for that particular outcome. Variables with >10% missing data were excluded. Similar analyses comparing discriminatory capability and calibration in cases of concordance versus discordance with ML were conducted but with these LR models instead of STS models.

RESULTS

Baseline Characteristics

A total of 243 142 patients undergoing isolated SAVR were included in this analysis. The baseline characteristics of the overall population as well as the development, validation, and performance of the ML models for each of the 7 outcomes is detailed in prior work.[9] In brief, the ML models were well calibrated and improved the discriminatory ability of the STS models for 5 of 7 outcomes (comparable performance between ML and STS for stroke and DSWI) in the overall study population.[9] The equal-patient-size tertile-based thresholds derived from the STS predicted risk of outcomes were also detailed previously.[9]

Rates and Predictors of Discordance in Predicted Risk Between ML and STS

Rates of discordance in predicted risk between ML and STS in the testing sets were as follows: 26.4% operative mortality, 34.4% composite of mortality and morbidity, 46.0% renal failure, 33.1% prolonged ventilation, 50.6% reoperation, 25.0% stroke, and 63.3% DSWI. A comparison of baseline preoperative characteristics between discordant and concordant patients revealed substantial differences for each of the 7 outcomes evaluated (Tables S1 through S7). With the exception of reoperation and DSWI, the STS predicted risk of the remaining outcomes was significantly lower in discordant patients (Tables S1 through S7). Similarly, several differences in intraoperative characteristics were noted as well (Tables S1 through S7).

In separate multivariable models that were created for each of the outcomes to identify predictors of discordance, there were distinct patterns that were noted (Table 1). For example, diabetes mellitus, hypertension, increasing serum creatinine, increasing ejection fraction, and emergent operative status were each associated with concordance in the vast majority of models (Table 1). Older age, female sex, White race, mild chronic lung disease, cardiogenic shock, and aortic valve insufficiency were each associated with discordance in the majority of models (Table 1). Of the intraoperative variables, mechanical valve placement was associated with concordance, whereas intraoperative blood product transfusion was associated with discordance in predicted risk by the ML and STS models for the majority of outcomes studied (Table 1).

Performance of the Models in Discordant Versus Concordant Cases

A comparison of calibration metrics demonstrated that the majority of metrics were improved in concordant cases as compared with discordant cases with the STS models (Figure 1 and Table 2). Similar trends in improvements in calibration metrics were observed with the ML models as well (Figure 2 and Table 3). A model that averaged the predicted risk between ML and STS again demonstrated improved calibration in the majority of outcomes with concordant patients (Table 4).

For all outcomes and all models with the exception of DSWI, the discriminatory capability or concordance index of the models was substantially higher in concordant cases (Figure 3 and Table 5). The greatest improvements in concordant versus discordant cases were in renal failure: ML model (concordance index, 0.660 [95% CI, 0.632–0.687] discordant versus 0.808 [95% CI, 0.794–0.822] concordant), STS model (concordance

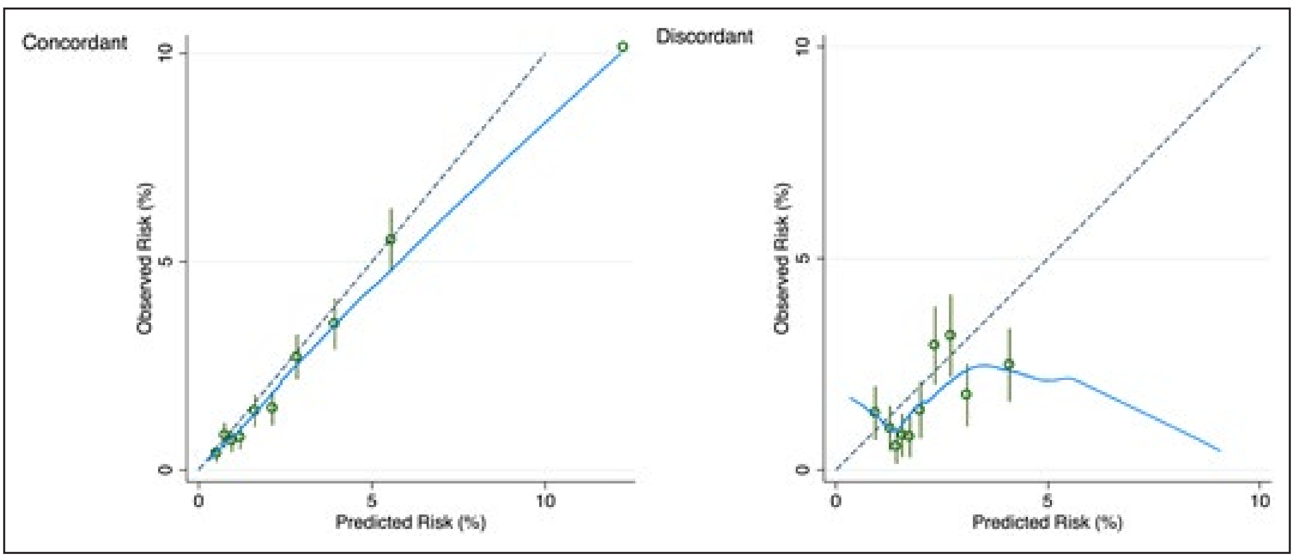

Figure 1. Improvement in calibration for Society of Thoracic Surgeons (STS) risk models for operative mortality in concordant cases.

Table 2. Improvement in Calibration Metrics of the STS Models in Cases of Concordance

STS Model		
Operative Mortality	**Discordant (n=12 615; 26.4%)**	**Concordant (n=35 191; 73.6%)**
Observed-to-expected ratio	0.770	0.866
Calibration-in-the-large	−0.267	−0.157
Slope of calibration curve	0.837	0.964
Composite of Mortality and Morbidity	**Discordant (n=16 441; 34.4%)**	**Concordant (n=31 397; 65.6%)**
Observed-to-expected ratio	0.742	0.868
Calibration-in-the-large	−0.354	−0.190
Slope of calibration curve	0.638	1.036
Renal Failure	**Discordant (n=21 506; 46.0%)**	**Concordant (n=25 264; 54.0%)**
Observed-to-expected ratio	0.399	0.749
Calibration-in-the-large	−0.953	−0.328
Slope of calibration curve	0.542	0.950
Prolonged Ventilation	**Discordant (n=15 861; 33.1%)**	**Concordant (n=32 110; 66.9%)**
Observed-to-expected ratio	0.672	0.802
Calibration-in-the-large	−0.438	−0.280
Slope of calibration curve	0.674	0.970
Reoperation	**Discordant (n=24 180; 50.6%)**	**Concordant (n=23 608; 49.4%)**
Observed-to-expected ratio	0.654	0.768
Calibration-in-the-large	−0.458	−0.289
Slope of calibration curve	0.813	1.088
Stroke	**Discordant (n=11 986; 25.0%)**	**Concordant (n=35 999; 75.0%)**
Observed-to-expected ratio	0.979	0.962
Calibration-in-the-large	−0.021	−0.039
Slope of calibration curve	0.647	0.916
Deep Sternal Wound Infection	**Discordant (n=9579; 63.3%)**	**Concordant (n=5559; 36.7%)**
Observed-to-expected ratio	1.149	0.839
Calibration-in-the-large	0.140	−0.177
Slope of calibration curve	0.502	0.246

STS indicates Society of Thoracic Surgeons.

index, 0.573 [95% CI, 0.549–0.576] discordant versus 0.797 [95% CI, 0.782–0.811] concordant), and combined ML/STS model (concordance index, 0.641 [95% CI, 0.614–0.669] discordant versus 0.807 [95% CI, 0.793–0.821] concordant) (each $P<0.001$) (Table 5). Excluding DSWI, the concordance indices ranged from 0.549 to 0.660 for discordant cases and from 0.674 to 0.808 for concordant cases (Table 5).

Findings Using De Novo LR Models

There were some differences noted in individual variables that were predictive of concordance or discordance between ML and LR as compared with ML and STS, although the majority remained similar (Table S8). The calibration metrics of the LR models were uniformly improved in cases of concordance with ML (Table S9). Although less pronounced, the majority of calibration metrics were improved in the ML models as well when concordant with LR (Table S10). The same findings held with the models developed by averaging the ML and LR predicted risk (Table S11). Similar to what was observed with the STS models, there was substantial improvement in concordance index in concordant cases across all model types and outcomes with the exception of DSWI (Table S12).

DISCUSSION

Risk modeling plays a vital role in cardiac surgery with important implications in program evaluation, quality improvement, patient prognostication, therapy selection, and clinical trial development. Historically, the STS risk models have demonstrated excellent calibration but only moderate discriminatory capability, findings that are fairly common in most population-level clinical risk models.[3] Therefore, although the models can accurately assign the rate of occurrence of an outcome for a population at hand, the models are less capable of identifying specific patients in whom that outcome will occur. This limits the ability to provide accurate individual patient counseling and decision making.

Substantial improvements in discriminatory capability of risk models to a point of achieving "state-of-the-art" performance with concordance indices >0.90 can potentially be obtained through several different mechanisms. Foremost, risk models are constrained by the available data points that are available for evaluation and inclusion in the models. If there are highly predictive elements, known or unknown, that are not captured within the data repository from which the model is built, this will likely constrain the performance of the model. Using ML techniques such as natural language processing and automated information extraction from electronic health records can help overcome the constraints of available data. This was demonstrated in a risk model for in-hospital mortality that achieved "state-of-the-art" predictive capability by taking advantage of the ability to analyze over 46 billion data points.[11] Translating these methods into large national clinical

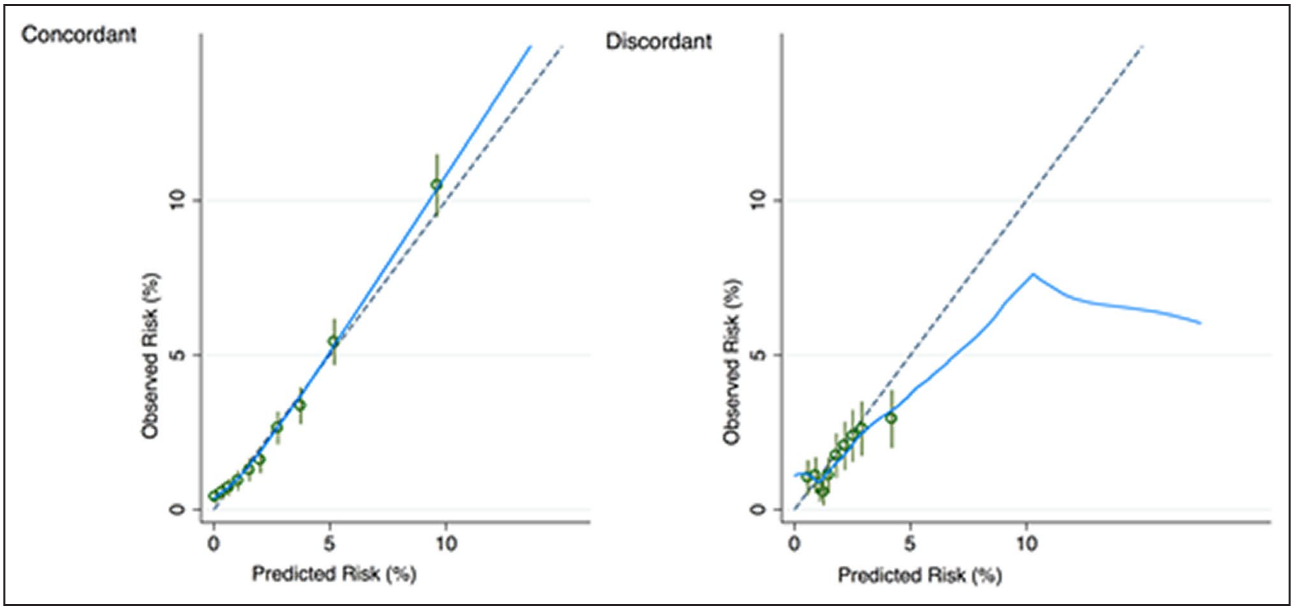

Figure 2. Improvement in calibration for machine learning risk models for operative mortality in concordant cases.

registries such as the STS present unique logistical challenges, and this remains an area of active investigation.

Another approach to improving risk model performance is to use different modeling strategies. Much of the literature describing both logistic regression and ML models for cardiac surgery, similar to other clinical fields, have focused on comparing isolated, singular approaches.[3,4,6] The concept of supplementing risk modeling approaches to determine potential utility is not necessarily novel in the ML world, but its application in the setting of large clinical registries such as the STS is new.

The main implication of the current analysis, which is the first of its kind to be performed in the STS national registry, is that supplementing ML and STS risk models allows us the ability to identify patient subsets where the STS risk models appear to be vulnerable. Our prior work demonstrated that the observed rates of outcomes in the training set fell within the range of predicted risk 100% of the time when there was concordance between the ML and STS models in predicted risk for each of the 7 outcomes studied in SAVR.[9] Concordance between the models does not appear to augment the predictive performance to "state-of-the-art," but rather, in those cases that are discordant, there appears to be a drastic decline in area under the receiver operating characteristic curve.

Also of interest is that there appear to be specific patterns and individual variables that predict greater likelihood of being discordant consistently across outcomes. This suggests that we can likely identify clusters of patients for whom existing models are likely to be less reliable. What to do with this information is a matter for debate and requires input from multiple stakeholders, including national societal leaders.

Currently, predicted risks are communicated to patients, clinicians, and to the STS as absolute values. One implication of the current analysis may be to communicate confidence levels in our estimated risks as well. For example, we have strong confidence that the estimated risk is X or that we are 95% confident that the risk will fall between X and Y; or perhaps excluding patients whose estimates are discordant and of low confidence is prudent for hospital and surgeon evaluation. Regardless, further investigation into understanding why and how ML and STS models are calculating risk differently and improving risk prediction in discordant patients is important for improving performance of the models in the overall population.

The STS risk models in addition to the ML models we developed for the current analysis included only preoperative variables. Clinicians appreciate the notion that the postoperative course of a cardiac surgical patient is also largely dictated by intraoperative events. Multiple prior reports have indeed demonstrated strong associations between intraoperative variables such as longer cardiopulmonary bypass times, longer aortic cross-clamp times, and intraoperative blood product transfusions with increased operative mortality and morbidity risk.[12–14]

Table 3. Improvement in Calibration Metrics of the ML Models in Cases of Concordance

ML Model		
Operative Mortality	**Discordant (n=12 615; 26.4%)**	**Concordant (n=35 191; 73.6%)**
Observed-to-expected ratio	0.860	1.017
Calibration-in-the-large	−0.154	0.016
Slope of calibration curve	0.806	0.987
Composite of Mortality and Morbidity	**Discordant (n=16 441; 34.4%)**	**Concordant (n=31 397; 65.6%)**
Observed-to-expected ratio	0.998	1.006
Calibration-in-the-large	−0.002	0.008
Slope of calibration curve	0.892	1.065
Renal Failure	**Discordant (n=21 506; 46.0%)**	**Concordant (n=25 264; 54.0%)**
Observed-to-expected ratio	0.820	1.043
Calibration-in-the-large	−0.203	0.040
Slope of calibration curve	0.812	0.937
Prolonged Ventilation	**Discordant (n=15 861; 33.1%)**	**Concordant (n=32 110; 66.9%)**
Observed-to-expected ratio	0.956	0.992
Calibration-in-the-large	−0.049	−0.011
Slope of calibration curve	0.954	1.043
Reoperation	**Discordant (n=24 180; 50.6%)**	**Concordant (n=23 608; 49.4%)**
Observed-to-expected ratio	1.029	1.017
Calibration-in-the-large	0.030	0.018
Slope of calibration curve	0.905	1.060
Stroke	**Discordant (n=11 986; 25.0%)**	**Concordant (n=35 999; 75.0%)**
Observed-to-expected ratio	0.949	1.006
Calibration-in-the-large	−0.054	0.006
Slope of calibration curve	0.406	0.945
Deep Sternal Wound Infection	**Discordant (n=9579; 63.3%)**	**Concordant (n=5559; 36.7%)**
Observed-to-expected ratio	3.311	1.091
Calibration-in-the-large	1.204	0.090
Slope of calibration curve	−0.021	0.474

ML indicates machine learning.

Table 4. Improvement in Calibration Metrics of Models Averaging ML and STS Risk in Cases of Concordance

Average Model		
Operative Mortality	**Discordant (n=12 615; 26.4%)**	**Concordant (n=35 191; 73.6%)**
Observed-to-expected ratio	0.812	0.935
Calibration-in-the-large	−0.212	−0.072
Slope of calibration curve	1.325	1.030
Composite of Mortality and Morbidity	**Discordant (n=16 441; 34.4%)**	**Concordant (n=31 397; 65.6%)**
Observed-to-expected ratio	0.851	0.932
Calibration-in-the-large	−0.188	−0.093
Slope of calibration curve	1.058	1.098
Renal Failure	**Discordant (n=21 506; 46.0%)**	**Concordant (n=25 264; 54.0%)**
Observed-to-expected ratio	0.537	0.871
Calibration-in-the-large	−0.640	−0.152
Slope of calibration curve	1.100	1.033
Prolonged Ventilation	**Discordant (n=15 861; 33.1%)**	**Concordant (n=32 110; 66.9%)**
Observed-to-expected ratio	0.789	0.887
Calibration-in-the-large	−0.258	−0.149
Slope of calibration curve	1.190	1.057
Reoperation	**Discordant (n=24 180; 50.6%)**	**Concordant (n=23 608; 49.4%)**
Observed-to-expected ratio	0.799	0.876
Calibration-in-the-large	−0.239	−0.145
Slope of calibration curve	1.147	1.146
Stroke	**Discordant (n=11 986; 25.0%)**	**Concordant (n=35 999; 75.0%)**
Observed-to-expected ratio	0.963	0.984
Calibration-in-the-Large	−0.038	−0.017
Slope of calibration curve	0.712	0.976
Deep Sternal Wound Infection	**Discordant (n=9579; 63.3%)**	**Concordant (n=5559; 36.7%)**
Observed-to-expected ratio	1.715	0.954
Calibration-in-the-Large	0.542	−0.047
Slope of calibration curve	0.566	0.484

ML indicates machine learning; and STS, Society of Thoracic Surgeons.

In the context of identifying potential reasons why discordance may exist in predicted risk between ML and STS approaches, we also performed a subanalysis in which we added intraoperative variables reliably coded in the STS registry to determine if any were predictors of discordance. It is conceivable that catastrophic intraoperative events, for example, would dramatically alter the postoperative risk of mortality and morbidity in a patient who was otherwise low risk when considering only baseline preoperative variables. Interestingly, we found that neither cardiopulmonary bypass time nor aortic cross-clamp time, both of which

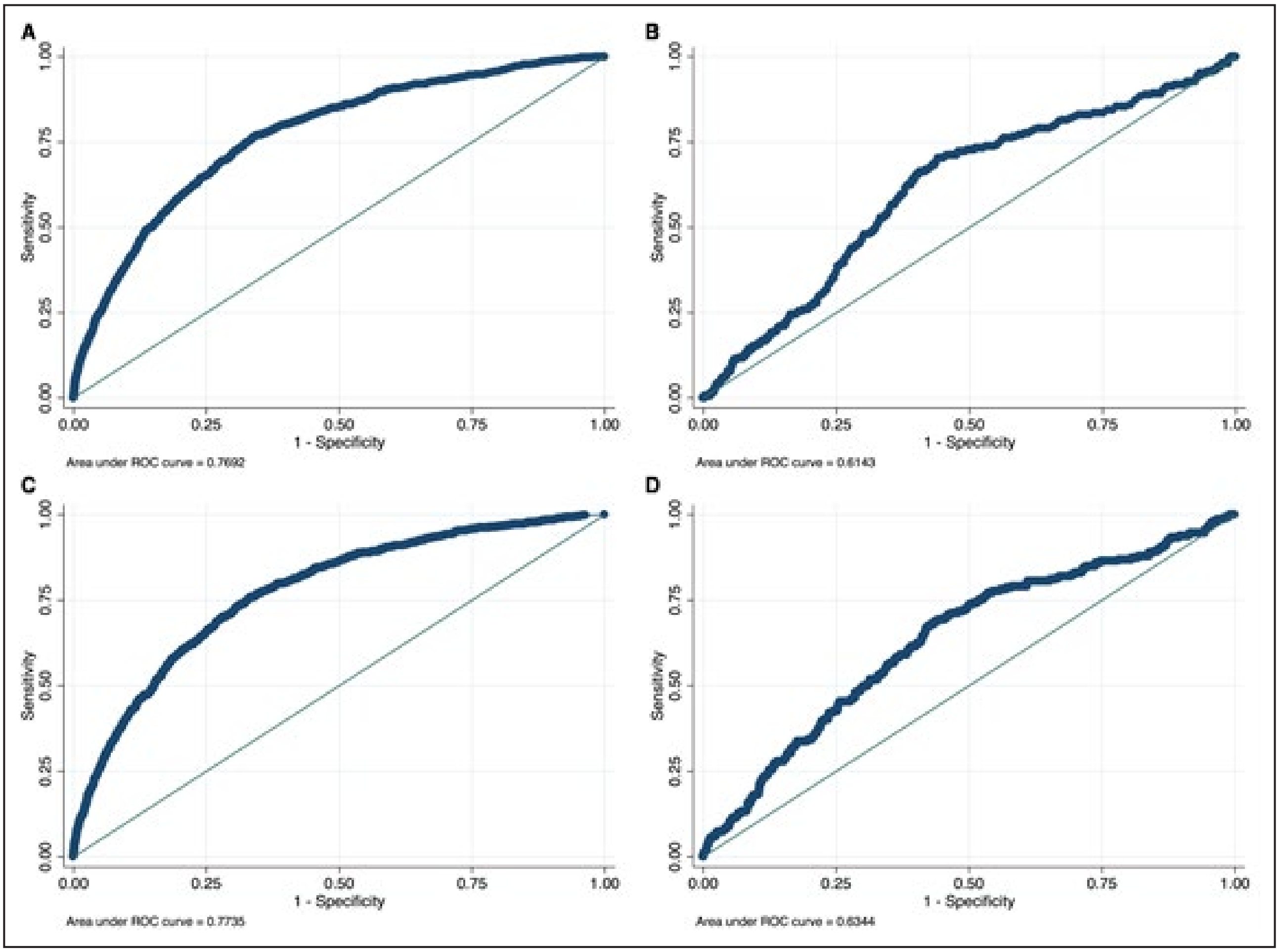

Figure 3. Improvement in area under the receiver operating characteristic (ROC) curve for the Society of Thoracic Surgeons models for operative mortality in (A) concordant vs (B) discordant cases, and for machine learning models in (C) concordant vs (D) discordant cases.

can be considered surrogates for intraoperative complications and operative efficiency, were predictive of discordance. Intraoperative blood transfusions, however, did reliably predict discordance in predicted risk for the majority of models.

Limitations

The current analysis evaluated only a specific ML algorithm. There are a plethora of other ML algorithms that exist and were not evaluated in this study, and therefore the generalizability of these results is unknown. Furthermore, we examined isolated SAVR using only the STS registry, and therefore whether these results extrapolate to other types of index cardiac operations remains to be elucidated. Other inherent limitations include the retrospective nature of the study design as well as errors in data entry, as is encountered with any multicenter registry.

CONCLUSIONS

This study of 243 142 patients undergoing isolated SAVR in the STS national database explored the utility of supplementing ML and STS risk models for operative mortality and major morbidity. The major finding was that in cases of discordant prediction, calibration was less reliable, and discriminatory capability as measured by concordance index was drastically reduced, as compared with cases of concordant prediction. In addition, distinct patterns were identified regarding variables that were reliably predictive of concordance or discordance in the majority of outcomes studied. Further investigation into methods of improving risk prediction in these subsets of patients for whom existing models are vulnerable appears prudent. These data highlight a potentially novel avenue to evaluate and refine risk modeling strategy in large clinical registries that carry profound implications in fields such as cardiac surgery.

Table 5. Improvement in Discriminatory Ability as Measured by Area Under the Receiver Operating Characteristic Curve in Cases of Concordance

Model	Concordance Index (95% CI)	Concordance Index (95% CI)	*P* Value
Operative Mortality	**Discordant (n=12 615; 26.4%)**	**Concordant (n=35 191; 73.6%)**	
ML	0.634 (0.595–0.673)	0.774 (0.759–0.788)	<0.001
STS	0.614 (0.576–0.653)	0.769 (0.754–0.784)	<0.001
Average of both models	0.650 (0.614–0.686)	0.775 (0.760–0.789)	<0.001
Composite of Mortality and Morbidity	**Discordant (n=16 441; 34.4%)**	**Concordant (n=31 397; 65.6%)**	
ML	0.590 (0.576–0.603)	0.734 (0.726–0.742)	<0.001
STS	0.563 (0.549–0.576)	0.726 (0.718–0.734)	<0.001
Average of both models	0.588 (0.575–0.601)	0.733 (0.726–0.741)	<0.001
Renal Failure	**Discordant (n=21 506; 46.0%)**	**Concordant (n=25 264; 54.0%)**	
ML	0.660 (0.632–0.687)	0.808 (0.794–0.822)	<0.001
STS	0.573 (0.543–0.603)	0.797 (0.782–0.811)	<0.001
Average of both models	0.641 (0.614–0.669)	0.807 (0.793–0.821)	<0.001
Prolonged Ventilation	**Discordant (n=15 861; 33.1%)**	**Concordant (n=32 110; 66.9%)**	
ML	0.628 (0.611–0.645)	0.769 (0.760–0.778)	<0.001
STS	0.580 (0.561–0.598)	0.759 (0.750–0.768)	<0.001
Average of both models	0.623 (0.605–0.640)	0.767 (0.758–0.776)	<0.001
Reoperation	**Discordant (n=24 180; 50.6%)**	**Concordant (n=23 608; 49.4%)**	
ML	0.574 (0.559–0.590)	0.686 (0.672–0.701)	<0.001
STS	0.549 (0.533–0.566)	0.674 (0.659–0.689)	<0.001
Average of both models	0.571 (0.555–0.588)	0.687 (0.672–0.701)	<0.001
Stroke	**Discordant (n=11 986; 25.0%)**	**Concordant (n=35 999; 75.0%)**	
ML	0.551 (0.505–0.597)	0.691 (0.670–0.711)	<0.001
STS	0.559 (0.512–0.607)	0.686 (0.665–0.706)	<0.001
Average of both models	0.570 (0.523–0.618)	0.690 (0.670–0.711)	<0.001
Deep Sternal Wound Infection	**Discordant (n=9579; 63.3%)**	**Concordant (n=5559; 36.7%)**	
ML	0.542 (0.451–0.632)	0.691 (0.526–0.855)	0.088
STS	0.560 (0.463–0.657)	0.531 (0.325–0.737)	0.396
Average of both models	0.571 (0.478–0.663)	0.647 (0.460–0.835)	0.247

ML indicates machine learning; and STS, Society of Thoracic Surgeons.

ARTICLE INFORMATION

Received October 8, 2020; accepted December 21, 2020.

Affiliations

From the Division of Cardiac Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA (A.K.); The Society of Thoracic Surgeons Research Center, Chicago, IL (R.H.H.); The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (J.K.M., A.W.D.); Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA (D.M.S.); and Department of Cardiovascular Surgery, Mayo Clinic, Rochester, MN (J.A.D.); Now with Division of Cardiothoracic Surgery, Medical University of South Carolina (A.K.).

Acknowledgments

The data for this research were provided by the Society of Thoracic Surgeons' National Database Participant Use File Research Program. Data analysis was performed at the University of Pittsburgh and Carnegie Mellon University.

Sources of Funding

None.

Disclosures

Dr Kilic is on the Medical Advisory Board of Medtronic, Inc. The remaining authors have no disclosures to report.

Supplementary Material

Tables S1–S12

REFERENCES

1. Caceres M, Braud RL, Garrett HE Jr. A short history of the Society of Thoracic Surgeons national cardiac database: perceptions of a practicing surgeon. *Ann Thorac Surg*. 2010;89:332–339.DOI: 10.1016/j.athoracsur.2009.09.045.
2. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr, Lobdell KW, Vassileva C, Wyler von Ballmoos MC, Thourani VH, 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;105:1411–1418.DOI: 10.1016/j.athoracsur.2018.03.002.

3. O'Brien SM, Feng L, He X, Xian Y, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr, Lobdell KW, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 2-statistical methods and results. *Ann Thorac Surg*. 2018;105:1419–1428.DOI: 10.1016/j.athoracsur.2018.03.003.
4. Kilic A, Goyal A, Miller JK, Gjekmarkaj E, Tam WL, Gleason TG, Sultan I, Dubrawksi A. Predictive utility of a machine learning algorithm in estimating mortality risk in cardiac surgery. *Ann Thorac Surg*. 2020;109:1811–1819.DOI: 10.1016/j.athoracsur.2019.09.049.
5. Allyn J, Allou N, Augustin P, Philip I, Martinet O, Belghiti M, Provenchere S, Montravers P, Ferdynus C. A comparison of a machine learning model with EuroSCORE II in predicting mortality after elective cardiac surgery: a decision curve analysis. *PLoS One*. 2017;12:e0169772. DOI: 10.1371/journal.pone.0169772.
6. Lee HC, Yoon HK, Nam K, Cho YJ, Kim TK, Kim WH, Bahk JH. Derivation and validation of machine learning approaches to predict acute kidney injury after cardiac surgery. *J Clin Med*. 2018;7:E322. DOI: 10.3390/jcm7100322.
7. Dokholyan RS, Muhlbaier LH, Falletta JM, Jacobs JP, Shahian D, Haan CK, Peterson ED. Regulatory and ethical considerations for linking clinical and administrative databases. *Am Heart J*. 2009;157:971–982.DOI: 10.1016/j.ahj.2009.03.023.
8. The Society of Thoracic Surgeons Cardiac Surgery Database: database data collection. Available at: https://www.sts.org/registries-research-center/sts-national-database/adult-cardiac-surgery-database/data-collection. Accessed November 14, 2019.
9. Kilic A, Goyal A, Miller JK, Gleason TG, Dubrawski A. Performance of a machine learning algorithm in predicting outcomes of aortic valve replacement. *Ann Thorac Surg*. 2021;111:503–510. DOI: 10.1016/j.athoracsur.2020.05.107.
10. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138.DOI: 10.1097/EDE.0b013e3181c30fb2.
11. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med*. 2018;1:18. DOI: 10.1038/s41746-018-0029-1.
12. Iino K, Miyata H, Motomura N, Watanabe G, Tomita S, Takemura H, Takamoto S. Prolonged cross-clamping during aortic valve replacement is an independent predictor of postoperative morbidity and mortality: analysis of the Japan Cardiovascular Surgery Database. *Ann Thorac Surg*. 2017;103:602–609.DOI: 10.1016/j.athoracsur.2016.06.060.
13. Chalmers J, Pullan M, Mediratta N, Poullis M. A need for speed? Bypass time and outcomes after isolated aortic valve replacement surgery. *Interact Cardiovasc Thorac Surg*. 2014;19:21–26.DOI: 10.1093/icvts/ivu102.
14. Vlot EA, Verwijmeren L, van de Garde EMW, Kloppenburg GTL, van Dongen EPA, Noordzij PG. Intra-operative red blood cell transfusion and mortality after cardiac surgery. *BMC Anesthesiol*. 2019;19:65. DOI: 10.1186/s12871-019-0738-2.

Supplemental Material

Table S1. Baseline preoperative and intraoperative characteristics of the discordant versus concordant cases for operative mortality.

	Discordant (n=12,615)	Concordant (n=35,191)	P-value
PRE-OPERATIVE			
Age (years)	67.7 ± 11.5	67.7 ± 13.2	0.79
Female	5,697 (45.2%)	13,505 (38.4%)	<0.001
Caucasian Race	11,127 (88.8%)	31,431 (89.8%)	0.001
Body Mass Index (kg/m^2)	30.7 ± 7.1	29.7 ± 6.5	<0.001
Body Surface Area (m^2)	1.98 ± 0.26	1.96 ± 0.25	<0.001
Dyslipidemia	9,160 (72.9%)	24,410 (69.5%)	<0.001
Diabetes Mellitus	4,344 (34.5%)	9,987 (28.4%)	<0.001
Hypertension	10,373 (82.4%)	27,059 (76.9%)	<0.001
Chronic Lung Disease			<0.001
None	9,717 (78.1%)	26,520 (75.8%)	
Mild	1,639 (13.2%)	4,401 (12.6%)	
Moderate	624 (5.0%)	2,059 (5.9%)	
Severe	253 (2.0%)	1,575 (4.5%)	
Yes, severity unknown	210 (1.7%)	425 (1.2%)	
Preoperative Dialysis	119 (1.0%)	945 (2.7%)	<0.001
Serum Creatinine (mg/dL)	1.05 ± 0.73	1.17 ± 1.00	<0.001
Immunosuppression	435 (3.5%)	1,442 (4.1%)	0.001
Infective Endocarditis	697 (5.5%)	2,063 (5.9%)	0.18
Peripheral Arterial Disease	929 (7.4%)	3,519 (10.0%)	<0.001
Cerebrovascular Disease	1,717 (13.7%)	5,115 (14.6%)	0.01
Family History of CAD	1,998 (16.2%)	5,268 (15.2%)	0.01
Number of Prior Open-Heart Surgeries			0.02
None	10,702 (84.9%)	30,258 (86.1%)	
One	1,740 (13.8%)	4,519 (12.9%)	
Two	136 (1.1%)	310 (0.9%)	
Three	22 (0.2%)	62 (0.2%)	
Four or More	6 (0.1%)	12 (0.03%)	

Previous MI	1,269 (10.1%)	3,813 (10.9%)	0.02
Recent Heart Failure in Past 2 Weeks	2,851 (33.1%)	6,588 (29.9%)	<0.001
Cardiogenic Shock	23 (0.2%)	255 (0.7%)	<0.001
Preoperative Intra-Aortic Balloon Pump	21 (0.2%)	119 (0.3%)	0.002
Aortic Valve Insufficiency			0.002
None	3,199 (26.7%)	8,844 (26.4%)	
Trivial/Trace	1,716 (14.3%)	4,610 (13.8%)	
Mild	3,041 (25.4%)	8,274 (24.7%)	
Moderate	1,915 (16.0%)	5,300 (15.8%)	
Severe	2,108 (17.6%)	6,440 (19.2%)	
Ejection Fraction (%)	56.7 ± 11.7	56.8 ± 11.5	0.17
Operative Urgency			<0.001
Elective	10,125 (80.3%)	27,434 (78.0%)	
Urgent	2,466 (19.6%)	7,414 (21.1%)	
Emergent	17 (0.1%)	312 (0.9%)	
Emergent Salvage	0 (0%)	18 (0.1%)	
STS Predicted Risk of Operative Mortality (%)	2.1 ± 0.9	3.2 ± 4.1	<0.001
INTRA-OPERATIVE			
Cardiopulmonary Bypass Time (min)	102.9 ± 39.2	103.3 ± 39.2	0.40
Aortic Cross-Clamp Time (min)	76.5 ± 28.8	76.6 ± 28.3	0.69
Lowest Temperature (degrees centigrade)	32.7 ± 2.7	32.7 ± 2.7	0.10
Lowest Hematocrit	24.7 ± 4.5	25.1 ± 4.8	<0.001
Highest Glucose	178.3 ± 48.5	176.2 ± 52.5	0.02
Ascending Aortic Calcification	353 (4.3%)	891 (4.2%)	0.83
Blood Product Transfusion	4,625 (36.8%)	13,853 (39.5%)	<0.001
Number of Units Transfused			
Red Blood Cells	1.7 ± 1.7	1.9 ± 1.7	<0.001
Platelets	1.1 ± 2.4	1.2 ± 2.7	0.03

Fresh Frozen Plasma	0.9 ± 1.8	1.0 ± 1.6	0.06
Cryoprecipitate	0.4 ± 1.8	0.4 ± 2.1	0.32
Mechanical Valve	1,117 (8.9%)	3,174 (9.0%)	0.58

CAD, coronary artery disease

Table S2. Baseline preoperative and intraoperative characteristics of the discordant versus concordant cases for the composite outcome of operative mortality or major morbidity.

	Discordant (n=16,441)	Concordant (n=31,397)	P-value
PRE-OPERATIVE			
Age (years)	70.9 ± 11.0	66.1 ± 13.3	<0.001
Female	7,015 (42.7%)	12,321 (39.2%)	<0.001
Caucasian Race	14,387 (88.2%)	28,198 (90.3%)	<0.001
Body Mass Index (kg/m^2)	30.3 ± 6.6	29.8 ± 6.6	<0.001
Body Surface Area (m^2)	1.97 ± 0.26	1.97 ± 0.25	0.01
Dyslipidemia	12,361 (75.3%)	21,137 (67.5%)	<0.001
Diabetes Mellitus	5,808 (35.4%)	8,462 (27.0%)	<0.001
Hypertension	14,007 (85.3%)	23,305 (74.3%)	<0.001
Chronic Lung Disease			<0.001
None	12,359 (75.7%)	23,813 (76.5%)	
Mild	2,318 (14.2%)	3,711 (11.9%)	
Moderate	943 (5.8%)	1,776 (5.7%)	
Severe	408 (2.5%)	1,480 (4.8%)	
Yes, severity unknown	304 (1.9%)	343 (1.1%)	
Preoperative Dialysis	123 (0.8%)	1,010 (3.2%)	<0.001
Serum Creatinine (mg/dL)	1.03 ± 0.49	1.19 ± 1.11	<0.001
Immunosuppression	657 (4.0%)	1,150 (3.7%)	0.07
Infective Endocarditis	587 (3.6%)	2,193 (7.0%)	<0.001
Peripheral Arterial Disease	1,701 (10.4%)	2,660 (8.5%)	<0.001
Cerebrovascular Disease	2,535 (15.5%)	4,241 (13.5%)	<0.001
Family History of CAD	2,315 (14.4%)	4,950 (16.0%)	<0.001

Number of Prior Open-Heart Surgeries			0.02
None	14,039 (85.5%)	27,155 (86.6%)	
One	2,187 (13.3%)	3,876 (12.4%)	
Two	155 (0.9%)	265 (0.8%)	
Three	31 (0.2%)	50 (0.2%)	
Four or More	11 (0.1%)	18 (0.1%)	
Previous MI	1,850 (11.3%)	3,312 (10.6%)	0.02
Recent Heart Failure in Past 2 Weeks	4,693 (36.6%)	4,701 (26.3%)	<0.001
Cardiogenic Shock	32 (0.2%)	252 (0.8%)	<0.001
Preoperative Intra-Aortic Balloon Pump	10 (0.1%)	146 (0.5%)	<0.001
Aortic Valve Insufficiency			<0.001
None	4,011 (25.6%)	8,137 (27.2%)	
Trivial/Trace	2,406 (15.4%)	4,009 (13.4%)	
Mild	4,140 (26.5%)	6,980 (23.4%)	
Moderate	2,472 (15.8%)	4,688 (15.7%)	
Severe	2,619 (16.7%)	6,061 (20.3%)	
Ejection Fraction (%)	57.1 ± 11.4	56.8 ± 11.5	0.04
Operative Urgency			<0.001
Elective	13,670 (83.2%)	23,867 (76.1%)	
Urgent	2,760 (16.8%)	7,209 (23.0%)	
Emergent	1 (0.01%)	282 (0.9%)	
Emergent Salvage	0 (0%)	22 (0.1%)	
STS Predicted Risk of the Composite Outcome of Operative Mortality or Major Morbidity (%)	16.9 ± 4.8	18.0 ± 12.6	<0.001
INTRA-OPERATIVE			
Cardiopulmonary Bypass Time (min)	101.2 ± 38.1	103.7 ± 39.2	<0.001
Aortic Cross-Clamp Time (min)	75.0 ± 27.8	77.2 ± 28.9	<0.001

Lowest Temperature (degrees centigrade)	32.7 ± 2.7	32.7 ± 2.7	0.05
Lowest Hematocrit	24.7 ± 4.5	25.3 ± 4.8	<0.001
Highest Glucose	177.3 ± 52.7	175.3 ± 46.8	0.01
Ascending Aortic Calcification	509 (4.1%)	704 (4.1%)	0.96
Blood Product Transfusion	6,044 (37.0%)	12,378 (39.6%)	<0.001
Number of Units Transfused			
Red Blood Cells	1.7 ± 1.5	1.9 ± 1.8	<0.001
Platelets	1.1 ± 2.6	1.2 ± 2.5	<0.001
Fresh Frozen Plasma	0.9 ± 1.5	1.1 ± 1.7	<0.001
Cryoprecipitate	0.4 ± 2.4	0.5 ± 2.0	0.61
Mechanical Valve	988 (6.0%)	3,202 (10.2%)	<0.001

CAD, coronary artery disease

Table S3. Baseline preoperative and intraoperative characteristics of the discordant versus concordant cases for postoperative renal failure.

	Discordant (n=21,506)	Concordant (n=25,264)	P-value
PRE-OPERATIVE			
Age (years)	71.7 ± 10.8	64.5 ± 13.5	<0.001
Female	8,955 (41.6%)	10,037 (39.7%)	<0.001
Caucasian Race	18,985 (88.9%)	22,889 (91.2%)	<0.001
Body Mass Index (kg/m^2)	30.1 ± 6.5	29.9 ± 6.7	0.007
Body Surface Area (m^2)	1.97 ± 0.26	1.97 ± 0.25	0.22
Dyslipidemia	16,438 (76.6%)	16,306 (64.8%)	<0.001
Diabetes Mellitus	8,241 (38.4%)	5,499 (21.8%)	<0.001
Hypertension	18,894 (87.9%)	17.454 (69.2%)	<0.001
Chronic Lung Disease			<0.001
None	15,819 (74.2%)	19,769 (78.9%)	
Mild	3,064 (14.4%)	2,659 (10.6%)	
Moderate	1,216 (5.7%)	1,408 (5.6%)	
Severe	805 (3.8%)	1,003 (4.0%)	
Yes, severity unknown	427 (2.0%)	224 (0.9%)	
Preoperative Dialysis	N/A	N/A	N/A
Serum Creatinine (mg/dL)	0.99 ± 0.24	1.07 ± 0.60	<0.001
Immunosuppression	857 (4.0%)	849 (3.4%)	<0.001
Infective Endocarditis	930 (4.3%)	1,515 (6.0%)	<0.001
Peripheral Arterial Disease	2,038 (9.5%)	2,171 (8.6%)	0.001
Cerebrovascular Disease	3,852 (18.0%)	2,739 (10.9%)	<0.001
Family History of CAD	3,171 (15.1%)	3,936 (15.8%)	0.03
Number of Prior Open-Heart Surgeries			<0.001
None	18,059 (84.0%)	22,156 (87.8%)	
One	3,187 (14.8%)	2,785 (11.0%)	
Two	191 (0.9%)	235 (0.9%)	
Three	41 (0.2%)	52 (0.25)	
Four or More	10 (0.1%)	21 (0.1%)	

Previous MI	2,440 (11.4%)	2,385 (9.5%)	<0.001
Recent Heart Failure in Past 2 Weeks	5,707 (35.0%)	3,356 (24.5%)	<0.001
Cardiogenic Shock	47 (0.2%)	208 (0.8%)	<0.001
Preoperative Intra-Aortic Balloon Pump	43 (0.2%)	80 (0.3%)	0.01
Aortic Valve Insufficiency			<0.001
None	5,449 (26.6%)	6,569 (27.5%)	
Trivial/Trace	3,129 (15.2%)	3,071 (12.9%)	
Mild	5,691 (27.7%)	5,468 (22.9%)	
Moderate	3,079 (15.0%)	3,793 (15.9%)	
Severe	3,178 (15.5%)	5,002 (20.9%)	
Ejection Fraction (%)	57.0 ± 11.5	57.0 ± 11.2	0.88
Operative Urgency			<0.001
Elective	17,048 (79.3%)	19,954 (79.0%)	
Urgent	4,437 (20.6%)	5,007 (19.8%)	
Emergent	14 (0.1%)	275 (1.1%)	
Emergent Salvage	0 (0%)	17 (0.1%)	
STS Predicted Risk of Renal Failure (%)	4.4 ± 2.2	4.5 ± 6.1	0.04
<u>INTRA-OPERATIVE</u>			
Cardiopulmonary Bypass Time (min)	101.9 ± 37.6	104.3 ± 40.1	<0.001
Aortic Cross-Clamp Time (min)	75.5 ± 27.5	77.7 ± 29.1	<0.001
Lowest Temperature (degrees centigrade)	32.7 ± 2.7	32.7 ± 2.6	0.04
Lowest Hematocrit	24.7 ± 4.6	25.6 ± 4.9	<0.001
Highest Glucose	177.3 ± 50.4	173.9 ± 46.6	<0.001
Ascending Aortic Calcification	756 (4.8%)	421 (3.2%)	<0.001
Blood Product Transfusion	8,415 (39.4%)	9,240 (36.7%)	<0.001
Number of Units Transfused			
Red Blood Cells	1.8 ± 1.7	1.9 ± 1.8	<0.001
Platelets	1.1 ± 2.6	1.2 ± 2.5	0.20

Fresh Frozen Plasma	0.9 ± 1.5	1.1 ± 1.8	<0.001
Cryoprecipitate	0.4 ± 2.1	0.4 ± 2.0	0.39
Mechanical Valve	1,225 (5.7%)	2,955 (11.7%)	<0.001

CAD, coronary artery disease

Table S4. Baseline preoperative and intraoperative characteristics of the discordant versus concordant cases for prolonged ventilation.

	Discordant (n=15,861)	Concordant (n=32,110)	P-value
PRE-OPERATIVE			
Age (years)	70.9 ± 11.2	66.2 ± 13.2	<0.001
Female	7,571 (47.7%)	11,857 (36.9%)	<0.001
Caucasian Race	14,065 (89.3%)	28,674 (89.8%)	0.10
Body Mass Index (kg/m^2)	30.4 ± 6.7	29.8 ± 6.7	<0.001
Body Surface Area (m^2)	1.97 ± 0.26	1.97 ± 0.25	0.49
Dyslipidemia	11,849 (74.9%)	21,718 (67.8%)	<0.001
Diabetes Mellitus	5,303 (33.5%)	8,925 (27.8%)	<0.001
Hypertension	13,410 (84.7%)	24,198 (75.5%)	<0.001
Chronic Lung Disease			<0.001
None	12,222 (77.6%)	24,070 (75.6%)	
Mild	2,176 (13.8%)	3,842 (12.1%)	
Moderate	738 (4.7%)	1,915 (6.0%)	
Severe	372 (2.4%)	1,535 (4.8%)	
Yes, severity unknown	241 (1.5%)	484 (1.5%)	
Preoperative Dialysis	118 (0.8%)	998 (3.1%)	<0.001
Serum Creatinine (mg/dL)	1.02 ± 0.47	1.20 ± 1.12	<0.001
Immunosuppression	500 (3.2%)	1,354 (4.2%)	<0.001
Infective Endocarditis	544 (3.4%)	2,248 (7.0%)	<0.001
Peripheral Arterial Disease	1,338 (8.5%)	3,165 (9.9%)	0.001
Cerebrovascular Disease	2,464 (15.6%)	4,402 (13.7%)	<0.001
Family History of CAD	2,341 (15.1%)	4,902 (15.5%)	0.29
Number of Prior Open-Heart Surgeries			<0.001
None	13,818 (87.2%)	27,534 (85.8%)	
One	1,879 (11.9%)	4,114 (12.8%)	
Two	121 (0.8%)	356 (1.1%)	
Three	23 (0.2%)	58 (0.2%)	
Four or More	5 (0.03%)	16 (0.1%)	

Previous MI	1,581 (10.0%)	3,435 (10.7%)	0.02
Recent Heart Failure in Past 2 Weeks	4,247 (35.8%)	5,118 (27.0%)	<0.001
Cardiogenic Shock	13 (0.1%)	228 (0.7%)	<0.001
Preoperative Intra-Aortic Balloon Pump	5 (0.03%)	112 (0.4%)	<0.001
Aortic Valve Insufficiency			<0.001
None	3,879 (25.8%)	8,214 (27.0%)	
Trivial/Trace	2,304 (15.3%)	3,919 (12.9%)	
Mild	4,092 (27.2%)	7,202 (23.6%)	
Moderate	2.358 (15.7%)	4,867 (16.0%)	
Severe	2,393 (15.9%)	6,273 (20.6%)	
Ejection Fraction (%)	57.4 ± 11.1	56.6 ± 11.6	<0.001
Operative Urgency			<0.001
Elective	13,352 (84.2%)	24,282 (75.7%)	
Urgent	2,500 (15.8%)	7,496 (23.4%)	
Emergent	1 (0.01%)	298 (0.9%)	
Emergent Salvage	0 (0%)	19 (0.1%)	
STS Predicted Risk of Prolonged Ventilation (%)	9.9 ± 3.7	11.8 ± 11.3	<0.001
INTRA-OPERATIVE			
Cardiopulmonary Bypass Time (min)	101.0 ± 38.0	104.2 ± 39.6	<0.001
Aortic Cross-Clamp Time (min)	75.2 ± 27.6	77.5 ± 29.2	<0.001
Lowest Temperature (degrees centigrade)	32.8 ± 2.6	32.7 ± 2.6	0.61
Lowest Hematocrit	24.5 ± 4.5	25.4 ± 4.9	<0.001
Highest Glucose	176.6 ± 56.2	176.2 ± 47.4	0.66
Ascending Aortic Calcification	453 (4.0%)	767 (4.2%)	0.34
Blood Product Transfusion	6,010 (38.1%)	12,523 (39.2%)	0.02
Number of Units Transfused			
Red Blood Cells	1.7 ± 1.6	1.9 ± 1.8	<0.001
Platelets	1.0 ± 2.1	1.2 ± 2.5	<0.001

Fresh Frozen Plasma	0.8 ± 1.5	1.1 ±1.7	<0.001
Cryoprecipitate	0.4 ± 2.5	0.5 ± 2.1	0.50
Mechanical Valve	1,088 (6.9%)	3,184 (9.9%)	<0.001

CAD, coronary artery disease

Table S5. Baseline preoperative and intraoperative characteristics of the discordant versus concordant cases for reoperation.

	Discordant (n=24,180)	Concordant (n=23,608)	P-value
PRE-OPERATIVE			
Age (years)	71.7 ± 11.2	63.7 ± 13.1	<0.001
Female	9,570 (39.6%)	9,798 (41.5%)	<0.001
Caucasian Race	21,094 (87.8%)	21,459 (91.4%)	<0.001
Body Mass Index (kg/m^2)	29.5 ± 6.5	30.4 ± 6.8	<0.001
Body Surface Area (m^2)	1.94 ± 0.26	1.99 ± 0.25	<0.001
Dyslipidemia	17,949 (74.5%)	15,496 (65.8%)	<0.001
Diabetes Mellitus	8,138 (33.7%)	6,104 (25.9%)	<0.001
Hypertension	20,062 (83.1%)	17,409 (73.9%)	<0.001
Chronic Lung Disease			<0.001
None	17,288 (72.1%)	19,057 (81.4%)	
Mild	3,602 (15.0%)	2,271 (9.7%)	
Moderate	1,587 (6.6%)	1,071 (4.6%)	
Severe	1,031 (4.3%)	815 (3.5%)	
Yes, severity unknown	471 (2.0%)	192 (0.8%)	
Preoperative Dialysis	393 (1.6%)	700 (3.0%)	<0.001
Serum Creatinine (mg/dL)	1.12 ± 0.81	1.15 ± 1.09	0.003
Immunosuppression	896 (3.7%)	1,001 (4.3%)	0.003
Infective Endocarditis	1,057 (4.4%)	1,729 (7.3%)	<0.001
Peripheral Arterial Disease	3,054 (12.7%)	1,370 (5.8%)	<0.001
Cerebrovascular Disease	4,337 (18.0%)	2,482 (10.5%)	<0.001
Family History of CAD	3,432 (14.5%)	3,798 (16.3%)	<0.001
Number of Prior Open-Heart Surgeries			<0.001
None	20,004 (82.8%)	20.955 (88.9%)	
One	3,870 (16.0%)	2,368 (10.1%)	
Two	240 (1.0%)	180 (0.8%)	
Three	36 (0.2%)	49 (0.2%)	
Four or More	9 (0.04%)	20 (0.1%)	

Previous MI	2,872 (11.9%)	2,163 (9.2%)	<0.001
Recent Heart Failure in Past 2 Weeks	6,273 (34.3%)	3,108 (25.0%)	<0.001
Cardiogenic Shock	51 (0.2%)	204 (0.9%)	<0.001
Preoperative Intra-Aortic Balloon Pump	18 (0.1%)	114 (0.5%)	<0.001
Aortic Valve Insufficiency			<0.001
None	5,896 (25.5%)	6,216 (27.9%)	
Trivial/Trace	3,406 (14.8%)	2,891 (13.0%)	
Mild	5,969 (25.9%)	5,170 (23.2%)	
Moderate	3,670 (15.9%)	3,433 (15.4%)	
Severe	4,150 (18.0%)	4,561 (20.5%)	
Ejection Fraction (%)	56.3 ± 12.0	57.5 ± 10.8	<0.001
Operative Urgency			<0.001
Elective	19,074 (78.9%)	18,458 (78.2%)	
Urgent	5,092 (21.1%)	4,838 (20.5%)	
Emergent	4 (0.02%)	292 (1.2%)	
Emergent Salvage	4 (0.02%)	14 (0.1%)	
STS Predicted Risk of Reoperation (%)	8.3 ± 2.0	7.5 ± 4.2	<0.001
INTRA-OPERATIVE			
Cardiopulmonary Bypass Time (min)	102.4 ± 38.7	104.5 ± 40.6	<0.001
Aortic Cross-Clamp Time (min)	75.5 ± 27.7	78.0 ± 29.7	<0.001
Lowest Temperature (degrees centigrade)	32.7 ± 2.7	32.8 ± 2.7	<0.001
Lowest Hematocrit	24.5 ± 4.6	25.8 ± 4.9	<0.001
Highest Glucose	175.6 ± 52.4	176.6 ± 47.6	0.26
Ascending Aortic Calcification	844 (4.8%)	376 (3.1%)	<0.001
Blood Product Transfusion	10,139 (42.1%)	8,496 (36.1%)	<0.001
Number of Units Transfused			
Red Blood Cells	1.8 ± 1.6	1.9 ± 1.9	<0.001

Platelets	1.1 ± 2.4	1.2 ± 2.5	<0.001
Fresh Frozen Plasma	0.9 ± 1.6	1.1 ± 1.8	<0.001
Cryoprecipitate	0.4 ± 1.9	0.4 ± 2.1	0.40
Mechanical Valve	1,299 (5.4%)	2,954 (12.5%)	<0.001

CAD, coronary artery disease

Table S6. Baseline preoperative and intraoperative characteristics of the discordant versus concordant cases for stroke.

	Discordant (n=11,986)	Concordant (n=35,999)	P-value
PRE-OPERATIVE			
Age (years)	67.4 ± 11.4	67.9 ± 13.3	<0.001
Female	5,127 (42.8%)	14,186 (39.4%)	<0.001
Caucasian Race	10,267 (86.3%)	32,445 (90.65)	<0.001
Body Mass Index (kg/m^2)	30.5 ± 6.9	29.8 ± 6.6	<0.001
Body Surface Area (m^2)	1.97 ± 0.24	1.97 ± 0.26	0.01
Dyslipidemia	8,435 (70.6%)	25,158 (70.1%)	0.24
Diabetes Mellitus	3,812 (31.9%)	10,496 (29.2%)	<0.001
Hypertension	9,468 (79.2%)	28,214 (78.5%)	0.12
Chronic Lung Disease			<0.001
None	8,855 (74.6%)	27,477 (76.9%)	
Mild	1,530 (12.9%)	4,466 (12.5%)	
Moderate	766 (6.5%)	1,909 (5.3%)	
Severe	502 (4.2%)	1,355 (3.8%)	
Yes, severity unknown	211 (1.8%)	516 (1.4%)	
Preoperative Dialysis	335 (2.8%)	675 (1.9%)	<0.001
Serum Creatinine (mg/dL)	1.16 ± 1.08	1.11 ± 0.86	<0.001
Immunosuppression	519 (4.4%)	1,366 (3.8%)	0.008
Infective Endocarditis	844 (7.1%)	1,875 (5.2%)	<0.001
Peripheral Arterial Disease	995 (8.3%)	3,390 (9.4%)	<0.001
Cerebrovascular Disease	1,232 (10.3%)	5,582 (15.6%)	<0.001
Family History of CAD	1,856 (15.8%)	5,542 (15.7%)	0.78
Number of Prior Open-Heart Surgeries			<0.001
None	10,525 (87.9%)	30,591 (85.1%)	
One	1,302 (10.9%)	4,970 (13.8%)	
Two	102 (0.9%)	341 (1.0%)	
Three	25 (0.2%)	53 (0.2%)	
Four or More	15 (0.1%)	11 (0.03%)	

Previous MI	1,265 (10.6%)	3,820 (10.6%)	0.90
Recent Heart Failure in Past 2 Weeks	2,446 (30.6%)	6,940 (30.5%)	0.12
Cardiogenic Shock	38 (0.3%)	224 (0.6%)	<0.001
Preoperative Intra-Aortic Balloon Pump	40 (0.3%)	76 (0.2%)	0.02
Aortic Valve Insufficiency			0.005
None	2,880 (25.4%)	9,170 (26.8%)	
Trivial/Trace	1,578 (13.9%)	4,767 (13.9%)	
Mild	2,779 (24.5%)	8,539 (24.9%)	
Moderate	1,855 (16.4%)	5,325 (15.6%)	
Severe	2,245 (19.8%)	6,431 (18.8%)	
Ejection Fraction (%)	56.4 ± 12.1	57.0 ± 11.1	<0.001
Operative Urgency			<0.001
Elective	8,956 (74.8%)	28,873 (80.2%)	
Urgent	2,994 (25.0%)	6,817 (18.9%)	
Emergent	21 (0.2%)	274 (0.8%)	
Emergent Salvage	2 (0.02%)	20 (0.1%)	
STS Predicted Risk of Stroke (%)	1.3 ± 0.5	1.5 ± 1.2	<0.001
INTRA-OPERATIVE			
Cardiopulmonary Bypass Time (min)	102.5 ± 39.2	102.9 ±39.2	0.32
Aortic Cross-Clamp Time (min)	76.4 ± 29.1	76.4 ± 29.0	0.78
Lowest Temperature (degrees centigrade)	32.8 ± 2.6	32.7 ± 2.7	0.10
Lowest Hematocrit	24.6 ± 4.5	25.2 ± 4.8	<0.001
Highest Glucose	175.8 ± 53.4	176.3 ± 50.9	0.62
Ascending Aortic Calcification	290 (3.8%)	864 (4.0%)	0.47
Blood Product Transfusion	4,567 (38.3%)	13,983 (39.0%)	0.18
Number of Units Transfused			
Red Blood Cells	1.7 ± 1.8	1.9 ± 1.7	<0.001
Platelets	1.2 ± 2.7	1.2 ± 2.6	0.94

Fresh Frozen Plasma	1.0 ± 1.6	1.0 ± 1.7	0.41
Cryoprecipitate	0.4 ± 2.1	0.4 ± 1.9	0.49
Mechanical Valve	969 (8.1%)	3,380 (9.4%)	<0.001

CAD, coronary artery disease

Table S7. Baseline preoperative and intraoperative characteristics of the discordant versus concordant cases for deep sternal wound infection.

	Discordant (n=9,579)	Concordant (n=5,559)	P-value
PRE-OPERATIVE			
Age (years)	69.5 ± 10.4	60.7 ± 13.5	<0.001
Female	3,215 (33.6%)	2,615 (47.0%)	<0.001
Caucasian Race	8,549 (90.6%)	4,762 (87.2%)	<0.001
Body Mass Index (kg/m^2)	31.6 ± 6.9	28.3 ± 5.8	<0.001
Body Surface Area (m^2)	2.05 ± 0.24	1.89 ± 0.25	<0.001
Dyslipidemia	7,334 (76.9%)	3,551 (64.2%)	<0.001
Diabetes Mellitus	3,280 (34.3%)	1,304 (23.5%)	<0.001
Hypertension	8,021 (84.0%)	3,936 (71.0%)	<0.001
Chronic Lung Disease			<0.001
None	6,820 (72.6%)	4,641 (85.6%)	
Mild	1,194 (12.7%)	351 (6.5%)	
Moderate	512 (5.5%)	138 (2.5%)	
Severe	374 (4.0%)	99 (1.8%)	
Yes, severity unknown	492 (5.2%)	195 (3.6%)	
Preoperative Dialysis	244 (2.6%)	75 (1.4%)	<0.001
Serum Creatinine (mg/dL)	1.16 ± 1.05	1.03 ± 0.75	<0.001
Immunosuppression	432 (4.5%)	218 (4.0%)	0.09
Infective Endocarditis	472 (4.9%)	515 (9.35)	<0.001
Peripheral Arterial Disease	751 (7.9%)	420 (7.6%)	0.52
Cerebrovascular Disease	1,576 (16.6%)	759 (13.8%)	<0.001
Family History of CAD	1,206 (13.2%)	718 (13.7%)	0.47
Number of Prior Open-Heart Surgeries			<0.001
None	8,656 (90.5%)	4,794 (86.4%)	
One	845 (8.8%)	692 (12.5%)	
Two	46 (0.5%)	54 (1.0%)	
Three	15 (0.2%)	5 (0.1%)	
Four or More	7 (0.1%)	4 (0.1%)	

Previous MI	1,013 (10.7%)	529 (9.6%)	0.04
Recent Heart Failure in Past 2 Weeks	3,289 (34.6%)	1,567 (28.5%)	<0.001
Cardiogenic Shock	54 (0.6%)	23 (0.4%)	0.21
Preoperative Intra-Aortic Balloon Pump	23 (0.2%)	6 (0.1%)	0.07
Aortic Valve Insufficiency			<0.001
None	2,107 (23.1%)	1,059 (19.9%)	
Trivial/Trace	1,458 (16.0%)	704 (13.2%)	
Mild	2,567 (28.15)	1,228 (23.1%)	
Moderate	1,476 (16.2%)	912 (17.1%)	
Severe	1,529 (16.7%)	1,417 (26.6%)	
Ejection Fraction (%)	57.3 ± 11.3	58.4 ± 10.1	<0.001
Operative Urgency			0.001
Elective	7,852 (82.0%)	4,445 (80.0%)	
Urgent	1,676 (17.5%)	1,072 (19.3%)	
Emergent	48 (0.5%)	35 (0.6%)	
Emergent Salvage	1 (0.01%)	6 (0.1%)	
STS Predicted Risk of Deep Sternal Wound Infection (%)	0.3 ± 0.2	0.2 ± 0.2	<0.001
INTRA-OPERATIVE			
Cardiopulmonary Bypass Time (min)	99.4 ± 37.7	100.6 ± 38.1	0.07
Aortic Cross-Clamp Time (min)	74.2 ± 26.8	75.6 ± 28.3	0.002
Lowest Temperature (degrees centigrade)	33.0 ± 2.6	32.9 ± 2.7	0.001
Lowest Hematocrit	25.6 ± 4.7	25.0 ± 4.9	<0.001
Highest Glucose	175.9 ± 48.9	177.2 ± 49.1	0.14
Ascending Aortic Calcification	231 (2.5%)	104 (1.9%)	0.03
Blood Product Transfusion	2,643 (27.9%)	1,663 (30.2%)	0.002
Number of Units Transfused			
Red Blood Cells	1.5 ± 1.7	1.6 ± 1.5	0.02
Platelets	1.0 ± 2.0	1.0 ± 2.2	0.37

Fresh Frozen Plasma	0.9 ± 1.6	0.8 ± 1.4	0.54
Cryoprecipitate	0.5 ± 1.7	0.5 ± 1.9	0.78
Mechanical Valve	718 (7.5%)	987 (17.8%)	<0.001

CAD, coronary artery disease

Table S8. Significant predictors of concordance and discordance in predicted risk between the ML and LR models. Intraoperative variables were added subsequently after identifying significant preoperative predictors.

Variable	Concordance							Discordance						
	Mort	Comp	Ren Fail	Pro Vent	Reop	Stroke	DSWI	Mort	Comp	Ren Fail	Pro Vent	Reop	Stroke	DSWI
Age (inc.)	x					x			x	x	x	x		x
Female					x			x	x		x			
White						x								
BMI (inc.)							x	x	x	x	x	x	x	
BSA (inc.)					x	x	x	x						
HLD		x	x				x	x				x	x	
Diabetes						x	x	x	x	x	x			
HTN								x	x	x	x			
Chronic Lung Disease - Mild				x			x			x		x	x	
Chronic Lung Disease - Moderate	x	x	x	x	x		x						x	
Chronic Lung Disease - Severe	x	x	x	x	x		x						x	
DIalysis	x	x		x	x		x							
Creatinine (inc.)	x	x	x	x	x		x						x	
Immuno	x						x							
Inf Endo			x		x		x	x	x		x		x	
PAD	x	x	x	x	x	x	x							
CVD	x					x	x			x				
FHCAD														
Redo	x	x	x	x		x	x							
Prior MI	x			x		x	x							
Shock	x	x	x	x	x	x	x							
IABP	x	x	x	x	x		x							
AV Insuff	x	x	x		x						x	x		
EF (inc.)	x	x	x	x										x
Urgent Status	x	x		x	x		x							

Emergent Status	x	x	x	x	x	x	x							
INTRA-OP *														
CPB Time (inc.)														
Aortic XC Time (inc.)														
Blood Tx								x				x	x	
Mech Valve									x	x		x		

*** Intraoperative variables were entered into the multivariable model only after fully executing the multivariable models using only preoperative variables.**

AV, aortic valve; BMI, body mass index; BSA, body surface area; cpb, cardiopulmonary bypass; CVD, cerebrovascular disease; EF, ejection fraction; FHCAD, family history of coronary artery disease; HLD, hyperlipidemia; HTN, hypertension; iabp, intra-aortic balloon pump; immuno, immunosuppressed; inc, increasing; inf endo, infective endocarditis; insuff, insufficiency; intra-op, intraoperative; mech, mechanical; mi, myocardial infarction; PAD, peripheral arterial disease; tx, transfusion; XC, cross-clamp

Table S9. Improvement in calibration metrics of the LR models in cases of concordance.

LR Model		
Operative Mortality	***Discordant (n=14,875; 33.3%)***	***Concordant (n=29,821; 66.7%)***
Observed-to-Expected Ratio	0.922	1.026
Calibration-in-the-Large	-0.072	0.027
Slope of Calibration Curve	0.463	1.071
Composite of Mortality and Morbidity	***Discordant (n=14,307; 33.4%)***	***Concordant (n=28,527; 66.6%)***
Observed-to-Expected Ratio	0.963	1.014
Calibration-in-the-Large	-0.043	0.018
Slope of Calibration Curve	0.529	1.076
Renal Failure	***Discordant (n=17,757; 42.5%)***	***Concordant (n=24,036; 57.5%)***
Observed-to-Expected Ratio	0.824	1.160
Calibration-in-the-Large	-0.199	0.157
Slope of Calibration Curve	0.024	1.110
Prolonged Ventilation	***Discordant (n=13,932; 32.5%)***	***Concordant (n=28,911; 67.5%)***
Observed-to-Expected Ratio	1.008	1.008

Calibration-in-the-Large	0.008	0.010
Slope of Calibration Curve	0.325	1.021
Reoperation	***Discordant (n=24,180; 50.6%)***	***Concordant (n=23,608; 49.4%)***
Observed-to-Expected Ratio	0.981	0.999
Calibration-in-the-Large	-0.019	-0.001
Slope of Calibration Curve	0.312	1.104
Stroke	***Discordant (n=11,131; 25.9%)***	***Concordant (n=31,870; 74.1%)***
Observed-to-Expected Ratio	1.038	0.991
Calibration-in-the-Large	0.038	-0.009
Slope of Calibration Curve	0.464	0.970
Deep Sternal Wound Infection	***Discordant (n=31,734; 67.7%)***	***Concordant (n=15,112; 32.3%)***
Observed-to-Expected Ratio	0.915	0.941
Calibration-in-the-Large	-0.089	-0.062
Slope of Calibration Curve	0.160	0.708

Table S10. Improvement in calibration metrics of the ML models in cases of concordance.

ML Model		
Operative Mortality	***Discordant (n=14,875; 33.3%)***	***Concordant (n=29,821; 66.7%)***
Observed-to-Expected Ratio	0.902	0.999
Calibration-in-the-Large	-0.106	-0.003
Slope of Calibration Curve	0.626	1.003
Composite of Mortality and Morbidity	***Discordant (n=14,307; 33.4%)***	***Concordant (n=28,527; 66.6%)***
Observed-to-Expected Ratio	0.956	1.019
Calibration-in-the-Large	-0.052	0.026
Slope of Calibration Curve	0.913	1.063
Renal Failure	***Discordant (n=17,757; 42.5%)***	***Concordant (n=24,036; 57.5%)***
Observed-to-Expected Ratio	0.810	1.047
Calibration-in-the-Large	-0.219	0.045
Slope of Calibration Curve	0.964	0.977
Prolonged Ventilation	***Discordant (n=13,932; 32.5%)***	***Concordant (n=28,911; 67.5%)***

Observed-to-Expected Ratio	0.901	1.022
Calibration-in-the-Large	-0.114	0.026
Slope of Calibration Curve	1.015	1.027
Reoperation	***Discordant (n=24,180; 50.6%)***	***Concordant (n=23,608; 49.4%)***
Observed-to-Expected Ratio	1.021	1.025
Calibration-in-the-Large	0.022	0.027
Slope of Calibration Curve	1.053	1.071
Stroke	***Discordant (n=11,131; 25.9%)***	***Concordant (n=31,870; 74.1%)***
Observed-to-Expected Ratio	0.919	0.997
Calibration-in-the-Large	-0.085	-0.003
Slope of Calibration Curve	0.424	0.999
Deep Sternal Wound Infection	***Discordant (n=31,734; 67.7%)***	***Concordant (n=15,112; 32.3%)***
Observed-to-Expected Ratio	0.111	0.249
Calibration-in-the-Large	-2.215	-1.407

Slope of Calibration Curve	0.386	-0.084

Table S11. Improvement in calibration metrics of models averaging ML and LR risk in concordant versus discordant cases.

Average Model		
Operative Mortality	***Discordant (n=14,875; 33.3%)***	***Concordant (n=29,821; 66.7%)***
Observed-to-Expected Ratio	0.917	1.012
Calibration-in-the-Large	-0.089	0.013
Slope of Calibration Curve	1.131	1.135
Composite of Mortality and Morbidity	***Discordant (n=14,307; 33.4%)***	***Concordant (n=28,527; 66.6%)***
Observed-to-Expected Ratio	0.960	1.016
Calibration-in-the-Large	-0.047	0.021
Slope of Calibration Curve	1.300	1.140
Renal Failure	***Discordant (n=17,757; 42.5%)***	***Concordant (n=24,036; 57.5%)***
Observed-to-Expected Ratio	0.817	1.100
Calibration-in-the-Large	-0.208	0.102
Slope of Calibration Curve	1.467	1.199
Prolonged Ventilation	***Discordant (n=13,932; 32.5%)***	***Concordant (n=28,911; 67.5%)***

Observed-to-Expected Ratio	0.951	1.015
Calibration-in-the-Large	-0.054	0.018
Slope of Calibration Curve	1.422	1.096
Reoperation	***Discordant (n=24,180; 50.6%)***	***Concordant (n=23,608; 49.4%)***
Observed-to-Expected Ratio	1.001	1.012
Calibration-in-the-Large	0.001	0.013
Slope of Calibration Curve	1.509	1.163
Stroke	***Discordant (n=11,131; 25.9%)***	***Concordant (n=31,870; 74.1%)***
Observed-to-Expected Ratio	0.976	0.994
Calibration-in-the-Large	-0.025	-0.006
Slope of Calibration Curve	0.605	1.032
Deep Sternal Wound Infection	***Discordant (n=31,734; 67.7%)***	***Concordant (n=15,112; 32.3%)***
Observed-to-Expected Ratio	0.198	0.394
Calibration-in-the-Large	-1.627	-0.939

Slope of Calibration Curve	0.464	0.049

Table S12. Improvement in discriminatory ability as measured by area under receiver-operating-characteristic curve in cases of concordance.

Model	***C-Index (95% Confidence Interval)***	***C-Index (95% Confidence Interval)***	***p-value***
Operative Mortality	***Discordant (n=14,875; 33.3%)***	***Concordant (n=29,821; 66.7%)***	
ML	0.614 (0.577-0.651)	0.777 (0.762-0.793)	<0.001
LR	0.556 (0.519-0.593)	0.765 (0.749-0.780)	<0.001
Average of Both Models	0.630 (0.594-0.665)	0.779 (0.763-0.794)	<0.001
Composite of Mortality and Morbidity	***Discordant (n=14,307; 33.4%)***	***Concordant (n=28,527; 66.6%)***	
ML	0.604 (0.590-0.612)	0.730 (0.722-0.738)	<0.001
LR	0.529 (0.515-0.543)	0.715 (0.707-0.724)	<0.001
Average of Both Models	0.599 (0.586-0.613)	0.727 (0.719-0.736)	<0.001
Renal Failure	***Discordant (n=17,757; 42.5%)***	***Concordant (n=24,036; 57.5%)***	
ML	0.712 (0.685-0.740)	0.805 (0.789-0.820)	<0.001
LR	0.505 (0.474-0.535)	0.769 (0.753-0.784)	<0.001
Average of Both Models	0.694 (0.666-0.722)	0.801 (0.786-0.817)	<0.001
Prolonged Ventilation	***Discordant (n=13,932; 32.5%)***	***Concordant (n=28,911; 67.5%)***	
ML	0.644 (0.626-0.662)	0.766 (0.757-0.775)	<0.001

LR	0.543 (0.524-0.562)	0.752 (0.743-0.762)	<0.001
Average of Both Models	0.632 (0.613-0.650)	0.764 (0.755-0.773)	<0.001
Reoperation	***Discordant (n=24,180; 50.6%)***	***Concordant (n=23,608; 49.4%)***	
ML	0.593 (0.573-0.612)	0.667 (0.654-0.681)	<0.001
LR	0.521 (0.500-0.541)	0.654 (0.641-0.668)	<0.001
Average of Both Models	0.580 (0.560-0.600)	0.665 (0.652-0.679)	<0.001
Stroke	***Discordant (n=11,131; 25.9%)***	***Concordant (n=31,870; 74.1%)***	
ML	0.548 (0.501-0.595)	0.700 (0.679-0.722)	<0.001
LR	0.542 (0.493-0.591)	0.692 (0.670-0.714)	<0.001
Average of Both Models	0.551 (0.503-0.599)	0.697 (0.676-0.719)	<0.001
Deep Sternal Wound Infection	***Discordant (n=31,734; 67.7%)***	***Concordant (n=15,112; 32.3%)***	
ML	0.573 (0.488-0.657)	0.470 (0.394-0.545)	0.075
LR	0.513 (0.437-0.589)	0.567 (0.490-0.645)	0.324
Average of Both Models	0.573 (0.488-0.658)	0.491 (0.420-0.563)	0.150