Title: Machine Learning Models for Mitral Valve Replacement:

A Comparative Analysis with the Society of Thoracic Surgeons Risk Score

Running Title: Machine Learning for MVS

Authors: Agni Orfanoudaki[1], PhD; Aikaterini Giannoutsou[2], BSc; Sabet Hashim[3], MD, FACS; Dimitris Bertsimas[4,5], PhD; Robert Carl Hagberg[3], MD.

Affiliations:

[1]Saïd Business School, University of Oxford, Park End Street, Oxford, UK, OX1 1HP

[2] USC Marshall School of Business, University of South California, 3670 Trousdale Pkwy, Los Angeles, CA, 90089

[3] Hartford HealthCare Heart & Vascular Institute, Hartford HealthCare, 80 Seymour Street, Hartford, CT, 06102

[4] Operations Research Center, Massachusetts Institute of Technology, Muckley Bldg, 1 Amherst St, Cambridge, MA, 02142

[5]Sloan School of Management, Massachusetts Institute of Technology, E62-560, Cambridge, MA, 02139

Keywords: mitral valve surgery; machine learning; artificial intelligence; heart.

Meeting Presentation:

STS 56th Annual Meeting, New Orleans, Louisiana, January 25-28, 2020

Funding and Disclosures:

Study funded by a grant to the Massachusetts Institute of Technology from Hartford HealthCare. The authors report no competing interests.

Word count: 4198

Abstract Word Count: 250

Corresponding Author:

Agni Orfanoudaki

Saïd Business School, University of Oxford

E1.33, Park End Street, Oxford, UK, OX1 1HP.

Email: agni.orfanoudaki@sbs.ox.ac.uk

Phone: 617-586-9202

Fax: 617-253-3601

Abstract

Background: Current Society of Thoracic Surgery (STS) risk models for predicting outcomes of mitral valve surgery (MVS) assume a linear and cumulative impact of variables. We evaluated post-operative MVS outcomes and designed mortality and morbidity risk calculators to supplement the STS risk score.

Methods: Data from the STS Adult Cardiac Surgery Database for MVS was used from 2008-2017. The data included 383,550 procedures and 89 variables. Machine learning (ML) algorithms were employed to train models in order to predict postoperative outcomes for MVS patients. Each model's discrimination and calibration performance were validated using unseen data against the STS risk score.

Results: Comprehensive mortality and morbidity risk assessment scores were derived from a training set of 287,662 observations. The Area Under the Curve (AUC) for mortality ranged from 0.77 to 0.83, leading to a 3% increase in predictive accuracy compared to the STS score. Logistic Regression and eXtreme Gradient Boosting achieved the highest AUC for prolonged ventilation (0.82) and deep sternal wound infection (0.78 and 0.77) respectively. EXtreme Gradient Boosting performed the best with an AUC of 0.815 for renal failure. For permanent stroke prediction all models performed similarly with an AUC around 0.67. The ML models led to improved calibration performance for mortality, prolonged ventilation, and renal failure, especially in cases of reconstruction/repair and replacement surgery.

Conclusions: The proposed risk models complement existing STS models in predicting mortality, prolonged ventilation, and renal failure, allowing healthcare providers to more accurately assess a patient's risk of morbidity and mortality when undergoing MVS.

Abbreviations Table:

Abbreviations	Meaning
ACSD	Adult Cardiac Surgery Database
ARDS	Adult Respiratory Distress Syndrome
AUC	Area Under ROC Curve
CART	Classification and Regression Trees
IABP	Intra-Aortic Balloon Pump procedure
Log.Reg	Multivariate Logistic Regression
L-OCT	Multivariate Logistic Regression with Optimal Classification Trees
ML	Machine Learning
MVS	Mitral Valve Surgery
OCT	Optimal Classification Trees
RF	Random Forest
ROC	Receiver Operator Characteristic
SD	Standard Deviation
STS	Society of Thoracic Surgeons
XGBoost	eXtreme Gradient Boosting

Introduction

Risk stratification has become a critical element in the practice of cardiac surgery, including mitral valve surgery (MVS), due to the risk of death and morbidity from intraoperative and postoperative complications.[1, 2] The risk/benefit ratio of surgery is sometimes difficult to predict and the decision to proceed with surgery on an individual basis can be complex.[3] Achieving a better understanding of preoperative risk factors and their interactions is key to reducing morbidity and mortality from MVS.[4]

The Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database (ACSD) and its risk calculator have been the state-of-the-art risk model for predicting operative mortality and morbidity after adult cardiac surgery since the early 1980s.[4] This model is used by physicians and patients as a tool for understanding the possible risks of surgery. As of November 15, 2018, in response to evolving changes in patient characteristics, risk profiles, surgical practice, and outcomes, the STS released an updated short-term risk calculator.[5, 6]

Even though the new calculator incorporates updated and richer datasets, it is still based on traditional statistical methods.[7] While useful, it assumes that the variables in the model interact in a linear and additive fashion. The mathematical and medical realities, however, suggest that the interaction of risk factors and markers of disease acuity may be far from linear, and that some variables gain or lose significance due to the absence or presence of other variables (see example in Supplemental Material).[8, 9]

Over the past years, the rapid increase of computational power has allowed the implementation of linear algebraic data analysis techniques. It has also led to the development and widespread use of more complex statistical algorithms that are called machine learning (ML) methods. ML algorithms use large datasets to establish powerful predictive models. Prior research has demonstrated their success in various segments of the medical field, including cardiac surgery.[10-13] Several studies have shown the superiority of ML methods over traditional logistic regression methods, such as the STS risk model.[14-16]

In this paper, we combine big data from the well-validated, national ACSD with ML to design and test multiple risk calculators for predicting MVS mortality and morbidity. We present a combination of linear and non-linear methods that deliver higher accuracy than the latest STS model. We provide a quantitative comparison of their discrimination and calibration performance as well as a tree-based application amenable to automatic integration into electronic health records (EHR).

Materials and Methods

Sample Population

We used the STS ACSD database for the years 2008-2017 including only those patients undergoing isolated MVS (annuloplasty, reconstruction/repair, replacement) which accounted for 383,550 entries. Patients that had other concomitant procedures were excluded. All preoperative variables were used to design and train our models, while postoperative endpoints, such as mortality, are presented as the outcomes predicted. Our dataset included the following types of surgery MVS: (1) Mitral Valve Replacement (170,542 observations), (2) Complex Mitral Valve Reconstruction/Repair (176,740 observations), and (3) Mitral Valve Annuloplasty Only (36,268 observations).

For each model 287,662 (75% of the observations) patients were leveraged for model derivation and algorithm training and 95,888 (25% of the observations) were used for model testing and validation.

Data bootstrapping was used to ensure that the reported performance is not due to a single random split of the data, remaining consistent across multiple data partitions. Five random splits of the data were conducted to ensure statistical significance of model performance. At each random split, data were partitioned in training and testing sets. Each model's performance was evaluated on the testing set for the specific data split (see Supplemental Figure 1). The training and testing cohorts were not mixed at any point during the process. At each split, the same data partition was used for all outcomes of interest. The random splitting of the data resulted in

variances in the outcome incidence rates among the two populations. The observations in the testing set were filtered a posteriori to those for which the STS risk score was available allowing direct comparisons to the ML models.

Outcomes

We include five endpoints; Mortality, Prolonged Ventilation, Renal Failure, Deep Sternal Wound Infection, and Permanent Stroke. All endpoints are in line with the STS risk model definitions[17]. Detailed descriptions are available in the Supplemental Material. Table 1 describes the outcome rates on both the training and testing set.

Predictor Variables

The original dataset includes over 300 demographic and preoperative variables from versions 2.61, 2.73, and 2.81 of the STS database (Supplemental Table 1). Each variable was defined the same as the STS database. Discrepancies in the definitions of variables corresponding to different versions of the database had been consolidated by the STS in the provided dataset. We kept numerical variables, such as laboratory results, in their raw numeric form. The variables used to design our predictive models are collected on each visit of a patient for MVS and can be grouped into Demographics, Medications, and Medical Data Risk Factors. Patient sociodemographic characteristics included age, sex, and race. We included all variables that are present in the STS Risk Score Calculator,[5, 6] as well as additional variables, not part of the STS calculator, such as the systolic diameter of the left ventricle, presence of tricuspid valve disease, and pulmonary hypertension, as defined in the STS database. Higher proportion of missing values were systematically found in variables that were not consistently recorded in all versions of the data. We included all risk factors that are present in the existing STS score. New variables were incorporated for which at most 30% of their values were present. The final processed dataset comprises 89 features. A full list can be found in Supplemental Table 1.

Missing Data Imputation

Missing values were imputed using a recently developed and novel machine-learning method called Optimal Impute[18]. This approach has been shown in multiple real-world datasets to lead to significant improvements in prediction accuracy compared to classical missing values imputation methods, including multiple imputation with chained equations.[19-22] Further information regarding the missing data imputation process is provided in the respective section of the Supplemental Material and at Supplemental Tables 2-3.

Predictive Methods

We employed a collection of well-established ML methods for binary classification predicting both the mortality and morbidity outcomes; Multivariate Logistic Regression (Log.Reg), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Optimal Classification Trees (OCT), and a combination of multivariate Log.Reg and OCT (L-OCT) (see the Supplementary Material for further details)[23-28]. Grid search was applied to tune the model parameters, so as to avoid overfitting. Specifically, for every algorithm a set of potential values was defined for each parameter that needed to be tuned. Every potential combination of these values was evaluated using K-Fold cross validation and the one with the best performance was selected. The Supplemental Material provides a more detailed description of the grid search and k-fold cross validation processes. Once the parameters were chosen, each model was trained on the entire training set and evaluated on the testing set. See Supplemental Table 4 for a detailed list of the tuned parameters.

Measurement of Model Performance

Data bootstrapping resulted in multiple models for each combination of endpoint and binary classification method (see Supplemental Figure 1). Specifically, data was partitioned five independent times in the derivation and validation cohorts, six algorithms were applied, and five

outcomes were measured, resulting in 150 models. We report the average model performance for every metric across all random data splits, outcomes, and algorithms. Randomization allowed us to calculate confidence intervals along our evaluation metrics and test whether the reported results are statistically significant.

The model performance was measured in the testing datasets, and compared against the STS score. The ability of each model to discriminate between the outcomes of interest was measured using the c-statistic, also known as the Area Under the ROC Curve (AUC). Receiver Operator Characteristic (ROC) and calibration curves were derived to directly compare the STS model and the best performing ML model for each type of surgery and outcome. Additional precision recall, ROC, and calibration curves were also created to compare the sensitivity, specificity, and precision of each algorithm for all tasks.

Interactive Application for Interpretability

A dynamic online and phone application was developed as the user-friendly interface of the models for direct use by providers. The application presents the results in the form of an interactive questionnaire. A user, dealing with a specific MVS patient, is initially prompted to select an outcome of interest. The user is then asked a series of questions about the presence or absence of certain preoperative variables. Figure 1 provides a visualization of the interface.

Ethical Oversight

Institutional Review Board approval for the study was obtained.

Results

The mortality and renal failure outcomes are present in 6.34% and 6.15% of the overall cohort respectively. A smaller proportion of patients experience deep sternal wound infection (0.20%) and permanent stroke (2.18%) after the surgery. On the contrary, 20.66% of the observations

were associated with prolonged ventilation. 53.85% of the sample population is male. The mean (SD) of age is 64.76 (13.06). The majority of the population included in this study is Caucasian (83.67%). The mean patient height and weight are 169.87 (11.36) cm and 80.89 (19.68) kg respectively. Supplemental Table 5 provides an overview of the primary risk factors considered for each model.

Table 2 summarizes the average discrimination performance of each model for all considered outcomes across five data partitions, compared to that of the STS predictions. For all tasks, except deep sternal wound infection, XGBoost is the model with the highest performance with Log.Reg being either at the same level or slightly worse (see also Supplemental Table 6). For the mortality outcome, Log.Reg along with XGBoost have the highest c-statistics of 0.825 and 0.826 respectively, outperforming OCT (0.787), RF (0.77), and the STS c-statistic (0.796). L-OCT achieves a c-statistic of 0.809 on this task, higher than the STS and, at the same time, more interpretable than XGBoost. Similarly, for predicting prolonged ventilation, Log.Reg and XGBoost were again the best with c-statistics of 0.817 and 0.818 respectively. L-OCT was associated with slightly worse performance (0.808) along with OCT (0.792), again outperforming the STS score (0.788). For renal failure, XGBoost performed the best with a c-statistic of 0.815. All other methods' AUCs were up to two percentage points lower except RF which performed significantly lower (0.745). In the prediction of deep sternal wound Infection (0.15% incidence rate), XGBoost and Log.Reg were again the best (0.784 and 0.771 respectively) outperforming all other methods, including the STS (0.75). Finally, for the prediction of permanent stroke all models, performed similarly, with a c-statistic around 0.67. ROC and precision and recall curves are presented in Supplemental Figures 2-3.

In Figure 2, we focus our analysis on direct comparisons of the discrimination performance of the STS risk score and the best performing ML method for each outcome and type of MVS surgery. Notice that for the cases of reconstruction/repair and replacement the ML model outperforms the STS model across all tasks, other than the case of permanent stroke where the curves are almost

identical. Nevertheless, our analysis shows that for annuloplasty only the STS has a more accurate model with respect to deep sternal wound infection and an equivalent one for permanent stroke. For all other outcomes, ML provides a more accurate estimation model.

Figure 3 depicts the respective results in terms of calibration performance. Generally, we notice that the best ML model tends to overestimate the actual risk while the STS underestimates the observed outcomes. For the outcome of deep sternal wound infection, ML is better calibrated for reconstruction/repair and replacement surgeries but has poorer performance in the cases of annuloplasty. For the outcomes of mortality and renal failure, ML provides a clear edge over the STS across all types of surgery while the calibration results seem equivalent for the events of permanent stroke and prolonged ventilation. Supplemental Table 7 and Supplemental Figure 4 present additional calibration results related to all ML methods considered.

Our analysis showed the key risk factors for the outcomes considered. Supplemental Table 8 displays the most significant variables for the Log.Reg models. Figure 4 shows the most important risk factors for the XGBoost models.

We were able to gather insights from the OCT models, as only a small subset of the variables is used in the determination of a patient's risk profile. Figure 5 shows an example of a mortality prediction model using the OCT algorithm. It provides simple predictions about future outcomes of MVS patients with comparable performance (mortality: 0.787, prolonged ventilation: 0.794) to the STS model (mortality: 0.796, prolonged ventilation: 0.788). With at most four questions, the user is able to estimate the final risk.

Conclusions

Discussion

To the best of our knowledge, we present one of the first studies to use advanced ML methods for the prediction of surgical risk in terms of both mortality and morbidity in the context of MVS[29]. Leveraging 10 years of data from the STS ACSD database, we developed clinically actionable

prognostic tools for mortality and morbidity prediction using well-established binary classification algorithms to shed light on the calculation of risk. The proposed models lead to better discrimination and calibration performance compared to the existing STS risk calculator in terms of predicting mortality, prolonged ventilation, and renal failure, and especially in cases of reconstruction/repair and replacement surgery. Combining the power of big data from the largest national database and the innovative logic of ML, we have designed new models that offer some advantages of being evidence-based, non-linear, interactive, and amenable to direct medical application.

Our results highlight that the nature and prevalence of certain variables may significantly impact the accuracy of the predictive models in a uniform way. Our analysis demonstrates that mortality is easier to predict compared to the occurrence of permanent stroke across all methods considered. This might be due to the higher frequency of death after an MVS procedure versus permanent strokes or deep sternal wound infection, empowering the algorithms with a higher number of cases where the adverse event was observed. In tasks of low incidence rate, the nature of the outcome might affect the accuracy of the models and their downstream performance. Deep sternal would infection is generally more predictable than permanent stroke even though it is more rarely observed.

Ensemble adaptive methods such as XGBoost outperform the rest of the models, providing a significant edge over the baseline STS model. We also note the high accuracy of the Log.Reg models that are comparable to XGBoost in all of the endpoints other than renal failure. Assuming an additive relationship between independent covariates, Log.Reg is able to provide high quality predictions. The STS risk models use Log.Reg. Thus, the superior performance of our models could be due to the newly added variables, the missing data imputation methods, or the tuning process of the hyperparameters. OCT, as a standalone method, achieves equivalent performance when compared to the STS risk calculator. Combining two of the most accurate and interpretable methods in the L-OCT models resulted in models that achieved both high performance and

transparency. Thus, due to their interpretable nature, we recommend the use of Log.Reg or L-OCT in all endpoints where their performance is comparable to XGBoost.

The latter characteristic has become of critical importance to the ultimate success of ML models in the medical field. Even though there has been a significant increase in the number of publications that leverage artificial intelligence and ML in medical applications, there has been minor integration of those applications into the healthcare system.[30] Unless the models provide actionable insight and guidelines for a clinician, they may not achieve any meaningful impact.[31] The FDA (2017) validated such concerns by mandating the use of interpretable ML models when it comes to medical decision making.[23]

For this reason, we focused our efforts on building the L-OCT models and a corresponding user-friendly tool for practitioners. Known findings that appeared as branching nodes in Figure 4 include the age of the patient > 72.5 years or the effect of an intra-aortic balloon pump (IABP). In this model, creatinine levels play a significant role in the determination of mortality risk as high values above 1.5 are associated with a mortality risk of 11.64%. Moreover, the mortality model distinguishes between different types of surgery indicating that mitral valve replacement operations are riskier (4.49%) than annuloplasty or mitral valve repair (1.76%). In addition, ML resulted in more accurate and better calibrated models for mitral valve reconstruction/repair and mitral valve replacement. To the contrary, in the case of only annuloplasty for the outcomes deep sternal wound infection and prolonged ventilation, the STS risk score remains better calibrated compared to the most accurate ML model.

Our efforts were subsequently focused on yielding actionable insights from other models such as Log.Reg and XGBoost. Our results show that age is consistently considered important, other than the case of deep sternal wound. For this morbidity, the last Hemoglobin A1c level plays the most significant role for the risk determination. Our results reveal the role that an IABP plays in terms

of outcome, signifying a sicker cohort of patients that require higher degree of attention and support.

We improve the AUC and calibration of the OCT when we create Log.Reg models for each leaf of the tree (L-OCT). An independent Log.Reg model is derived for all patients that had an IABP procedure with baseline risk of 25.79%. Similarly, separate Log.Reg models were developed for patients who did not undergo an IABP placed with creatinine levels >1.5. This model combines the tree-based structure of OCT with the traditional Log.Reg method, rendering accurate and personalized estimations of risk.

Limitations

Central to the limitations of our study lies in the fact that the power of machine-learning prediction depends on the accuracy and comprehensiveness of the data it uses, in this case the ACSD.[11] As such, systematic biases resulting from the ACSD data collection methodology and its changes over the years of data collection might exist. Another limitation refers to causality between the variables and the outcomes, which is still not proven despite the high degree of connectivity between the two. Therefore, interpretability and actionability on the relevant variables remains controversial. The ML models created in this study confirm medical insights that have been previously discussed in the medical literature. The most predictive risk factors identified by the ML models would need to be taken into consideration in conjunction with the results from the Log.Reg models, the STS Risk Score, and prior relevant findings of the medical literature. In addition, we believe that only by including new variables in the data, that have not been previously measured, will we be able to uncover new interactions between the risk factors. Moreover, outcomes with limited incidence such as deep sternal wound infection remain challenging endpoints for these advanced methods. Concomitant procedures during the MVS were not considered in our analysis. The inclusion of more cases in the future may result in stronger performance. The sensitivity and accuracy results of the non-linear algorithms, such as XGBoost, were slightly compromised compared to AUC since the cross-validation procedure was optimized

based on the latter metric. Future research could focus on a retrospective external validation of the models with new observations from subsequent years that were not included in this study. The proposed models would need to be validated with more contemporaneous entries of the database and with the most recent version of the STS risk score. Another direction could be the direct comparison of ML algorithms to the existing STS score without the inclusion of new features. Such an approach would be able to quantify the exact benefit of ML in the MVS risk prediction task.

ML methods are able to provide more accurate risk prediction models compared to the existing STS risk calculator in terms of predicting mortality and some morbidity for MVS, especially for more complex mitral valve operations such as mitral valve reconstruction/repair and replacement. Our results confirm the fact that the use of artificial intelligence in the field of medical prediction can be enhanced by incorporating interpretable, user-friendly tools which clinicians can understand and incorporate in practice.

Acknowledgements:

The data for this research were provided by The Society of Thoracic Surgeons' National Database Participant User File Research Program. Data analysis was performed at the investigators' institutions. The authors thank Suzi Birz and Michele Blackwelder for her administrative assistance in coordinating the effort between the two institutions and Barry Stein for supporting the whole team endeavor and vision. The authors also acknowledge the continuous support provided by Dr. David Shahian throughout this investigation and the valuable guidance and feedback of Dr. Robert Habib. We would also like to thank the reviewers from the *Journal of Cardiac Surgery*, whose comments significantly improved the paper.

References:

1. Kouchoukos NT, Ebert PA, Grover FL, et al. Report of the Ad Hoc Committee on Risk Factors for Coronary Artery Bypass Surgery. *The Annals of thoracic surgery* 1988; 45(3):348-349.
2. Kilic A, Acker MA, Gleason TG, et al. Clinical outcomes of mitral valve reoperations in the United States: an analysis of The Society of Thoracic Surgeons National Database. *The Annals of thoracic surgery* 2019; 107(3):754-759.
3. Nashef SA, Roques F, Hammill BG, et al. Validation of European system for cardiac operative risk evaluation (EuroSCORE) in North American cardiac surgery. *European journal of cardio-thoracic surgery* 2002; 22(1):101-105.
4. Shroyer ALW, Coombs LP, Peterson ED, et al. The Society of Thoracic Surgeons: 30-day operative mortality and morbidity risk models. *The Annals of thoracic surgery* 2003; 75(6):1856-1865.
5. O'Brien SM, Feng L, He X, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 2—statistical methods and results. *The Annals of thoracic surgery* 2018; 105(5):1419-1428.
6. Shahian DM, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 adult cardiac surgery risk models: part 1—background, design considerations, and model development. *The Annals of thoracic surgery* 2018; 105(5):1411-1418.
7. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis: Springer, 2015.
8. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *The New England journal of medicine* 2017; 376(26):2507.
9. Larsson SC, Wallin A, Wolk A, et al. Differing association of alcohol consumption with different stroke types: a systematic review and meta-analysis. *BMC medicine* 2016; 14(1):178.

10. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* 2016; 375(13):1216.
11. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019; 380(14):1347-1358.
12. Allyn J, Allou N, Augustin P, et al. 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(1):e0169772.
13. Krittanawong C, Zhang H, Wang Z, et al. Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology* 2017; 69(21):2657-2664.
14. Bertsimas D, Kallus N, Weinstein AM, et al. Personalized Diabetes Management Using Electronic Medical Records. *Diabetes Care* 2017; 40(2):210-217.
15. Bertsimas D, Dunn J, Velmahos GC, et al. Surgical Risk Is Not Linear: Derivation and Validation of a Novel, User-friendly, and Machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator. *Ann Surg* 2018; 268(4):574-583.
16. Bertsimas D, Orfanoudaki A, Weiner RB. Personalized Treatment for Coronary Artery Disease Patients: A Machine Learning Approach. *arXiv preprint arXiv:1910.08483* 2019.
17. Shih T, Paone G, Theurer PF, et al. The society of thoracic surgeons adult cardiac surgery database version 2.73: more is better. *The Annals of thoracic surgery* 2015; 100(2):516-521.
18. Bertsimas D, Pawlowski C, Zhuo YD. From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research* 2017; 18(1):7133-7171.
19. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine* 2011; 30(4):377-399.

20. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical Risk Is Not Linear: Derivation and Validation of a Novel, User-friendly, and Machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator. *Ann Surg*. Oct 2018;268(4):574-583.

21. Bertsimas D, Dunn J, Pawlowski C, et al. Applied informatics decision support tool for mortality predictions in patients with cancer. *JCO clinical cancer informatics*. 2018;2:1-11.

22. Bertsimas D, Kung J, Trichakis N, Wang Y, Hirose R, Vagefi PA. Development and validation of an optimized prediction of mortality for candidates awaiting liver transplantation. *American Journal of Transplantation*. 2019;19(4):1109-1118.

23. Administration USFaD. Clinical and Patient Decision Support Software - Guidance for Industry and Food and Drug Administration Staff. *Center for Devices and Radiological Health* 2017.

24. Breiman L. Classification and regression trees: Routledge, 2017.

25. Breiman L. Random Forests. *Machine learning* 2001; 45(1):5-32.

26. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining: ACM, 2016. pp. 785-794.

27. Bertsimas D, Dunn J. Optimal classification trees. *Machine Learning* 2017; 106(7):1039-1082.

28. Bertsimas D, Dunn J. Machine Learning under a Modern Optimization Lens. Dynamic Ideas, Belmont, 2019.

29. Kilic, A., Goyal, A., Miller, J.K., Gjekmarkaj, E., Tam, W.L., Gleason, T.G., Sultan, I. and Dubrawksi, A. Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery. *The Annals of thoracic surgery* 2019.

30. Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? *Jama* 2019; 321(23):2281-2282.

31. Ghassemi M, Naumann T, Schulam P, et al. Opportunities in machine learning for healthcare. *arXiv preprint arXiv:1806.00388* 2018.

Figure Legends:

Fig.1: Example of the MVS mortality and morbidity risk calculator's interface.

Fig.2: The ROC curves for the STS risk model and the best performing ML method according to the AUC metric for all outcomes and types of surgery.

Fig.3: The calibration curves for the STS risk model and the best performing ML method according to the AUC metric for all outcomes and types of surgery.

Fig.4: Top 10 important predictors of each outcome selected by the XGBoost algorithm. The average information gain in gini indices is illustrated across five splits of the data.

Fig.5: The OCT model for mortality risk within 30 days from MVS.

Tables:

Table 1: Postoperative outcomes on the training and testing sets. Note that the results remain the same across all five random splits of the data as outcome stratification is used.

Sample Size	Overall Cohort	Training Set	Testing Set
Total Observations	383,550	287,661	95,889
Deaths	24,331 (6.34%)	20,161 (7.00%)	4,170 (4.34%)
Prolonged Ventilation	79,254 (20.66%)	63,867 (22.20%)	15,387 (16.04%)
Renal Failure	23,584 (6.15%)	19,125 (6.64%)	4,459 (4.65%)
Deep Sternal Wound	755 (0.20%)	607 (0.20%)	148 (0.15%)
Permanent Stroke	8,362 (2.18%)	6,495 (2.25%)	1,867 (1.94%)

Table 2: AUC and 95% Confidence intervals (CI) for all methods and outcomes as measured in the validation population (testing set). P-values were calculated using the DeLong statistical test comparing the AUC of each ML model with the STS score for all tasks. All results are averaged across five different splits of the data.

	Mortality			**Prolonged Ventilation**		
	Mean	CI: 95%	p-value	Mean	CI: 95%	p-value
STS Score	0.796	(0.794, 0.8)	NA	0.788	(0.786, 0.79)	NA
Log.Reg	0.825	(0.823, 0.828)	0.001	0.817	(0.816, 0.819)	0.001
XGBoost	0.826	(0.824, 0.829)	0.001	0.819	(0.818, 0.82)	0.001
RF	0.77	(0.767, 0.774)	0.995	0.806	(0.805, 0.808)	1.000
OCT	0.787	(0.784, 0.792)	0.995	0.792	(0.789, 0.797)	0.037
L-OCT	0.808	(0.807, 0.811)	0.001	0.808	(0.803, 0.814)	0.001
	Deep Sternal Wound Infection			**Permanent Stroke**		
	Mean	CI: 95%	p-value	Mean	CI: 95%	p-value
STS Score	0.7507	(0.741, 0.761)	NA	0.6798	(0.673, 0.687)	NA
Log.Reg	0.784	(0.774, 0.795)	0.032	0.676	(0.67, 0.682)	0.812
XGBoost	0.771	(0.751, 0.791)	0.170	0.6808	(0.677, 0.685)	0.422
RF	0.5208	(0.512, 0.53)	1.000	0.5752	(0.568, 0.582)	1.000
OCT	0.698	(0.669, 0.729)	0.970	0.6525	(0.644, 0.661)	1.000
L-OCT	0.723	(0.676, 0.77)	0.726	0.6713	(0.665, 0.677)	0.987
	Renal Failure					
	Mean	CI: 95%	p-value			
STS Score	0.7909	(0.79, 0.792)	NA			
Log.Reg	0.798	(0.796, 0.802)	0.009			
XGBoost	0.815	(0.812, 0.819)	0.001			
RF	0.745	(0.742, 0.748)	1.000			
OCT	0.789	(0.783, 0.795)	0.630			
L-OCT	0.8086	(0.805, 0.813)	0.400			

Figures:

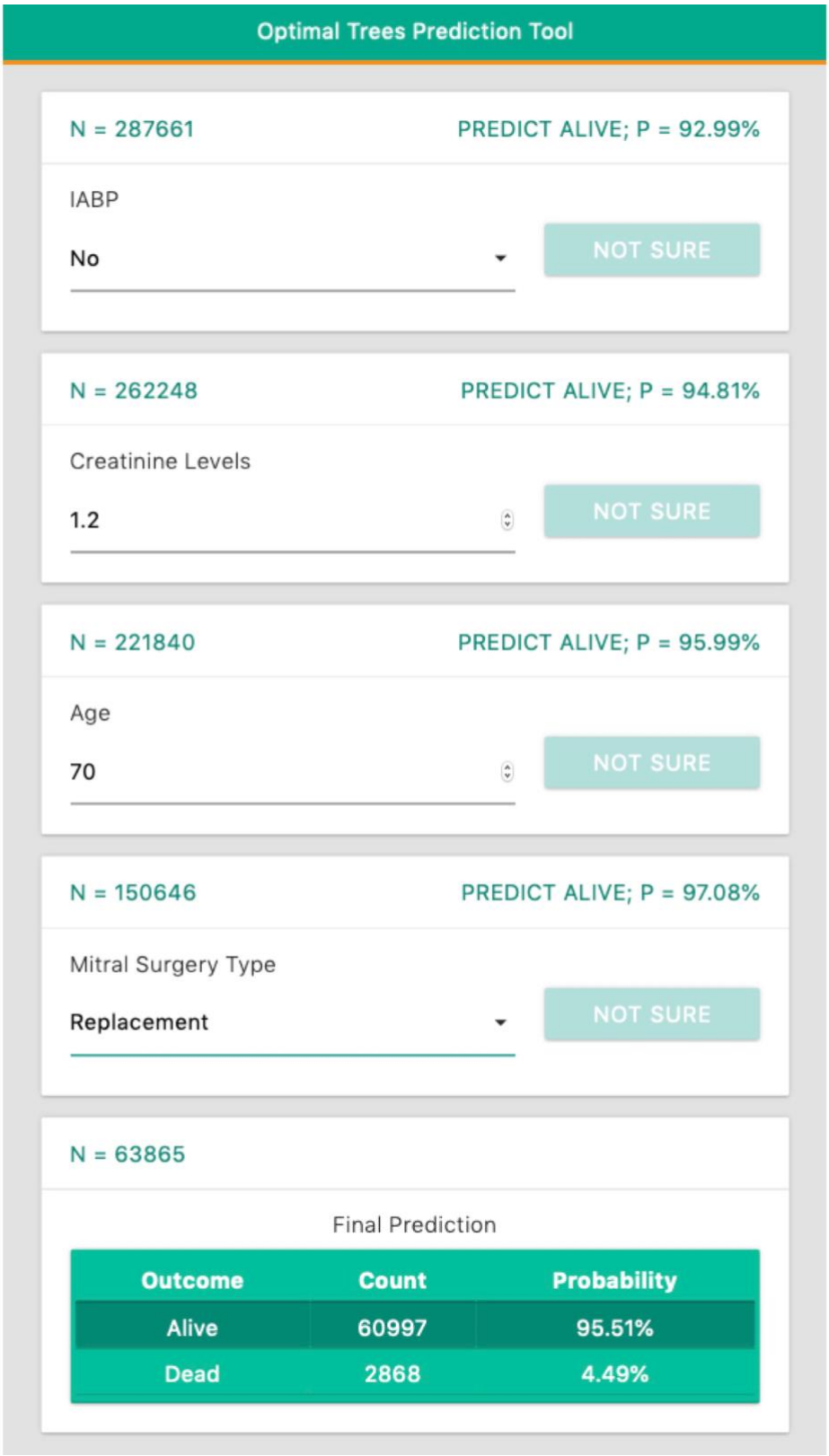

Figure 1: Example of the mitral valve surgery mortality and morbidity risk calculator's interface.

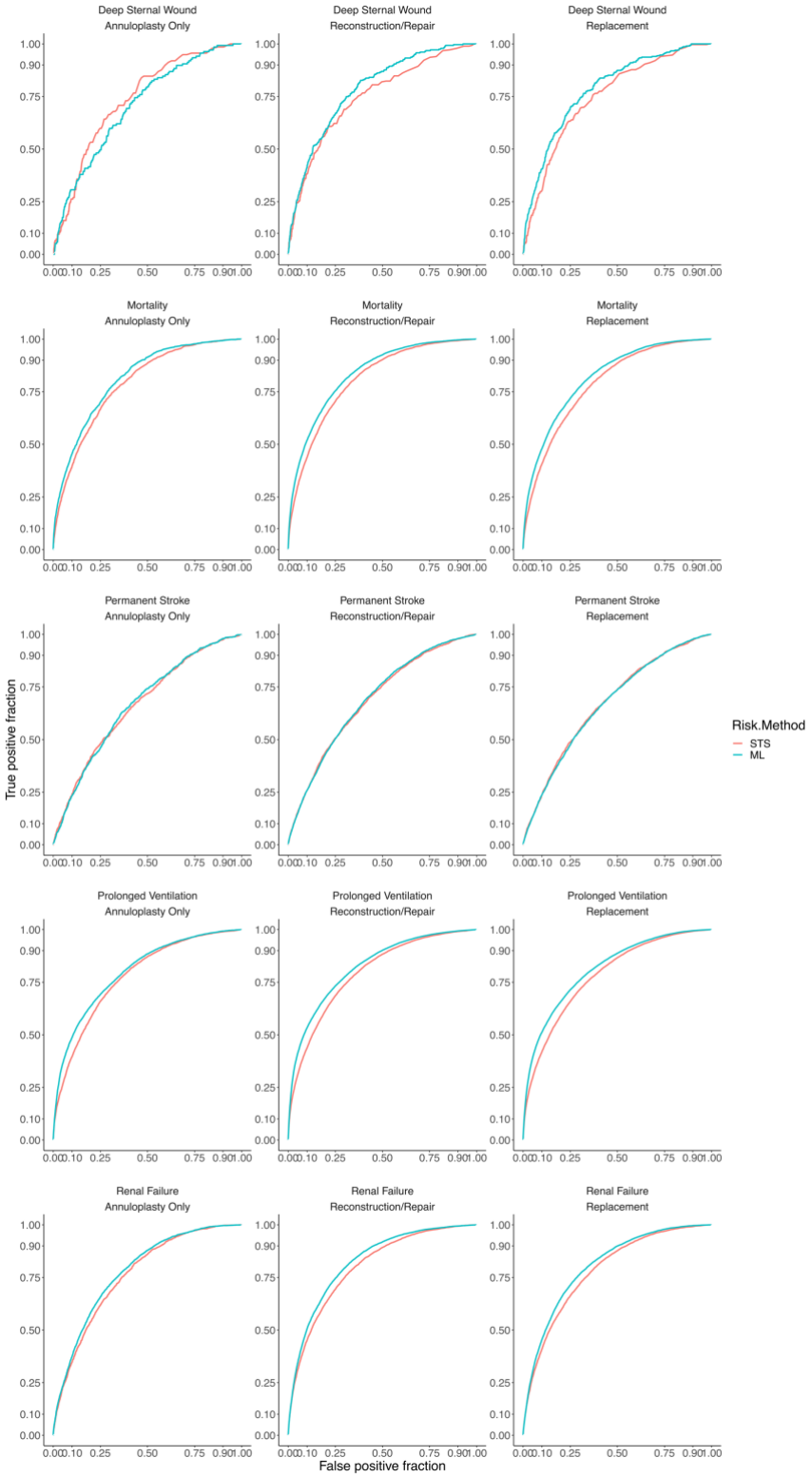

Figure 2: The ROC curves for the Society of Thoracic Surgeons risk model and the best performing machine learning method according to the AUC metric for all outcomes and types of surgery.

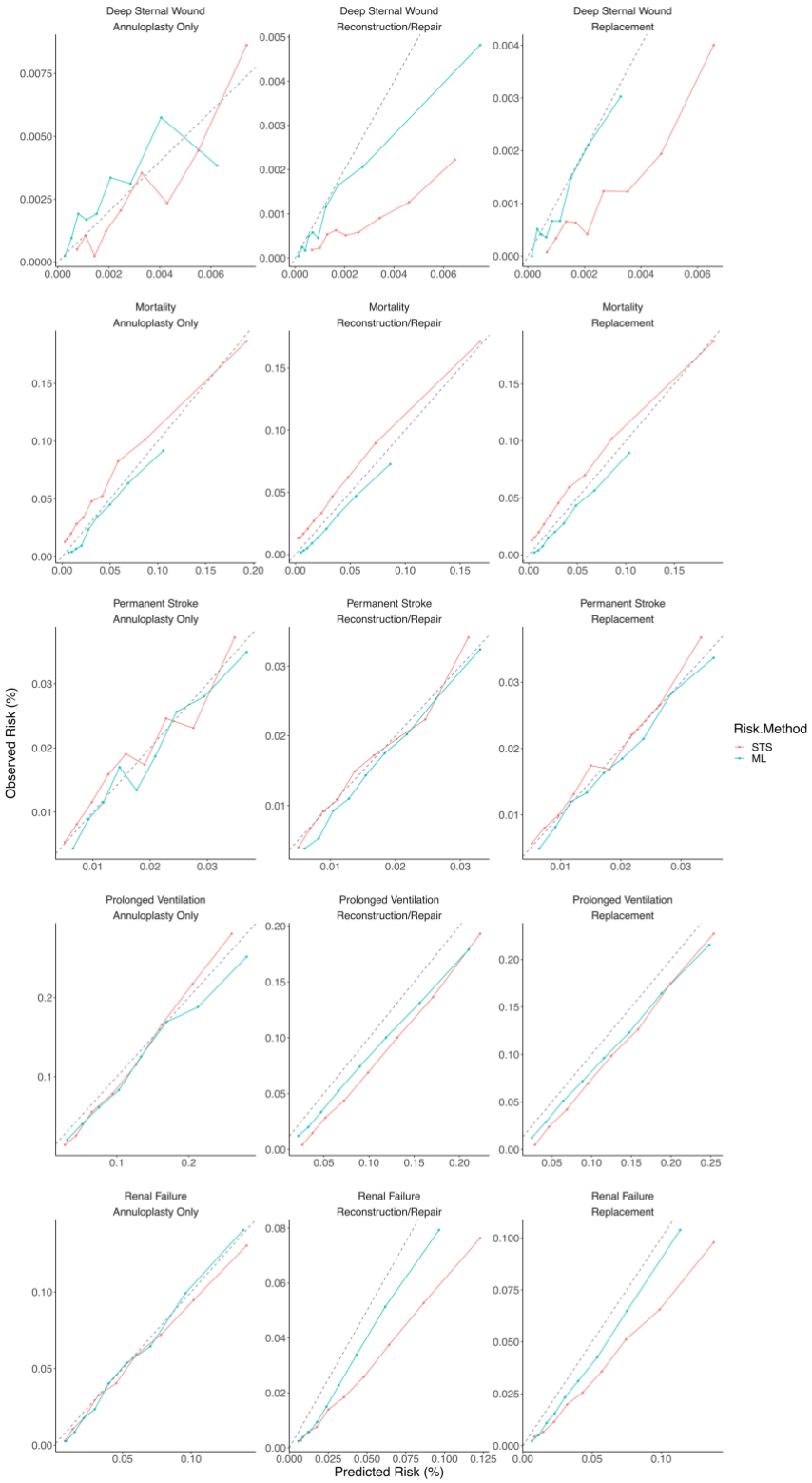

Figure 3: The calibration curves for the Society of Thoracic Surgeons risk model and the best performing machine learning method according to the AUC metric for all outcomes and types of surgery.

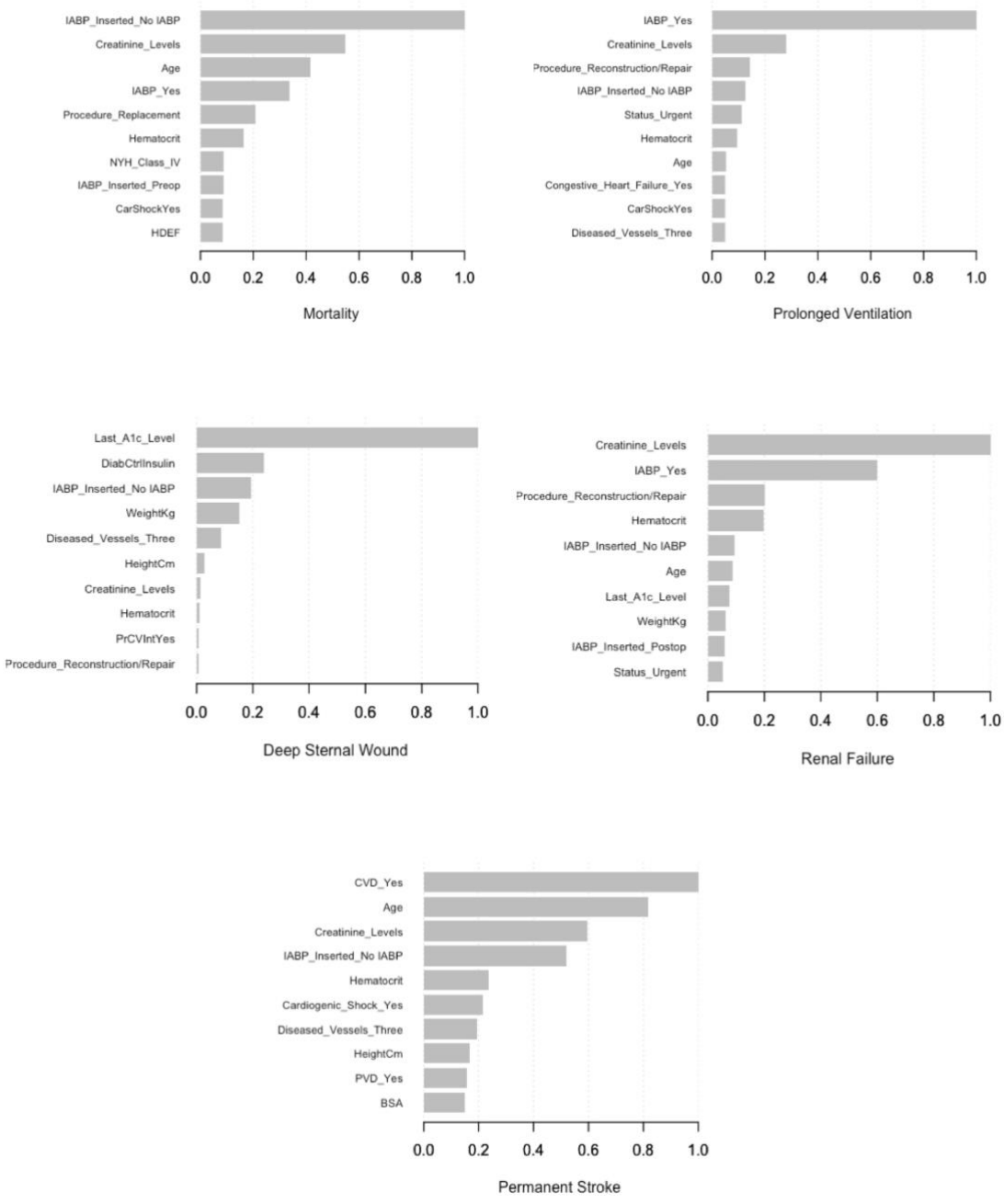

Figure 4: Top 10 important predictors of each outcome selected by the XGBoost algorithm. The average information gain in gini indices is illustrated across five splits of the data.

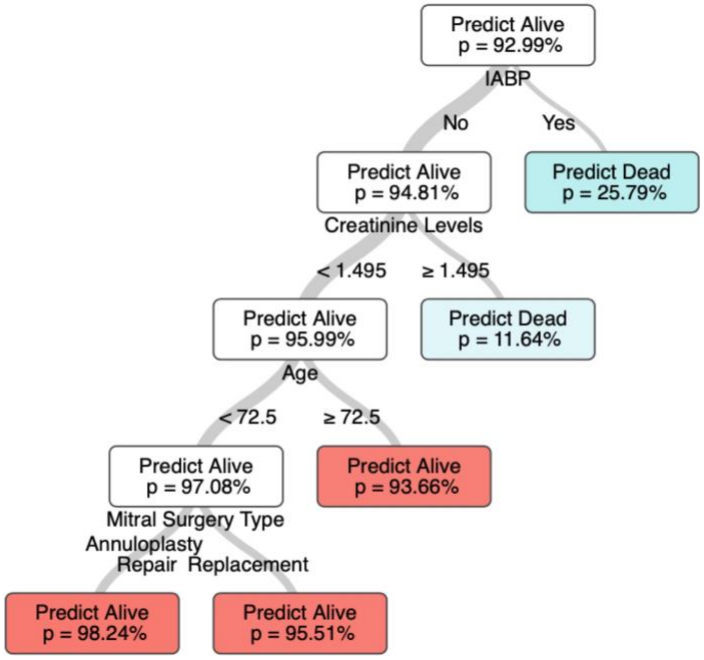

Figure 5: The Optimal Classification Trees model for mortality risk within 30 days from mitral valve surgery.