HHS Public Access

Author manuscript

J Thorac Cardiovasc Surg. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

J Thorac Cardiovasc Surg. 2023 April ; 165(4): 1449–1459.e15. doi:10.1016/j.jtcvs.2021.09.010.

Prediction of operative mortality for patients undergoing cardiac surgical procedures without established risk scores

Chin Siang Ong, MBBS, PhD[a], **Erik Reinertsen, MD, PhD**[b,c,d], **Haoqi Sun, PhD**[e], **Philicia Moonsamy, MD, MPH**[a], **Navyatha Mohan, MD**[a], **Masaki Funamoto, MD**[a], **Tsuyoshi Kaneko, MD**[f], **Prem S. Shekar, MD**[f], **Stefano Schena, MD, PhD**[g], **Jennifer S. Lawton, MD**[g], **David A. D'Alessandro, MD**[a], **M. Brandon Westover, MD, PhD**[e,h], **Aaron D. Aguirre, MD, PhD**[b,c,i,j], **Thoralf M. Sundt, MD**[a]

[a]Division of Cardiac Surgery, Massachusetts General Hospital and Corrigan Minehan Heart Center, Boston, Mass

[b]Division of Cardiology, Massachusetts General Hospital and Corrigan Minehan Heart Center, Boston, Mass

[c]Center for Systems Biology, Massachusetts General Hospital, Boston, Mass

[d]Research Laboratory for Electronics, Massachusetts Institute of Technology, Cambridge, Mass

[e]Division of Clinical Neurophysiology, Department of Neurology, Massachusetts General Hospital, Boston, Mass

[f]Division of Cardiac Surgery, Brigham and Women's Hospital, Boston, Mass

[g]Division of Cardiac Surgery, Johns Hopkins Hospital, Baltimore, Md

[h]Clinical Data AI Center, Massachusetts General Hospital, Boston, Mass

[i]Wellman Center for Photomedicine, Massachusetts General Hospital and Harvard Medical School, Boston, Mass.

[j]Healthcare Transformation Lab, Massachusetts General Hospital, Boston, Mass

Abstract

Objective: Current cardiac surgery risk models do not address a substantial fraction of procedures. We sought to create models to predict the risk of operative mortality for an expanded set of cases.

Address for reprints: Aaron D. Aguirre, MD, PhD, Division of Cardiology, Massachusetts General Hospital, 55 Fruit St, Thier 212, Boston, MA 02114 (Aguirre.Aaron@mgh.harvard.edu).

Chin Siang Ong and Erik Reinertsen contributed equally to this work.

Conflict of Interest Statement

The authors reported no conflicts of interest.

The *Journal* policy requires editors and reviewers to disclose conflicts of interest and to decline handling or reviewing manuscripts for which they may have a conflict of interest. The editors and reviewers of this article have no conflicts of interest.

Code Availability

Code used for this work has been made available on GitHub. The repository is available at https://github.com/aguirre-lab/sts-ml.

Methods: Four supervised machine learning models were trained using preoperative variables present in the Society of Thoracic Surgeons (STS) data set of the Massachusetts General Hospital to predict and classify operative mortality in procedures without STS risk scores. A total of 424 (5.5%) mortality events occurred out of 7745 cases. Models included logistic regression with elastic net regularization (LogReg), support vector machine, random forest (RF), and extreme gradient boosted trees (XGBoost). Model discrimination was assessed via area under the receiver operating characteristic curve (AUC), and calibration was assessed via calibration slope and expected-to-observed event ratio. External validation was performed using STS data sets from Brigham and Women's Hospital (BWH) and the Johns Hopkins Hospital (JHH).

Results: Models performed comparably with the highest mean AUC of 0.83 (RF) and expected-to-observed event ratio of 1.00. On external validation, the AUC was 0.81 in BWH (RF) and 0.79 in JHH (LogReg/RF). Models trained and applied on the same institution's data achieved AUCs of 0.81 (BWH: LogReg/RF/XGBoost) and 0.82 (JHH: LogReg/RF/XGBoost).

Conclusions: Machine learning models trained on preoperative patient data can predict operative mortality at a high level of accuracy for cardiac surgical procedures without established risk scores. Such procedures comprise 23% of all cardiac surgical procedures nationwide. This work also highlights the value of using local institutional data to train new prediction models that account for institution specific practices.

Keywords

cardiac surgery; risk prediction; machine learning; operative mortality

Mortality risk stratification models have been used to guide quality improvement and to inform clinician and patient decision-making in cardiac surgery since the late 1980s.[1-3] Previously, creation of these risk models required a large data set addressing a carefully selected uniform set of procedures. For example, the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database, with more than 6 million records,[4] has been used to create 3 mortality risk models: coronary artery bypass graft (CABG), valve, and CABG/valve.[2,5] However, these models do not accommodate a substantial fraction of cardiac surgical procedures at many quaternary care centers. Cases without STS risk models (eg, tricuspid valve procedures, aortic procedures, left ventricular assist device placement, etc) account for 23% of all procedures nationally[4] and more than 40% of all procedures at many quaternary care centers. There is, therefore, a need for new risk models that encompass the diversity and complexity of modern cardiac surgical care.

Machine learning (ML) techniques might be useful to develop risk models for these heterogeneous procedures. ML models can capture relationships between input variables ("features") that might improve prediction of outcomes compared with traditional regression models. Provided there are sufficient data inputs, institution-specific models can be generated as well, enabling even more locally relevant predictions. Moreover, such institution-specific risk models can in principle learn from new data generated from every case and be updated yearly or quarterly, in contrast to the widely used statewide or national risk models that might be revised only once per decade.

We therefore trained and evaluated 4 supervised ML models to: (1) estimate the probability, and (2) classify (eg, predict the outcome) of operative mortality of procedures with no national STS risk models using data from our institution and evaluated using data from 2 other institutions. Additionally, we present simplified models with limited preoperative feature sets that maintain good performance.

METHODS

Patient Population

Between 2002 and 2019, 9248 procedures were performed at the Massachusetts General Hospital (MGH; Boston, Mass) for which no STS risk models exist using STS version 2.9 definitions. These “other” case types comprise 44.5% of all cardiac surgical procedures at MGH (Tables 1 and E1). Patients younger than the age of 18 years, or those who underwent procedures involving circulatory arrest (n = 1503), were excluded, leaving a class-imbalanced data set of 7745 procedures with 5.5% (n = 424) operative mortality (see Table 1, which includes preoperative and intraoperative variables). Similar STS data sets were obtained from Brigham and Women’s Hospital (BWH; Boston, MA; n = 7529; 2002-2019; Table E2) and Johns Hopkins Hospital (JHH; Baltimore, Md; n = 9706; 1997-2019; Table E3) for external validation. Relevant approvals from the respective institutional review boards were obtained before the commencement of the study.

Creation of Risk Models

Variables found to be meaningful in the existing STS Risk models for operative mortality[5] and used in the online calculator[6] were extracted from each institutional STS data set. Because the study time frame spanned multiple STS versions, to harmonize data from different STS versions, for variables for which definitions became more granular in subsequent STS versions, earlier definitions, that were less granular, were used. Detailed methodology regarding data preprocessing, in particular handling of missing data, is reported in Appendix E1. Only preoperative variables were used for modeling and the list is presented in Table E4. Four supervised ML risk models were trained: 2 linear models (logistic regression with elastic net regularization [LogReg], and support vector machine [SVM]) and 2 nonlinear models (random forest [RF], and extreme gradient boosted trees [XGBoost][7]). Hyperparameter optimization was performed via fivefold nested cross-validation[8] (not non-nested cross-validation[9]; see Appendix E1). For each patient, each model estimated the probability of operative mortality, ranging between 0 to 1. This nested cross-validation approach resulted in distinct training, validation, and test sets, such that the same subjects are never used for training and evaluation of models. Discrimination and calibration performance on either cross-validated test set or external cohort data was determined via plotting receiver operating characteristic (Figure 1) and precision-recall curves (Figure E1) and computing a variety of metrics. Probability calibration was performed via Platt scaling; if calibrated probabilities achieved expected-to-observed event (E/O) ratios and calibration slope (CS) was closer to 1 than uncalibrated probabilities; the former instead of the latter were used for the calculation of metrics and plots (Figure 1).[10]

Model performance is reported as median (lower bound-upper bound) using 95% confidence intervals (Table 2). Model performance was estimated and compared by bootstrapping (sampling with replacement) estimated probabilities of test set data 10,000 times. Features were ranked according to absolute variable importance or coefficient, from highest to lowest (Table 3). Detailed methodology is in Appendix E1.

External validation.—External validation was performed by using models trained on patients from MGH to estimate mortality risk on data sets from BWH and JHH. Institution-specific risk models were also trained and assessed on BWH and JHH cohorts using the nested cross-validation method as previously described.

Parsimonious model.—To evaluate discriminative performance versus number of top variables, logistic regression risk and calibration models were trained on MGH "other" case types using forward feature selection to select a smaller subset of variables. To assess external validity, these parsimonious models were applied to BWH and JHH cohorts. Additional details are in Appendix E1.

RESULTS

All Models Achieve Similar Performance for Models Trained and Evaluated at MGH

Discriminative performance of LogReg (area under receiver operating characteristic curve [AUC], 0.82; E/O, 1.00) was comparable with SVM (AUC, 0.82; $P = .36$; E/O, 1.00), RF (AUC, 0.83; $P = .09$; E/O, 1.00) and XGBoost (AUC, 0.82; $P = .80$; E/O, 0.99; Table 2). Across models, the most predictive variables of operative mortality for patients who underwent cardiac surgical procedures that did not have STS risk scores were cardiogenic shock and clinical status (elective, urgent, emergent, emergent salvage; Table 3). Receiver operating characteristic curves and calibration plots are shown in Figure 1.

Risk Models and ML Methodology Generalize to Cohorts From 2 Different Institutions

We assessed the generalizability of the 4 MGH-trained predictive models using STS data from BWH and JHH (Table 2). For BWH cases, the highest AUC was 0.81 (RF). For JHH cases, the highest AUCs were 0.79 (Log-Reg/RF; Table 2). For BWH-specific risk models trained on BWH data and JHH-specific risk models trained on JHH data, LogReg/RF/XGBoost had the highest AUCs of 0.81 and 0.82, respectively (Figure 2).

For BWH, AUCs were not significantly different for all models trained on BWH data, compared with models trained on MGH data (Table 2). For JHH, however, AUCs were significantly higher for models trained on JHH data, compared with models trained on MGH data (Table 2). Full classification performance metrics of each model are presented in Tables E5-E9.

A Parsimonious Model Selects a Subset of Variables and Maintains Good Discriminatory Performance and Calibration

To assess the relationship between number of predictive variables and performance, we trained a parsimonious model on the other case types at MGH using a subset of variables

selected via forward feature selection. The mean AUC across test set data reached an asymptote at 0.82 with 8 variables (Figure 3, A, Table 4). The model achieved similar discrimination but was less well calibrated than the nonparsimonious models, with a Brier score of 0.05, CS of 1.34, and E/O of 0.77. When this parsimonious model was externally validated on the BWH and JHH cohorts, the AUCs were both 0.79 (Figure 3, B-C). The CS was 1.25 and E/O was 0.88 in the BWH cohort, but the CS was 1.63 and E/O was 0.57 in the JHH cohort.

DISCUSSION

This study shows novel risk classification and probability estimation models for heterogeneous cardiac surgical procedures. Furthermore, we show institution-specific risk predictions by training and applying these models to institution-specific data, which could improve the accuracy of preoperative informed consent. Such an approach provides clinicians with new tools to stratify risk and facilitate clinical decision-making across the full spectrum of cardiac surgical procedures. This might prove valuable in selecting recommendations among interventional options, a particularly important issue in the current era with the rapid development of multiple complementary treatment modalities including open and catheter-based technologies. It will also enable quality assessment and quality improvement efforts focused on these complex procedures.

There are many advantages to models trained on the national STS database, which are designed for frequent cardiac surgical procedures, such as CABG surgery, valve surgery, or both. Our intent was not to directly compare to the national STS models. In this present study, we focused specifically on procedures with no STS risk scores. The published performance for the STS risk models, however, can serve as a useful benchmark of ballpark performance for this work. The STS risk models have a summary AUC of 0.77 in a meta-analysis of 22 studies[11] and AUC of 0.75-0.79 (STS 2008 model) to 0.76-0.80 (STS 2018 model).[2] Our models achieved a highest mean AUC of 0.83.

The top 5 variables that predicted operative mortality differed slightly according to model and institution. However, the most consistent top variables across model and institution were cardiogenic shock and clinical status (Table 3). Accurately predicting postoperative outcomes and identifying factors that affect mortality influence postoperative management by affecting the threshold to perform additional tests or interventions in patients predicted to be at high risk of mortality. Furthermore, in this era of advanced temporary mechanical support, these data might inform discussions and decision-making regarding type and duration of such support.

Institution-specific risk models are also a potentially powerful application of our approach. Additional variables, even if not recorded in the national STS database, could be added to a local institutional database, and models can be retrained on these data to investigate if such variables contribute to risk prediction. We have shown a flexible platform for prediction that, with enough cases, can incorporate institution-specific and surgeon-specific “intuition” about planned procedures. The institution-specific risk models described herein enable such modifications to existing national risk models. This approach can also be used to train

models for a smaller subset of procedures (eg, procedures performed by each surgeon at an institution), larger superset of procedures (eg, in the national STS), or a specific group of procedures (eg, aortic procedures).

Benefits of Plurality of Models, Different Performance of Models, and Selection of the Optimal Model for the Specific Institution

Because of the ubiquity of linear models in surgical outcomes research, in particular logistic regression,[12] we used a logistic regression model with elastic net regularization (LogReg), as a baseline model against which to compare a linear model (SVM) and 2 nonlinear models (RF/XGBoost). The current STS calculator uses logistic regression.[2] The best model might differ from institution to institution, in a different subset or superset analysis, or if other variables are used. This highlights the importance of rigorously evaluating more than 1 model because of heterogeneity of data, variables, and patient case mix of each institution.

Generalizability of the MGH model.—We applied MGH-trained models to predict operative mortality for BWH and JHH cohorts. The MGH-trained XGBoost risk model achieved the highest or comparably highest AUCs for BWH and JHH. The performance of all 4 models trained on BWH data was similar to models trained on MGH data and applied to the BWH cohort. However, for JHH, LogReg and SVM models trained on JHH data significantly outper-formed models trained on MGH data and applied to the JHH cohort. These results suggest institution-specific risk models can sometimes outperform a general model applied to a cohort from a different institution. However, the differences in performance are small, suggesting our cross-validation approach prevents models from overfitting to data on which they were trained, and enables generalization to cohorts from other institutions. On the flip side, the fact that RF and XGBoost models trained in MGH perform equally well when applied on another institution in another state (JHH), shows the potential superiority of tree-based methods in capturing complex interactions between variables compared with simpler models like LogReg and SVM.

Parsimonious model.—Compared with the model trained on all preoperative features, a parsimonious model using 7 features selected via forward feature selection achieved lower discriminative performance and underestimated mortality (MGH, JHH), compared with the full models. However, a parsimonious model with fewer variables might help predict operative mortality in scenarios in which limited patient data are available and might provide an easy-to-use screening risk calculator for busy clinicians who cannot curate and enter the full set of STS input variables. The variables selected via forward feature selection overlapped with the most predictive features as ranked in the models developed using the full set of variables.

Discrimination versus calibration.—For clinical applications, prediction models should have adequate (1) discrimination (eg, they can be used to discern between patients who will have vs will not have an event), and (2) calibration (eg, they accurately predict absolute risk).[13] To assess discrimination of our binary classifiers, we report receiver operating characteristics and their associated AUCs (also known as “c-statistic”). To assess calibration, we report calibration curves (actual vs estimated risk), CSs, and E/O ratios.

We also performed Platt scaling[10] on the final predicted probabilities if doing so improved the aforementioned calibration metrics. We found the discriminatory ability of our models to be on par, if not better than, reported risk prediction models in postoperative outcomes literature. Moreover, the calibration of our models before Platt scaling was excellent in all 3 cohorts, meaning on average, across many different risk groups, the estimated posterior probability of postoperative mortality equaled the actual prevalence of mortality. One exception was RF, which, even after Platt scaling, slightly underpredicted mortality in lower-risk patients and slightly overpredicted mortality in higher-risk patients. We note the importance of calibration in the context of a surgeon discussing surgery with a patient.

Label imbalance.—In the MGH cohort, 424 mortality events occurred (5.5%) of 7745 cases (Table 1). This imbalance in labels can pose a challenge to the training of supervised learning algorithms. Because of the low prior probability of the outcome in the minority class, models can be biased toward the majority class, perform poorly for the minority class, yet achieve high overall accuracy.[14,15] To investigate if label imbalance detracted from model performance, we implemented the synthetic minority oversampling technique to generate synthetic new training samples for the minority class by randomly interpolating between existing minority samples and their neighbors.[16] This method can improve performance in the setting of label imbalance, but it did not in our experiments across a variety of classification metrics including area under the precision-recall curve and geometric mean of the sensitivity and specificity scores (data not shown). The lack of improvement conferred by the synthetic minority oversampling technique suggests patients in the minority class are sufficiently close to each other in the high-dimensional feature space and sufficiently far from patients in the majority class for the decision boundary to be adequately learned, despite the imbalance of labels. This is also consistent with the consistent discrimination performance achieved by models trained at MGH and applied to BWH and JHH data. Other methods have been developed to more robustly train models using imbalanced data, and the application of these techniques to improve the prediction of cardiac surgical outcomes is an important area of future work.[17,18]

Study Limitation

In this study, we trained and evaluated 4 supervised ML algorithms (LogReg, SVM, RF, and XGBoost). Other ML algorithms can be used to predict operative mortality of patients who undergo cardiac surgeries that lack STS risk scores, including but not limited to multilayer perceptron,[19] K-nearest neighbors, naive Bayes classifier, quadratic discriminant analysis, and adaptive boosting,[20] but are beyond the scope of this present study. Another limitation of this study is that we only used variables that were clinically meaningful in the STS models version 2.9, the STS classification that was current at the time of the work.[6] Additional variables in the newer STS version 4.20, or other confounders that are not recorded in these variables, might affect the prediction of operative mortality after cardiac surgery. This is reflected in slight differences in external validation results at 2 different hospitals in different states.

CONCLUSIONS

We developed well calibrated risk models to predict operative mortality of patients who undergo procedures for which risk scores have not yet been developed (Figure 4). Another contribution is validation of these methods and models, using data from multiple institutions. To enable our systematic approach to model-building at other institutions, we make our code publicly available to readers who might wish to develop risk prediction models calibrated to their own data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

A.D. A. and M.B.W. acknowledge funding from Controlled Risk Insurance Company/Risk Management Foundation. A.D.A. was also supported for this work by the MGH Hassenfeld Award. M.B.W. was supported by the Glenn Foundation for Medical Research and the American Federation for Aging Research through a Breakthroughs in Gerontology Grant; through the American Academy of Sleep Medicine through an AASM Foundation Strategic Research Award; by the Football Players Health Study at Harvard University; from the Department of Defense through a subcontract from Moberg ICU Solutions, Inc, and by grants from the National Institutes of Health (1R01NS102190, 1R01NS102574, 1R01NS107291, and 1RF1AG064312). C.S.O., P.M., and N.M. acknowledge support from the Massachusetts General Hospital Corrigan Minehan Heart Center.

The authors thank the following persons for institutional STS data collection: Laura Collier (MGH), Christine Smith (MGH), Christine Tetrault-Angelini (MGH), Linda Denning (BWH), Diane Alejo (JHH), Joseph Dinatale (JHH), and Kimberly Behrens (JHH). Images, logos, and icons used in the Central Image are licensed from shutterstock.com or obtained from clipartsworld.com, iconfinder.com, freeiconspng.com, flaticon.com, Wikipedia, and the respective institutions (MGH, BWH, JHH).

Abbreviations and Acronyms

AUC	area under receiver operating characteristic curve
BWH	Brigham and Women’s Hospital
CABG	coronary artery bypass graft
CS	calibration slope
E/O	expected-to-observed event
JHH	Johns Hopkins Hospital
LogReg	logistic regression with elastic net regularization
MGH	Massachusetts General Hospital
ML	machine learning
RF	random forest
STS	Society of Thoracic Surgeons
SVM	support vector machine

XGBoost	extreme gradient boosted trees

References

1. Nilsson J, Algotsson L, Höglund P, Lührs C, Brandt J. Comparison of 19 preoperative risk stratification models in open-heart surgery. Eur Heart J. 2006;27:867–74. [PubMed: 16421172]
2. O'Brien SM, Feng L, He X, Xian Y, Jacobs IP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 2-statistical methods and results. Ann Thorac Surg. 2018;105:1419–28. [PubMed: 29577924]
3. Kilic A, Goyal A, Miller JK, Gjekmarkaj E, Tam WL, Gleason TG, et al. Predictive utility of a machine learning algorithm in estimating mortality risk in cardiac surgery. Ann Thorac Surg. 2020;109:1811–9. [PubMed: 31706872]
4. D'Agostino RS, Jacobs JP, Badhwar V, Fernandez FG, Paone G, Wormuth DW, et al. The Society of Thoracic Surgeons adult cardiac surgery database: 2018 update on outcomes and quality. Ann Thorac Surg. 2018;105:15–23. [PubMed: 29233331]
5. Shahian DM, Jacobs JP, Badhwar V, Kurlansky PA, Furnary AP, Cleveland JC Jr, 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. [PubMed: 29577925]
6. Society of Thoracic Surgeons. About the STS Risk Calculator v2.9. Available at: https://riskcalc.sts.org/stswebriskcalc/assets/About%20the%20STS%20Risk%20Calculator%20v29.pdf. Accessed September 27, 2021.
7. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery; 2016. 785–94.
8. King RD, Orhobor OI, Taylor CC. Cross-validation is safe to use. Nat Mach Intell. 2021;3:276.
9. Scikit-learn. Nested versus non-nested cross-validation. Available at: https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html. Accessed September 27, 2021.
10. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: International Conference on Machine Learning. New York, NY: Association for Computing Machinery; 2005:625–32.
11. Sullivan PG, Wallach JD, Ioannidis JP. Meta-analysis comparing established risk prediction models (EuroSCORE II, STS Score, and ACEF Score) for perioperative mortality during cardiac surgery. Am J Cardiol. 2016;118:1574–82. [PubMed: 27687052]
12. Normand SL. Some old and some new statistical tools for outcomes research. Circulation. 2008;118:872–84. [PubMed: 18711024]
13. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA. 2017;318:1377–84. [PubMed: 29049590]
14. Ishwaran H, Blackstone EH. Commentary: dabblers: beware of hidden dangers in machine-learning comparisons. J Thorac Cardiovasc Surg. August 31, 2020 [Epub ahead of print].
15. Ishwaran H, O'Brien R. Commentary: the problem of class imbalance in biomedical data. J Thorac Cardiovasc Surg. 2021;161:1940–1. [PubMed: 32711988]
16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
17. O'Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced data. Pattern Recognit. 2019;90:232–49. [PubMed: 30765897]
18. Dai W, Ng K, Severson K, Huang W, Anderson F, Stultz C. Generative oversampling with a contrastive variational autoencoder. In: 2019 IEEE International Conference on Data Mining (ICDM). Los Alamitos, CA: IEEE Computer Society; 2019:101–9.
19. Bakshi K The Universal Approximation Theorem for neural networks. Available at: https://www.techleer.com/articles/449-the-universal-approximation-theorem-for-neural-networks/. Accessed September 27, 2021.

20. Scikit-learn. Classifier comparison. Available at: https://scikit-leam.org/stable/auto_examples/classification/plot_classifier_comparison.html. Accessed September 27, 2021.

CENTRAL MESSAGE

Machine learning models can predict operative mortality at a high level of accuracy for cardiac surgical procedures without STS scores.

PERSPECTIVE

Accurate risk assessment before cardiac surgery is essential but current models are limited to a subset of frequently performed procedures. We used supervised machine learning to predict operative mortality for an expanded set of cardiac surgical procedures without STS scores.

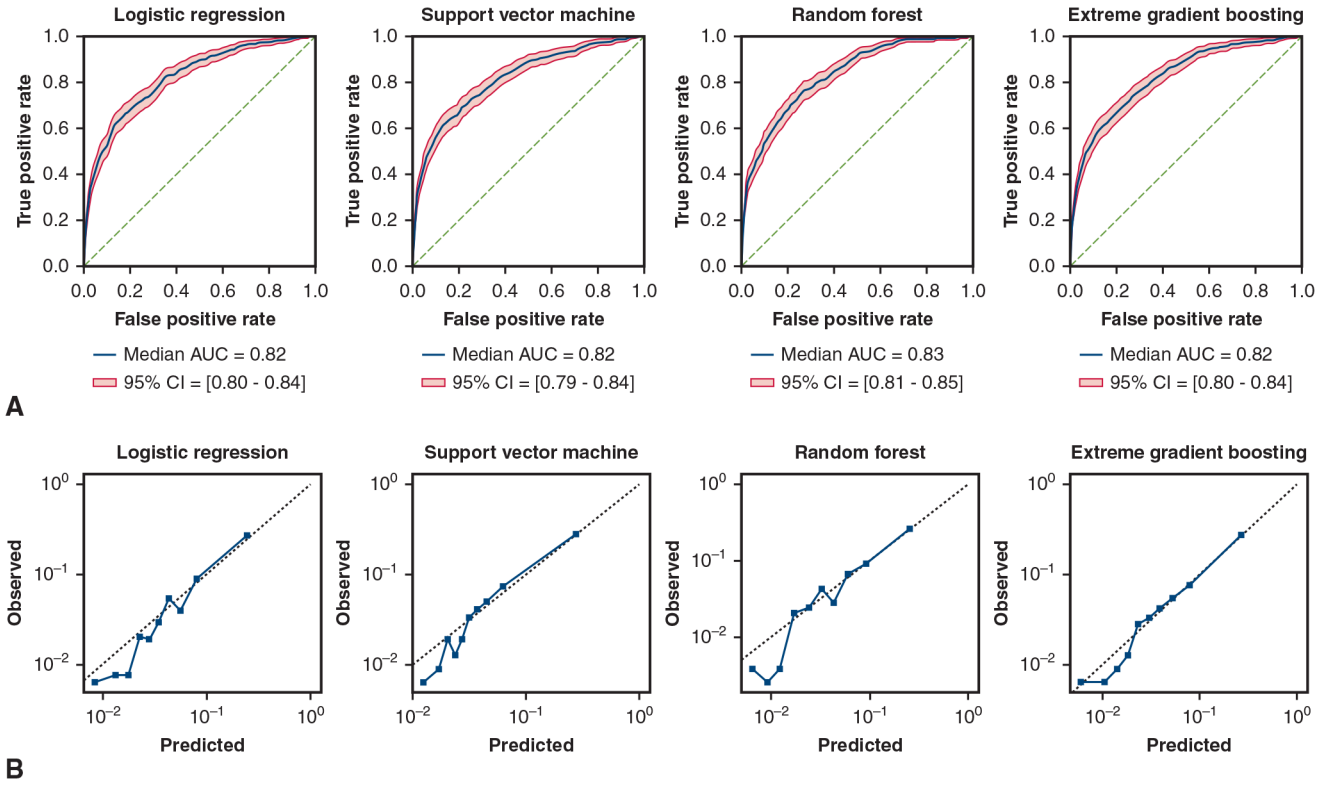

FIGURE 1.

Receiver operating characteristic (A) and calibration curves (B) of mortality risk models for cardiac surgical procedures without Society of Thoracic Surgeons risk scores at the Massachusetts General Hospital using preoperative variables. *AUC*, Area under the receiver operating characteristic curve; *CI*, confidence interval.

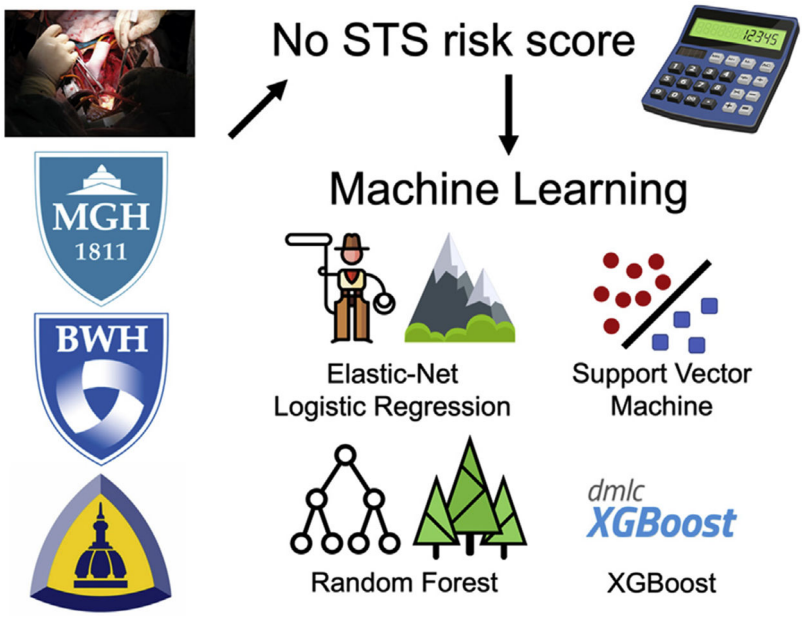

FIGURE 2.
Machine learning can accurately predict mortality for procedures without STS scores.

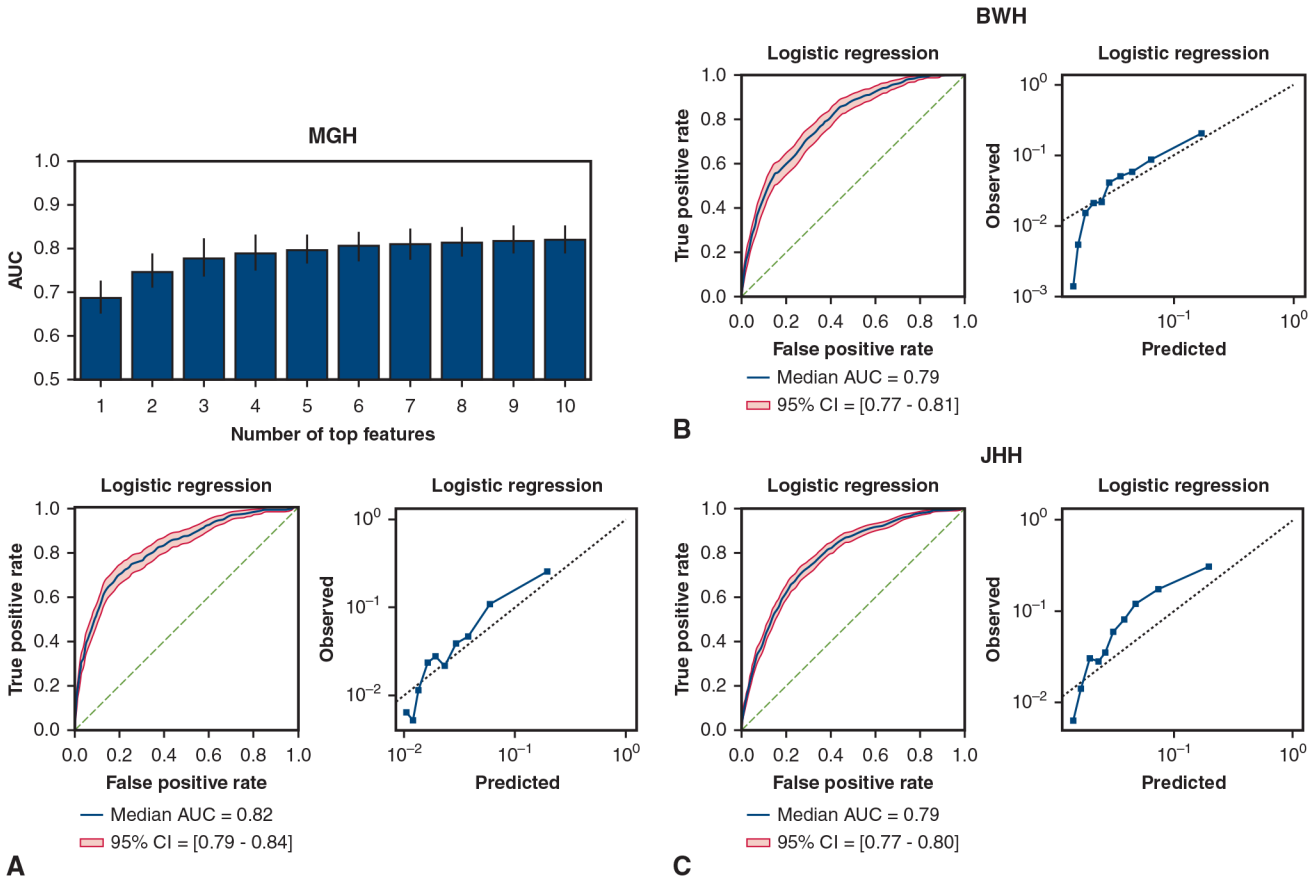

FIGURE 3.

Parsimonious model trained on subset of preoperative variables from the Massachusetts General Hospital (*MGH*; A; top variables, receiver operating characteristic and calibration curves shown). MGH-trained model applied to the Brigham and Women's Hospital (*BWH*) cohort (B) and the Johns Hopkins Hospital (*JHH*) cohort (C); receiver operating characteristic and calibration curves shown. *AUC*, Area under the receiver operating characteristic curve; *CI*, confidence interval.

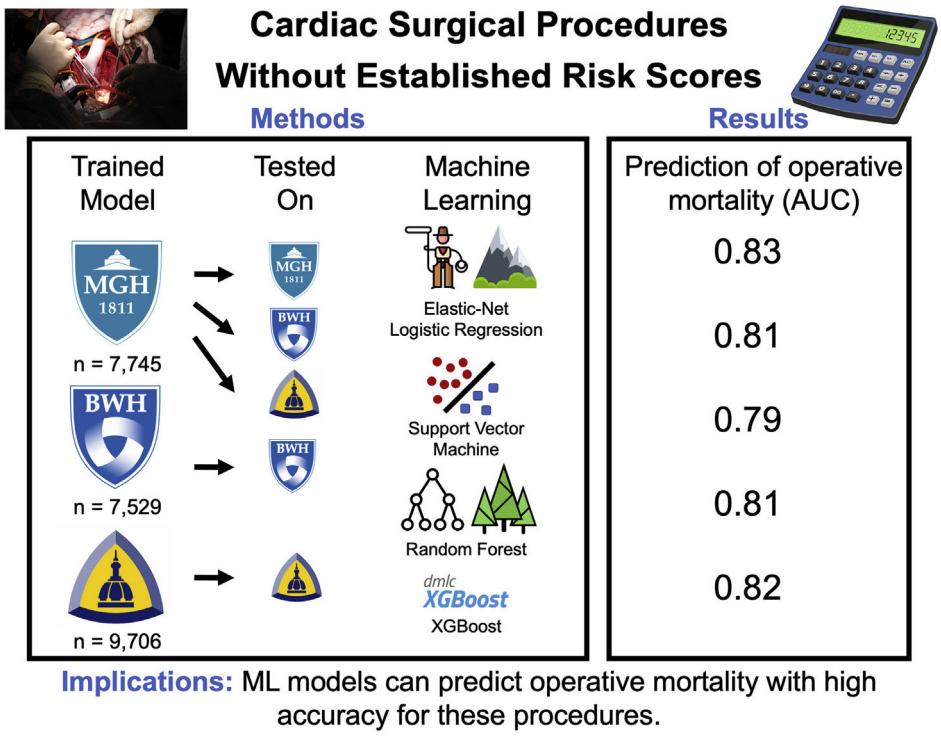

FIGURE 4.

Machine learning (*ML*) models can predict operative mortality at a high level of accuracy for cardiac surgical procedures without established risk scores. *MGH*, Massachusetts General Hospital; *BWH*, Brigham and Women's Hospital; *AUC*, area under the receiver operating characteristic curve; *XGBoost*, extreme gradient boosted trees.

TABLE 1.

Patient characteristics of the MGH cohort (n = 7745)

Variable	Operative mortality		
	No (n = 7321)	Yes (n = 424)	*P* value
Sex, n (%)			.428
Male	4433 (60.6)	248 (58.5)	
Female	2888 (39.4)	176 (41.5)	
Median age (IQR), y	64 (52-74)	67 (58-77)	<.001
Race, n (%)			<.001
Caucasian	6466 (89.1)	343 (82.5)	
Black	215 (3.0)	26 (6.2)	<.001
Asian	157 (2.2)	12 (2.9)	.422
American Indian/Alaskan Native	14 (0.2)	0 (0)	1.000
Native Hawaiian/Pacific Islander/Other	278 (3.8)	20 (4.8)	.383
Hispanic, Latino, or Spanish ethnicity, n (%)	300 (4.1)	24 (5.7)	.001
Median weight (IQR), kg	79.3 (67.0-91.8)	77.0 (64.8-90.3)	.013
Median height (IQR), cm	170.2 (162.6-178.0)	168.0 (160.0-177.0)	.003
Hypertension, n (%)	4631 (63.3)	297 (70.0)	<.001
Diabetes, n (%)	1507 (20.6)	135 (31.8)	<.001
Median last creatinine level (IQR), mg/dL	1.1 (0.9-1.3)	1.3 (1.0-1.9)	<.001
Renal failure undergoing dialysis, n (%)	115 (1.6)	55 (13.0)	<.001
Chronic lung disease, n (%)			<.001
No	6009 (82.3)	302 (71.4)	
Mild	764 (10.5)	61 (14.4)	
Moderate	247 (3.4)	20 (4.7)	
Severe	242 (3.3)	36 (8.5)	
Documented, unknown severity	39 (0.5)	4 (0.9)	
Immunocompromised state, n (%)	577 (7.9)	58 (13.7)	<.001
Peripheral arterial disease, n (%)	874 (11.9)	84 (19.8)	<.001
Cerebrovascular disease, n (%)	1462 (20.0)	114 (26.9)	.001
Previous cerebrovascular accident, n (%)	702 (9.6)	65 (15.3)	<.001

Variable	Operative mortality		*P* value
	No (n = 7321)	Yes (n = 424)	
Previous cardiac interventions, n (%)	3069 (42.0)	262 (61.8)	<.001
Cardiogenic shock, n (%)*			<.001
Yes	115 (1.6)	101 (23.8)	
No	7106 (97.2)	284 (67.0)	
Yes, at the time of the procedure	84 (1.1)	35 (8.3)	
Yes, not at the time of the procedure but within the previous 24 h	7 (0.1)	4 (0.9)	
Preoperative CPR, n (%)*			<.001
Yes	22 (0.3)	18 (4.2)	
No	7284 (99.6)	401 (94.6)	
Yes, within 1 h of the start of the procedure	3 (0.0)	2 (0.5)	
Yes, more than 1 h but less than 24 h of the start of the procedure	2 (0.0)	3 (0.7)	
NYHA classification, n (%)			<.001
Class I	792 (11.4)	28 (6.7)	
Class II	1100 (15.9)	39 (9.4)	
Class III	1150 (16.6)	65 (15.7)	
Class IV	1098 (15.9)	184 (44.3)	
Inotropes within 48 h, n (%)	471 (6.5)	164 (38.7)	<.001
Steroids within 24 h, n (%)	316 (4.3)	39 (9.2)	<.001
Glycoprotein IIb/IIIa inhibitor within 24 h, n (%)	19 (0.3)	5 (1.2)	.003
ADP inhibitors within 5 d, n (%)	75 (1.0)	24 (5.7)	<.001
Median last hematocrit (IQR), %	38.8 (33.3-42.4)	32.8 (27.7-38.7)	<.001
Median last white blood cell count (IQR), × 10^9/L	7.1 (5.9-8.91	8.5 (6.4-14.4)	<.001
Median last platelet count (IQR), × 10^9/L	210 (167-259)	173 (112-244)	<.001
Diseased major coronary arteries, n (%)			<.001
None	4737 (68.8)	194 (49.0)	
1	640 (9.3)	45 (11.4)	
2	488 (7.1)	46 (11.6)	
3	1023 (14.9)	111 (28.0)	
Median ejection fraction (IQR), %	60 (50-67)	54 (25-64)	<.001
Aortic stenosis, n (%)	2391 (32.8)	118 (28.0)	.091

Variable	Operative mortality		
	No (n = 7321)	Yes (n = 424)	*P* value
Mitral stenosis, n (%)	503 (6.9)	33 (7.8)	.374
Aortic valve insufficiency, n (%)			.015
None	2393 (32.8)	119 (28.1)	
Trivial/trace	1685 (23.1)	113 (26.7)	
Mild	1754 (24.1)	121 (28.6)	
Moderate	724 (9.9)	41 (9.7)	
Severe	590 (8.1)	20 (4.7)	
Mitral valve insufficiency, n (%)			<.001
None	790 (10.8)	41 (9.7)	
Trivial/trace	2071 (28.4)	82 (19.3)	
Mild	1954 (26.8)	126 (29.7)	
Moderate	1205 (16.5)	105 (24.8)	
Severe	1171 (16.0)	61 (14.4)	
Tricuspid valve insufficiency, n (%)			<.001
None	1008 (13.8)	45 (10.6)	
Trivial/trace	2464 (33.8)	76 (17.9)	
Mild	2031 (27.8)	131 (30.9)	
Moderate	1090 (14.9)	103 (24.3)	
Severe	514 (7.0)	56 (13.2)	
Preoperative status, n (%)			<.001
Elective	4465 (61.1)	99 (23.3)	
Urgent	2548 (34.9)	178 (42.0)	
Emergent	291 (4.0)	125 (29.5)	
Emergent salvage	6 (0.1)	22 (5.2)	
Incidence of cardiac surgery, n (%)			<.001
First	4613 (63.4)	242 (57.3)	
First reoperation	1410 (19.4)	100 (23.7)	
Second reoperation	267 (3.7)	28 (6.6)	
Third reoperation	61 (0.8)	7 (1.7)	
Fourth or more reoperation	25 (0.3)	4 (0.9)	

Variable	Operative mortality		
	No (n = 7321)	Yes (n = 424)	*P* value
Median cross-clamp time (IQR), minutes	87 (23-127)	52.5 (0-150)	<.001
Median cardiopulmonary bypass time (IQR), minutes	123 (77-170)	139.0 (0-237)	.008
CABG performed, n (%) *			.002
Yes	595 (8.1)	48 (11.3)	
No	6437 (87.9)	349 (82.3)	
Yes, planned	278 (3.8)	25 (5.9)	
Yes, unplanned because of surgical complication	3 (0.0)		
Yes, unplanned because of unsuspected disease or anatomy	6 (0.1)	2 (0.5)	
Valve procedure performed, n (%)	4577 (62.5)	190 (44.8)	<.001

IQR, Interquartile range; *CPR*, cardiopulmonary resuscitation; *NYHA*, New York Heart Association; *ADP*, adenosine diphosphate receptor inhibitors; *CABG*, coronary artery bypass graft.

*The categories for these variables were changed in the more recent STS data specifications to increase granularity, resulting in multiple categories for the “Yes” group. These “Yes” categories are combined into a single group, if the variable is subsequently used for ML.

TABLE 2.

Comparison of model performances by AUC

Model	Trained model: MGH Tested on: MGH	Trained model: MGH Tested on: BWH	Trained model: BWH Tested on: BWH	*P* value (MGH trained model vs BWH trained model)	Trained model: MGH Tested on: JHH	Trained model: JHH Tested on: JHH	*P* value (MGH trained model vs JHH trained model)
LogReg	0.82 (0.80-0.84)	0.80 (0.78-0.82)	0.81 (0.79-0.83)	.75	0.79 (0.78-0.81)	0.82 (0.80-0.83)	**.04**
SVM	0.82 (0.79-0.84)	0.78 (0.75-0.80)	0.79 (0.77-0.81)	.34	0.76 (0.74-0.78)	0.81 (0.79-0.82)	<**.01**
RF	0.83 (0.81-0.85)	0.81 (0.79-0.83)	0.81 (0.79-0.83)	.88	0.79 (0.78-0.81)	0.82 (0.81-0.83)	<**.01**
XGBoost	0.82 (0.80-0.84)	0.79 (0.77-0.81)	0.81 (0.79-0.83)	.09	0.77 (0.75-0.78)	0.82 (0.80-0.83)	<**.01**
P value (LogReg vs SVM)	.36	<**.01**	**.01**		<**.01**	<**.01**	
P value (LogReg vs RF)	.09	.39	.59		.75	.16	
P value (LogReg vs XGBoost)	.80	.09	.25		<**.01**	.51	

Bold values indicate statistically significant results ($P < .05$). *MGH*, Massachusetts General Hospital; *BWH*, Brigham and Women's Hospital; *JHH*, Johns Hopkins Hospital; *LogReg*, logistic regression with elastic net regularization; *SVM*, support vector machine; *RF*, random forest; *XGBoost*, extreme gradient boosted trees.

TABLE 3.

Top 5 important variables selected by models trained on preoperative variables

Model		MGH		BWH		JHH
LogReg	**1**	Cardiogenic shock	**1**	Elective procedure	**1**	Elective procedure
	2	Inotropes within 48 h	**2**	Cardiogenic shock	**2**	Emergent salvage procedure
	3	Elective procedure	**3**	Emergent salvage procedure	**3**	Cardiogenic shock
	4	Emergent procedure	**4**	Emergent procedure	**4**	Preoperative CPR
	5	Age	**5**	Previous cardiovascular interventions	**5**	NYHA class IV
SVM	**1**	Elective procedure	**1**	Elective procedure	**1**	Elective procedure
	2	Cardiogenic shock	**2**	Urgent procedure	**2**	Emergent salvage procedure
	3	Urgent procedure	**3**	Emergent salvage procedure	**3**	Cardiogenic shock
	4	No infective endocarditis	**4**	Cardiogenic shock	**4**	Urgent procedure
	5	Inotropes within 48 h	**5**	No infective endocarditis	**5**	Preoperative CPR
RF	**1**	Cardiogenic shock	**1**	Last creatinine level	**1**	Elective procedure
	2	Last WBC count	**2**	Last WBC count	**2**	Last creatinine level
	3	Last creatinine level	**3**	Last hematocrit	**3**	Emergent salvage procedure
	4	Emergent procedure	**4**	Weight	**4**	Cardiogenic shock
	5	Inotropes within 48 h	**5**	Age	**5**	Ejection fraction
XGBoost	**1**	Cardiogenic shock	**1**	Elective procedure	**1**	Cardiogenic shock
	2	Elective procedure	**2**	Cardiogenic shock	**2**	Elective procedure
	3	Inotropes within 48 h	**3**	Last creatinine level	**3**	Emergent salvage procedure
	4	Urgent procedure	**4**	Emergent procedure	**4**	NYHA class IV
	5	Dialysis	**5**	Previous cardiovascular interventions	**5**	Last creatinine level

MGH, Massachusetts General Hospital; *BWH*, Brigham and Women's Hospital; *JHH*, Johns Hopkins Hospital; *LogReg*, logistic regression with elastic net regularization; *CPR*, cardiopulmonary resuscitation; *NYHA*, New York Heart Association; *SVM*, support vector machine; *RF*, random forest; *WBC*, white blood cell; *XGBoost*, extreme gradient boosted trees.

TABLE 4.

Preoperative variables added via forward feature selection to parsimonious model trained on MGH others cohort

Rank	Mean AUC ± SD	Variable
1	0.69 ± 0.04	Elective procedure
2	0.75 ± 0.04	Last creatinine level
3	0.78 ± 0.04	Urgent procedure
4	0.79 ± 0.04	Previous cardiovascular interventions
5	0.8 ± 0.03	Heart failure
6	0.81 ± 0.03	Chronic lung disease
7	0.81 ± 0.03	Inotropes within 48 h
8	0.82 ± 0.03	Sex
9	0.82 ± 0.03	No significant coronary obstructive disease
10	0.82 ± 0.03	Cardiogenic shock

The variable at a given rank indicates a logistic regression model trained with this variable as well as all previous variables (eg, rank 3 indicates a model was trained with the “urgent procedure,” “last creatinine level,” and “elective procedure” variables) that achieved the highest AUC compared to any other set of that number of variables. *AUC*, Area under the receiver operating characteristic curve; *SD*, standard deviation.