Title

Parsimonious Machine Learning Models to Predict Resource Use in Cardiac Surgery Across a Statewide Collaborative

Author

Verma, Arjun

Publication Date

2022-06-03

Undergraduate

Parsimonious Machine Learning Models to Predict Resource Use in Cardiac Surgery Across a Statewide Collaborative

Arjun Verma[1], Yas Sanaiha MD[1], Joseph Hadaya MD[1], Anthony Jason Maltagliati MD[2], Zachary Tran MD[1], Ramin Ramezani PhD[3], Richard J. Shemin MD[4], Peyman Benharash MD[1,4], and the University of California Cardiac Surgery Consortium*

[1]Cardiovascular Outcomes Research Laboratories (CORELAB), University of California Los Angeles, Los Angeles, CA
[2]Department of Surgery, Harbor-UCLA Medical Center, Los Angeles, CA
[3]Department of Computer Science, University of California Los Angeles, Los Angeles, CA
[4]Division of Cardiac Surgery, University of California Los Angeles, Los Angeles, CA
*Contributing Consortium Members Listed in Supplemental Table 1

Presented at the 47th Annual Meeting of the Western Thoracic Surgical Association, Phoenix, AZ, September 30th – October 2nd, 2021

Disclosures: Richard J. Shemin serves as a consultant to the Edwards Lifesciences Advisory Board and as a Co-Principal Investigator on the PARTNER II Trial. The other authors have no conflict of interest to disclose.

Funding: None

Acknowledgements: None

IRB approval: #16-000558 (Approved 5/6/2016, Renewed 4/15/2020)

Corresponding Author

Peyman Benharash, MD
10833 Le Conte Avenue
UCLA Center for Health Sciences, Room 62-249
Los Angeles, CA 90095
Phone: 310-206-6717
Fax: 310-206-5901
Pbenharash@mednet.ucla.edu

Word Count: 3189

Abbreviations and Acronyms

AKI	acute kidney injury
AUC	area under the receiver operating characteristic
CABG	coronary artery bypass grafting
CI	confidence interval
COVID-19	coronavirus disease 2019
EF	ejection fraction
GBM	gradient boosted machine
ICU	intensive care unit
INR	international normalized ratio
IQR	interquartile range
LOS	length of stay
ML	machine learning
RF	random forest
R^2	coefficient of determination
STS	Society of Thoracic Surgeons
UCCSC	University of California Cardiac Surgery Consortium
XGBoost	extreme gradient boosting

Central Message (196/200)

Compared to traditional regression, machine learning yielded superior performance in the prediction of length of stay, mortality, acute kidney injury and reoperation following cardiac operations.

Perspective Statement (379/405)

This study outlined the development of machine learning (ML) models to predict length of stay (LOS) following cardiac operations. Several clinical, operative and hospital characteristics were found to be associated with increased LOS. Taken together, our findings suggest that ML models may be used to inform case scheduling strategies during times of limited hospital capacity.

Central Picture Legend (69/90)

Observed length of stay versus predictions by machine learning model.

Abstract (250/250)

Objective

We sought to several develop parsimonious machine learning (ML) models to predict resource utilization and clinical outcomes following cardiac operations using only preoperative factors.

Methods

All patients undergoing coronary artery bypass grafting and/or valve operations were identified in the 2015-2021 University of California Cardiac Surgery Consortium repository. The primary endpoint of the study was length of stay (LOS). Secondary endpoints included 30-day mortality, acute kidney injury (AKI), reoperation, postoperative blood transfusion and duration of intensive care unit admission (ICU LOS). Linear regression, gradient boosted machines (GBM), random forest (RF), extreme gradient boosting (XGBoost) predictive models were developed. The coefficient of determination (R^2) and area under the receiver operating characteristic (AUC) were used to compare models. Important predictors of increased resource use were identified using SHapley summary plots.

Results

Compared to all other modeling strategies, GBM demonstrated the greatest performance in the prediction of LOS (R^2 0.42), ICU LOS (R^2 0.23) and 30-day mortality (AUC 0.69). Advancing age, reduced hematocrit and multiple-valve procedures were associated with increased LOS and ICU LOS. Furthermore, the GBM model best predicted AKI (AUC 0.76), while RF exhibited greatest discrimination in the prediction of postoperative transfusion (AUC

0.73). We observed no difference in performance between modeling strategies for reoperation (AUC 0.80).

Conclusion

Our findings affirm the utility of ML in the estimation of resource use and clinical outcomes following cardiac operations. We identified several risk factors associated with increased resource use, which may be used to guide case scheduling in times of limited hospital capacity.

Keywords: cardiac surgery, resource utilization, length of stay, machine learning, COVID-19, pandemic, hospital capacity

Introduction

The novel coronavirus disease 2019 (COVID-19) pandemic has placed unprecedented strain on healthcare systems, influencing the allocation of personnel and resources. Several groups have reported cardiac surgery case volume reductions of 45-94%, with significant regional variability.[1–4] Subject to rates of "reopening" and patients' desire to proceed with elective surgery, the projected time to equilibrium between back-logged cases and ongoing surgical need is estimated to be 12-22 months.[5–7] Furthermore, recovery from cessation of elective cases requires a nuanced approach to management of deferred and newly presenting patients as well as ongoing demands for perioperative resources. With estimates that operating volume must exceed 120% of baseline to accommodate deferred patients while concurrently preventing excess waitlist morbidity, rapid and accurate prediction of hospital bed occupancy and resource utilization are especially crucial.[6]

The Society of Thoracic Surgeons (STS), among others, has successfully implemented risk models to provide canonical estimates for parameters such as mortality, postoperative complications and prolonged length of stay (LOS).[8] However, as demonstrated by several reports of poor calibration when applied at the institutional level, these predictive tools are often cumbersome and require numerous data fields to yield a predicted risk without accounting for local variations in clinical practice.[9–11] Furthermore, most available models predict prolonged LOS in a binary manner, rather than an estimate of the actual duration of hospitalization in days.[8,12–14] The classification of LOS into prolonged and routine may reduce generalizability and limit the application of such tools in the acute care setting.[15]

Machine learning (ML) algorithms allow for complex modeling of non-linear relationships between predictive factors and have demonstrated superior discrimination and calibration in several clinical applications.[16–18] Therefore, we sought to develop ML-based models to predict length of stay, 30-day mortality and select complications using an academic, state-wide database. We hypothesized that a parsimonious ML model, containing few explanatory covariates, would yield superior discrimination and calibration compared to traditional linear and logistic regression.

Methods

Study Population

All adults (≥ 18 years) who underwent coronary artery bypass grafting (CABG) and/or valve operations were identified from the 2015-2021 University of California Cardiac Surgery Consortium (UCCSC) repository. Founded in 2013, the UCCSC is a collaborative among five academic hospitals across California. Data elements, including those submitted to the STS, are prospectively collected in compliance with policies of individual institutions and the University of California System-wide Review Board (IRB #16-000558, approved 5/6/2016, renewed 4/15/2020). Patient written consent for the publication of the study data was waived by the IRB due to the de-identified nature of the UCCSC.

Patients were stratified by the class of operation performed: isolated CABG, isolated valve, concomitant CABG/valve and multi-valve operations. Those who required left ventricular assist device implantation, extracorporeal membrane oxygenation or transcatheter procedures were excluded to maintain cohort homogeneity. Moreover, records with missing values for

overall and intensive care unit (ICU) LOS as well as 30-day mortality were excluded (Supplemental Figure 1). Patients with LOS or ICU LOS greater than the 95th percentile (>30 days for LOS, >259 hours in ICU) were similarly excluded.

Variable and Outcome Definitions

The primary endpoint was overall LOS. Mortality at 30-days, acute kidney injury (AKI), postoperative blood transfusion, reoperation and ICU LOS were also considered. Patient comorbidities, operative characteristics, and complications including AKI, postoperative blood transfusion and reoperation, were defined in accordance with the STS Adult Cardiac Database dictionary.[19] Annual operative caseload, number of adult cardiac surgeons, total number of low acuity and cardiothoracic ICU beds were tabulated for each institution. Variables with missing values in >20% of patients were not considered for inclusion. For retained features with missing data, values were imputed using the median and mode for continuous and categorical variables, respectively. The number of records with missing data for each variable is reported in Supplemental Table 2.

Modeling Techniques

We compared three ML models to traditional, multivariable linear and logistic regression: gradient boosted machines (GBM), extreme gradient boosting (XGBoost) and random forest (RF). These algorithms autonomously generate a large set of decision trees to capture nuanced patterns in the training dataset. In the case of RF, development of one decision tree occurs independently from the other and the final output of the model is the arithmetic mean of the output from each decision tree. However, XGBoost and GBM develop decision trees in a

stepwise manner to compensate for errors of the prior trees, and the output is the weighted average of each decision tree's estimate.[20] A brief schematic highlighting the differences between boosting and bagging classifiers is shown in Supplemental Figure 2. Hyperparameters, which are used to control the learning process of ML models, were selected using the GridSearchCV function in the Python *sklearn* library. This technique exhaustively evaluates a wide range of hyperparameters and selects values which optimize model performance. Selected hyperparameters for each model are shown in Supplemental Table 3.

Model Development

Thirty-seven preoperative patient and hospital characteristics were chosen as candidate predictors. Clinical variables were selected from the STS risk score variable list based on clinical relevance and are listed in Table 1.[8] Hospital factors were incorporated to account for variation in practice across participating institutions. Variable selection was performed using recursive feature elimination (RFE), a ML technique that is used to reduce collinearity and eliminate covariates with low variance. In RFE, cross validation is used to exhaustively evaluate variable sets of different sizes and select the best collection of features. Given that transportability and ease of use is an important aspect of risk tools, we identified smallest set of variables which retained maximum predictive performance. This algorithm was independently applied using linear regression and GBM to ascertain any differences between modeling strategies. Selected variables were used for all subsequent model development (Supplemental Table 4). We also compared the performance of ML against the STS risk scores for 30-day mortality, AKI and reoperation.

The derivation cohort consisted of operations performed before March 2020, while the remainder comprised the validation dataset. To obtain cross-validated performance metrics, models were fit using 50% of the derivation cohort and tested using the remainder. This process was repeated 100 times to acquire model performance metrics, which are reported as means with 95% confidence intervals (95% CI). To account for potential differences in case-mix due to the COVID-19 pandemic, we assessed the stability of model performance in the pre- (derivation) and post-COVID-19 (validation) eras.

Model Evaluation and Interpretation

Linear regression, GBM, RF and XGBoost models were compared using the coefficient of determination (R^2) between observed and predicted values. Binary classifiers were evaluated using the area under the receiver operating characteristic (AUC). The accuracy of probabilistic predictions was assessed using the Brier score, for which lower values denote superior calibration. Model R^2 and Brier scores were analyzed using a paired t-test, which allowed for comparison of model performance across cross-validation folds. Similarly, model AUCs were compared using DeLong's test, which specifically accounts for the impact of model evaluation on a common test set. SHapley additive values were calculated to estimate the marginal impact of each covariate on the output of a decision tree model.[17]

Baseline characteristics are reported as means with standard deviation or medians with interquartile range (IQR), as appropriate. Means were analyzed using the adjusted Wald's test, while medians with the Mann-Whitney U test. Categorical variables are reported as frequencies and were compared using the Pearson's chi-squared test. Statistical significance was set at $\alpha<0.05$. Statistical analysis was conducted using Stata 16.0 (StataCorp, TX) and Python version

3.9 (Python Software Foundation, Wilmington, DA). The *sklearn, shap* and *xgboost* packages of Python were used to develop and assess ML models as described above.[21,22]

Results

Population Characteristics

Across five participating centers, 6,316 patients met study criteria. The study cohort was predominantly male (72.5%), with mean age of 63 years. A significant proportion of patients had pre-existing medical conditions such as diabetes, congestive heart failure and atrial fibrillation (Table 1). The most frequent operation was isolated CABG (50.5%), followed by isolated valve (33.3%) and concomitant CABG/valve operations (10.6%). The majority of operations were performed electively. Over the study period, the highest volume center performed 1,205 operations, while the lowest volume center performed 626. The 30-day mortality rate was 0.9%. Overall, 27.7% of patients received postoperative transfusions, and 1.5% developed AKI. Median LOS was 8 days (IQR 6-13) with a median ICU LOS of 74 hours (IQR 47-116).

Comparison of baseline characteristics and outcomes between the derivation and validation cohorts is shown in Tables 1 and 2. Patients in the validation cohort were marginally older (64±13 vs 63±13 years, $p<0.001$) and had greater rates of congestive heart failure (45.0 vs 33.8%, $p<0.001$) and peripheral vascular disease (11.1 vs 8.3%, $p=0.003$). Valve operations were more frequent in the validation group, compared to derivation. While rates of 30-day mortality and AKI were similar, the incidence of reoperation (6.9 vs 9.1%, $p=0.014$) and postoperative blood transfusion (23.1 vs 28.8%, $p<0.001$) was lower in the validation cohort. The distribution

of LOS and ICU LOS was statistically different between the derivation and validation datasets (Table 2).

Variable Selection

Recursive feature elimination was applied to 37 candidate variables to determine the optimal covariate set in the prediction of overall LOS. Figure 1 demonstrates the cross validated R^2 versus the number of covariates included in each model. The GBM model outperformed linear regression, regardless of feature set size. Notably, after the inclusion of 23 features, no appreciable increase in performance was observed from the GBM or linear regression model. Thus, all models were developed using the 23 features which were most strongly associated with LOS (Supplementary Table 4).

Resource Utilization

Linear regression, GBM, RF and XGBoost models were developed to predict in-hospital LOS. Compared to linear regression, the GBM model yielded the greatest R^2 (0.42 vs 0.41, $p<0.001$). As shown in Supplemental Figure 3, predictions by the GBM model were more strongly correlated with observed values for LOS, compared to linear regression. While the difference in cross-validated R^2 between the two strategies was subtle, the GBM model greatly outperformed linear regression in the validation dataset (R^2 0.47 vs 0.42, Table 4). When assessing cumulative model error in the validation cohort, the GBM model resulted in a 197-day reduction in error across all patients relative to linear regression.

The GBM model was interpreted using SHapley summary plots, and the most salient predictors of LOS were ranked by their relative importance (y-axis). Figure 2 depicts how high

(red dot) and low (blue dot) feature values corresponded to a change in LOS prediction. Elective admission had the highest feature importance and was associated with significantly decreased LOS. In addition, we found decreased hematocrit and serum albumin to increase the estimated LOS. Certain procedures, such as concomitant CABG/valve and multi-valve operations, were found to confer longer LOS. Notably, an increased number of floor beds conferred greater estimated LOS (Figure 2).

In the prediction of ICU LOS, the GBM model demonstrated significantly increased cross-validated R^2 compared to linear regression (0.23 vs 0.15, $p<0.001$). However, in the validation dataset, the XGBoost model demonstrated the highest performance (Table 4). Decreased pre-operative creatinine, low EF and pre-existing congestive heart failure were associated with greater predicted ICU LOS. Notably, increased annual hospital volume and a higher number of low acuity beds were associated with lower estimated ICU LOS (Figure 3).

Clinical Outcomes

The GBM, RF and XGBoost models outperformed logistic regression in the prediction of 30-day mortality (AUC 0.69 vs 0.67, $p<0.001$). Furthermore, the GBM and RF models outperformed logistic regression and XGBoost in the prediction of AKI (Table 3). While postoperative blood transfusion was best predicted by GBM and XGBoost, all modeling strategies displayed similar discrimination in the estimation of reoperation (Table 3). The STS risk score for 30-day mortality and AKI outperformed ML models. However, ML displayed greater discrimination than the STS model in the prediction of reoperation (Table 3). These comparisons were consistent when evaluating the Brier score for each model (Supplemental Tables 5 and 6).

Discussion

Reliable estimation of hospitalization duration remains a challenge for surgeons and administrators alike. The present study developed several parsimonious ML models to develop a readily useful prediction instrument for LOS. This work entails one of the largest applications of ML to discretely model LOS using a multi-center, academic dataset. Compared to linear and logistic regression, we found ML algorithms to exhibit the best performance for prognostication of LOS, 30-day mortality, AKI, postoperative transfusion and ICU LOS. Using autonomous techniques, we identified several key predictors of increased resource use including existing comorbidities, decreased preoperative hematocrit and serum albumin. And finally, we noted a significant impact of hospital characteristics on ICU LOS, suggesting the need for incorporation of center-specific characteristics in predictive tools.

Several clinical characteristics, including preoperative anemia, renal dysfunction and operative complexity, were associated with increased overall and ICU length of stay. These findings are expected since laboratory values such as hematocrit, INR, creatinine and albumin are incorporated in virtually every clinical risk score calculator.[8] Moreover, these clinical factors influence the development of postoperative complications, including pneumonia and AKI, which are drivers of hospital LOS and costs.[8,12,17,23] SHapley interpretation revealed that more complex operations were associated with greater LOS. The relatively higher incidence of complications in the setting of complex cardiac surgery, such as pacemaker placement, need for blood transfusion, and a greater need for ICU-level care, may explain this observation. Taken together, our findings validate the utilization of ML methods to reduce bias, enhance external validity and autonomously select features associated with increasing LOS. Furthermore, our results

demonstrate that during times of limited hospital capacity, clinical characteristics such as organ dysfunction and operative complexity should be considered when predicting hospitalization duration.[24]

In addition to patient factors, we found certain hospital structural characteristics to influence ICU LOS. For example, increasing cardiac institutional volume and a greater number of low acuity beds was associated with reduced ICU LOS. Several factors may contribute to this important finding. Greater institutional cardiac surgery volume may represent greater expertise, the presence of standardized care pathways, and more efficient hospital throughput for these cases. Moreover, greater availability of low-acuity beds may lead to less delay in transitioning out of the ICU when clinical milestones are met.[15] Consistent with this notion, several prior studies have demonstrated wide variation in hospital practices that may influence LOS, such as expedited discharge after lung resection and CABG.[25,26] A nationwide study of minimally invasive esophagectomy in the Netherlands demonstrated great heterogeneity in ICU LOS, pointing to differences in use of early extubation protocols and analgesic modalities as contributing factors.[27] Investigation at a broader scale is necessary to confirm the generalizability of our findings and to identify modifiable practice patterns that increase LOS.

In the present work, ML models exhibited superior accuracy in the prognostication of overall and ICU length of stay, compared to linear regression. A single-center study similarly compared linear regression and artificial neural networks, finding the latter to have enhanced LOS prediction for patients undergoing isolated CABG.[28] Furthermore, LaFaro and colleagues used a sample of 185 cardiac surgical patients to show that artificial neural networks yield more accurate estimates of ICU LOS compared to linear regression.[29] The improved performance of

ML models is likely attributable to their ability to capture non-linear interactions between covariates and outcomes of interest. While the decision tree structure evaluates such interactions autonomously, linear regression models can only accommodate explicitly included interaction terms, making the development of an equivalent model cumbersome and more prone to bias. Our findings are in congruence with the growing body of literature which demonstrates increased performance of ML models in the clinical setting.[16–18] Thus, ML algorithms should be considered as a viable and potentially superior alternative modeling approach in surgical care applications.

Although ML methods outperformed linear strategies for prediction of reoperation, the STS models outperformed ML for 30-day mortality and AKI. This observation is most attributable to the large sample used to derive the STS risk scores as well as the incorporation of over 100 data fields.[14] Nonetheless, the STS models are limited to operations either involving CABG or single valve replacement, and do not provide risk estimates for aortic surgery or multi-valve procedures. Such operations present a more heterogeneous risk profile and may reduce the performance of predictive models. We opted to include such operations in our modeling attempts to develop a tool which accurately reflects the case-mix at our five academic institutions. Indeed, procedures not accounted for by the STS comprised approximately 5% of our study cohort. Regardless, ML approaches are gradually being incorporated into the STS models to provide more bespoke estimates, an effort which will certainly improve risk prediction across cases performed in the United States.

The predictive models developed in the present work have considerable utility in the clinical and administrative settings. Their mode of application is tunable to an institution's needs, and the insights that they provide have the potential to enhance clinical outcomes. A landmark

randomized control trial by Shimabukuro and colleagues found the implementation of ML models to reduce ICU mortality and LOS, demonstrating that such tools can tangibly improve clinical outcomes and decrease resource utilization.[30] Our group has chosen to make the ML models with the greatest R^2 and AUC available for public use. This online tool may be used by clinicians when evaluating patient risk or by administrators who wish to apply our predictive model at the programmatic level. However, a model which continuously incorporates postoperative events into the estimated LOS would be most pertinent to patient care in the perioperative setting. Further efforts to develop such tools are warranted.

Given the premium placed on low-acuity and ICU beds during the COVID-19 pandemic, hospitals transiently reduced surgical volume. Prachand et al. proposed the widely used MeNTS framework, highlighting several key factors, such as OR time, estimated LOS, and anticipated blood loss, when determining resource allocation.[31] In the event of significant reduction in operating capacity, the development of algorithms that balance risk associated with delay in operative management as well as estimated resource use may be necessary. Our proposed ML based models may better inform decisions about scheduling and optimizing case mix to ensure sufficient hospital throughput. With wide availability of ML present and use of few explanatory variables, prospective studies may readily determine the pragmatic impact of such models in optimizing hospital efficiency.

The present study has several limitations. As a multi-center study confined to a group of academic centers, our findings are not generalizable to the cardiac surgical population at large. In addition, while the consortium makes a concerted effort to homogenize practice patterns across participating institutions, certain clinical factors may vary by center and surgeon, such as the

threshold for blood transfusion. Transfer status was similarly not captured in the UCCSC and could not be accounted for in our predictive models. Furthermore, despite the relatively large size of the dataset, prospective application of the ML models is required to externally validate their utility. Nonetheless, we used robust statistical methods and a sparse set of autonomously selected variables to enhance the generalizability.

In conclusion, we have demonstrated the superior performance of machine leaning models in providing accurate predictions for length of stay using a multi-institutional, cardiac surgical database. Derived from few variables, such models can estimate resource use and better inform projected hospital census. Leveraging the information derived from machine learning models may be especially useful in reducing the impact of pandemic related disruptions in cardiac surgical programs.

References

1. Gaudino M, Chikwe J, Hameed I, Robinson NB, Fremes SE, Ruel M. Response of Cardiac Surgery Units to COVID-19. *Circulation*. 2020;142(3):300-302. doi:10.1161/CIRCULATIONAHA.120.047865

2. Farrington WJ, Robinson NB, Rahouma M, et al. Cardiac Surgery Outcomes in an Epicenter of the COVID-19 Pandemic. *Semin Thorac Cardiovasc Surg*. Published online January 12, 2021. doi:10.1053/J.SEMTCVS.2021.01.005

3. Ad N, Luc JGY, Nguyen TC, et al. Cardiac surgery in North America and coronavirus disease 2019 (COVID-19): Regional variability in burden and impact. *J Thorac Cardiovasc Surg*. 2021;162(3):893-903.e4. doi:10.1016/J.JTCVS.2020.06.077

4. George I, Salna M, Kobsa S, et al. The rapid transformation of cardiac surgery practice in the coronavirus disease 2019 (COVID-19) pandemic: insights and clinical strategies from a centre at the epicentre. *Eur J Cardio-Thoracic Surg*. 2020;58(4):667-675. doi:10.1093/EJCTS/EZAA228

5. Bose SK, Dasani S, Roberts SE, et al. The Cost of Quarantine: Projecting the Financial Impact of Canceled Elective Surgery on the Nation's Hospitals. *Ann Surg*. 2021;273(5):844-849. doi:10.1097/SLA.0000000000004766

6. Salenger R, Etchill EW, Ad N, et al. The Surge After the Surge: Cardiac Surgery Post–COVID-19. *Ann Thorac Surg*. 2020;110(6):2020-2025. doi:10.1016/J.ATHORACSUR.2020.04.018

7. Engelman DT, Lother S, George I, et al. Ramping Up Delivery of Cardiac Surgery During the COVID-19 Pandemic: A Guidance Statement From The Society of Thoracic Surgeons COVID-19 Task Force. *Ann Thorac Surg*. 2020;110(2):712-717. doi:10.1016/J.ATHORACSUR.2020.05.002

8. O'Brien SM, Feng L, He X, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 2—Statistical Methods and Results. *Ann Thorac Surg*. 2018;105(5):1419-1428. doi:10.1016/J.ATHORACSUR.2018.03.003

9. Chan V, Ahrari A, Ruel M, Elmistekawy E, Hynes M, Mesana TG. Perioperative Deaths After Mitral Valve Operations May Be Overestimated by Contemporary Risk Models. *Ann Thorac Surg*. 2014;98(2):605-610. doi:10.1016/J.ATHORACSUR.2014.05.011

10. Kirmani BH, Mazhar K, Saleh HZ, et al. External validity of the Society of Thoracic Surgeons risk stratification tool for deep sternal wound infection after cardiac surgery in a UK population. *Interact Cardiovasc Thorac Surg*. 2013;17(3):479-484. doi:10.1093/ICVTS/IVT222

11. Sharkawi MA, Shah PB, Zenati M, Kaneko T, Ramadan R. Structural Heart Underclassification of Predicted Risk of Mortality Using the Latest Society of Thoracic Surgeons Risk Models. Published online 2021. doi:10.1080/24748706.2021.1902596

12. Daghistani TA, Elshawi R, Sakr S, Ahmed AM, Al-Thwayee A, Al-Mallah MH.

Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *Int J Cardiol*. 2019;288:140-147. doi:10.1016/J.IJCARD.2019.01.046

13. Alshakhs F, Alharthi H, Aslam N, Khan IU, Elasheri M. Predicting Postoperative Length of Stay for Isolated Coronary Artery Bypass Graft Patients Using Machine Learning. *Int J Gen Med*. 2020;13:751. doi:10.2147/IJGM.S250334

14. Shahian DM, Jacobs JP, Badhwar V, et al. The Society of Thoracic Surgeons 2018 Adult Cardiac Surgery Risk Models: Part 1—Background, Design Considerations, and Model Development. *Ann Thorac Surg*. 2018;105(5):1411-1418. doi:10.1016/J.ATHORACSUR.2018.03.002

15. Messaoudi N, Cocker J De, Stockman B, Bossaert LL, Rodrigus IER. Prediction of Prolonged Length of Stay in the Intensive Care Unit After Cardiac Surgery: The Need for a Multi-institutional Risk Scoring System. *J Card Surg*. 2009;24(2):127-133. doi:10.1111/J.1540-8191.2008.00716.X

16. Kilic A, Goyal A, Miller JK, et al. Predictive Utility of a Machine Learning Algorithm in Estimating Mortality Risk in Cardiac Surgery. *Ann Thorac Surg*. 2020;109(6):1811-1819. doi:10.1016/J.ATHORACSUR.2019.09.049

17. Tseng P-Y, Chen Y-T, Wang C-H, et al. Prediction of the development of acute kidney injury following cardiac surgery by machine learning. *Crit Care 2020 241*. 2020;24(1):1-13. doi:10.1186/S13054-020-03179-9

18. Lee H-C, Yoon H-K, Nam K, et al. Derivation and Validation of Machine Learning Approaches to Predict Acute Kidney Injury after Cardiac Surgery. *J Clin Med 2018, Vol 7, Page 322*. 2018;7(10):322. doi:10.3390/JCM7100322

19. Adult Cardiac Surgery Database Data Collection | STS. Accessed September 9, 2021. https://www.sts.org/registries-research-center/sts-national-database/adult-cardiac-surgery-database/data-collection

20. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min*. doi:10.1145/2939672

21. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. *Adv Neural Inf Process Syst*. 2017;30. Accessed September 10, 2021. https://github.com/slundberg/shap

22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830. Accessed September 10, 2021. http://jmlr.org/papers/v12/pedregosa11a.html

23. Almashrafi A, Elmontsri M, Aylin P. Systematic review of factors influencing length of stay in ICU after adult cardiac surgery. *BMC Heal Serv Res 2016 161*. 2016;16(1):1-12. doi:10.1186/S12913-016-1591-3

24. Tzeng C-WD, Teshome M, Katz MHG, et al. Cancer Surgery Scheduling During and After the COVID-19 First Wave: The MD Anderson Cancer Center Experience. *Ann Surg*.

2020;272(2):e106. doi:10.1097/SLA.0000000000004092

25. Tran Z, Chervu N, Williamson C, et al. The Impact of Expedited Discharge on 30-Day Readmission Following Lung Resection: A National Study. *Ann Thorac Surg*. Published online April 18, 2021. doi:10.1016/J.ATHORACSUR.2021.04.009

26. Afflu DK, Seese L, Sultan I, et al. Very Early Discharge After Coronary Artery Bypass Grafting Does Not Affect Readmission or Survival. *Ann Thorac Surg*. 2021;111(3):906-913. doi:10.1016/J.ATHORACSUR.2020.05.159

27. Voeten DM, van der Werf LR, Gisbertz SS, et al. Postoperative intensive care unit stay after minimally invasive esophagectomy shows large hospital variation. Results from the Dutch Upper Gastrointestinal Cancer Audit. *Eur J Surg Oncol*. 2021;47(8):1961-1968. doi:10.1016/J.EJSO.2021.01.005

28. Triana AJ, Vyas R, Shah AS, Tiwari V. Predicting Length of Stay of Coronary Artery Bypass Grafting Patients Using Machine Learning. *J Surg Res*. 2021;264:68-75. doi:10.1016/J.JSS.2021.02.003

29. LaFaro RJ, Pothula S, Kubal KP, et al. Neural Network Prediction of ICU Length of Stay Following Cardiac Surgery Based on Pre-Incision Variables. *PLoS One*. 2015;10(12):e0145395. doi:10.1371/JOURNAL.PONE.0145395

30. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. *BMJ Open Respir Res*. 2017;4(1):e000234. doi:10.1136/BMJRESP-2017-000234

31. Prachand VN, Milner R, Angelos P, et al. Medically Necessary, Time-Sensitive Procedures: Scoring System to Ethically and Efficiently Manage Resource Scarcity and Provider Risk During the COVID-19 Pandemic. *J Am Coll Surg*. 2020;231(2):281-288. doi:10.1016/J.JAMCOLLSURG.2020.04.011

Tables:

Table 1: Baseline patient characteristics of the study cohort. *CABG: Coronary Artery Bypass Grafting; SD: Standard Deviation.*

Parameter	Overall n=6,316	Derivation n=5,028	Validation n=1,288	P-Value
Age (years, mean±SD)	63±13	63±13	64±13	<0.001
Elective Admission (%)	58.5	58.3	59.3	0.52
Female (%)	27.5	27.7	26.6	0.45
Height (centimeters, mean±SD)	171±11	171±11	171±10	0.29
Weight (kilograms, mean±SD)	82±19	82±19	81±20	0.57
Ethnicity (%)	19.7	19.9	19.3	0.68
Operative Type (%)				
Isolated CABG	50.5	51.3	47.4	0.012
Isolated Valve Operation	33.3	31.3	41.2	<0.001
CABG + Valve	10.6	11.1	8.3	0.003
Multiple Valve	5.8	6.3	3.6	<0.001
Medical Conditions (%)				
Atrial Fibrillation	17.6	17.6	17.7	0.91
Cancer	6.9	7.1	6.4	0.37
Cerebrovascular Disease	17.0	17.2	16.2	0.38
Severe Lung Disease	3.3	3.2	3.9	0.23
Congestive Heart Failure	36.1	33.8	45.0	<0.001
Diabetes	38.1	37.6	39.7	0.18
Home Oxygen	3.0	3.1	2.8	0.59
Hypertension	77.4	77.2	78.0	0.57
Infectious Endocarditis	5.9	5.9	6.1	0.84
Liver Disease	6.4	6.8	5.0	0.017
Peripheral Vascular Disease	9.0	8.5	11.1	0.003
Prior Myocardial Infarction	31.2	31.7	29.3	0.09
Laboratory Values (mean±SD)				
Hematocrit (% blood volume)	39±6	39±6	39±6	0.01
International Normalized Ratio	1.13±0.3	1.13±0.3	1.12±0.2	0.26
Serum Albumin (g/dL)	3.9±0.6	3.9±0.6	3.9±0.6	0.008
Preoperative Creatinine (mg/dL)	1.4±1.7	1.4±1.6	1.5±1.9	<0.001
Ejection Fraction (%)	56±12	56±12	57±12	0.21
Hospital of Operation (%)				
Center 1	32.7	31.7	36.5	0.001
Center 2	24.0	24.7	21.4	0.014
Center 3	19.1	19.1	18.9	0.84
Center 4	14.0	14.9	10.5	<0.001
Center 5	10.2	9.5	12.7	<0.001

Table 2: Resource utilization and clinical outcomes stratified by derivation and validation cohorts.

Outcome	Overall n=6,316	Derivation n=5,028	Validation n=1,288	P-Value
Resource Use (median, IQR)				
Length of Stay (days)	8 [6-13]	8 [6-13]	8 [5-12]	0.008
ICU Length of Stay (hours)	74 [47-116]	75 [47-117]	68 [43-99]	<0.001
Clinical Endpoints (%)				
Mortality	0.9	1.0	0.7	0.39
Acute Kidney Injury	1.5	1.5	1.7	0.54
Postoperative Transfusion	27.7	28.8	23.1	<0.001
Reoperation	8.6	9.1	6.9	0.014

Table 3: Cross-validated model performance metrics for each algorithm and outcome. Reported as means with 95% confidence intervals. *[a]Models with continuous output were evaluated using the coefficient of determination (R^2), while binary classifiers[b] were assessed with the area under the receiver operating characteristic (AUC). GBM: Gradient Boosted Machine. ICU: Intensive Care Unit. STS: Society of Thoracic Surgeons Risk Score. RF: Random Forest. XGBoost: Extreme Gradient Boosting.*

Outcome	Linear	Logistic	GBM	RF	XGBoost	S
Resource Use (R2, 95% CI)						
Length of Stay	0.41 (0.41-0.41)	-	0.42 (0.42-0.42)	0.41 (0.40-0.41)	0.42 (0.42-0.42)	
ICU Length of Stay	0.15 (0.15-0.15)	-	0.23 (0.23-0.23)	0.21 (0.21-0.21)	0.22 (0.22-0.22)	
Clinical Endpoint (AUC, 95% CI)						
Mortality	-	0.67 (0.67-0.68)	0.69 (0.68-0.70)	0.69 (0.68-0.70)	0.69 (0.69-0.70)	0.91 (0
Acute Kidney Injury	-	0.67 (0.67-0.68)	0.76 (0.75-0.77)	0.76 (0.76-0.77)	0.74 (0.73-0.75)	0.84 (0
Postoperative Transfusion	-	0.71 (0.71-0.72)	0.73 (0.73-0.73)	0.71 (0.71-0.71)	0.73 (0.73-0.74)	
Reoperation	-	0.81 (0.80-0.81)	0.8 (0.79-0.80)	0.80 (0.80-0.80)	0.79 (0.79-0.80)	0.76 (0

Table 4: Performance of each algorithm when predicting resource utilization and clinical outcomes in the validation cohort. *Regressions[a] were evaluated using the coefficient of determination (R^2), while binary classifiers[b] were assessed with the area under the receiver operating characteristic (AUC). GBM: Gradient Boosted Machine. ICU: Intensive Care Unit. STS: Society of Thoracic Surgeons Risk Score. RF: Random Forest. XGBoost: Extreme Gradient Boosting.*

Outcome	**Linear**	**Logistic**	**GBM**	**RF**	**XGBoost**	**STS**
Resource Use (R^2)						**-**
Length of Stay	0.42	-	0.47	0.47	0.47	**-**
ICU Length of Stay	0.017	-	0.078	0.054	0.082	
Clinical Endpoint (AUC)						
Mortality	-	0.68	0.68	0.7	0.72	0.91
Acute Kidney Injury	-	0.77	0.79	0.8	0.8	0.84
Postoperative Transfusion	-	0.69	0.68	0.68	0.67	**-**
Reoperation	-	0.78	0.79	0.8	0.78	0.76

Figures:

Figure 1: Coefficient of determination (R^2) versus covariate set size in the prediction of in-hospital length of stay. *GBM: Gradient Boosted Machine; LR: Linear Regression.*

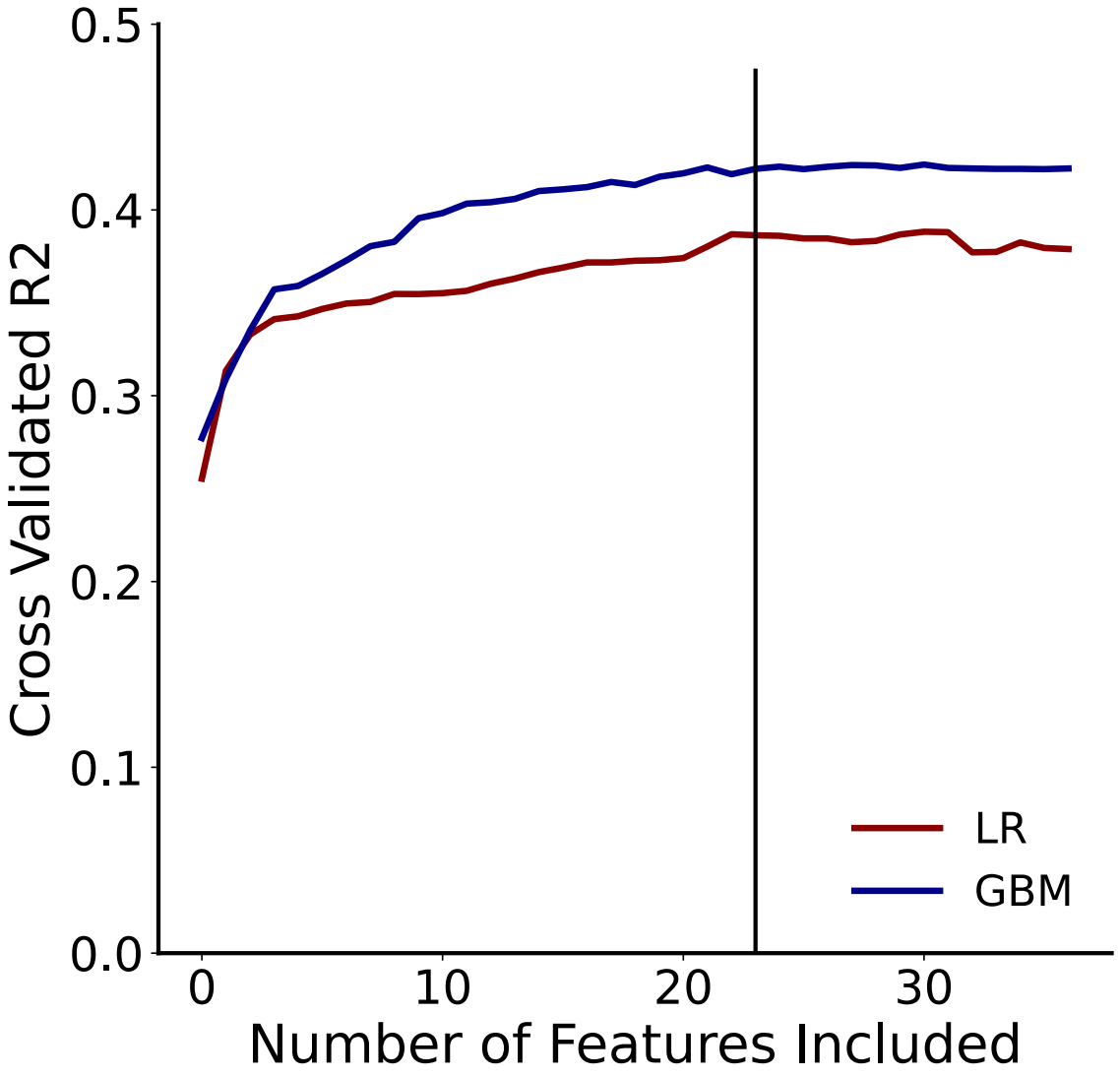


Figure 2: Interpretation of GBM-based model for prediction of length of stay (LOS, days) using SHapley summary plots. Y axis is ordered by increasing feature importance, and the x-axis is the marginal effect of each parameter on predicted LOS. Red dots show the impact of high feature values on predicted LOS, while blue dots show the impact of low feature values.

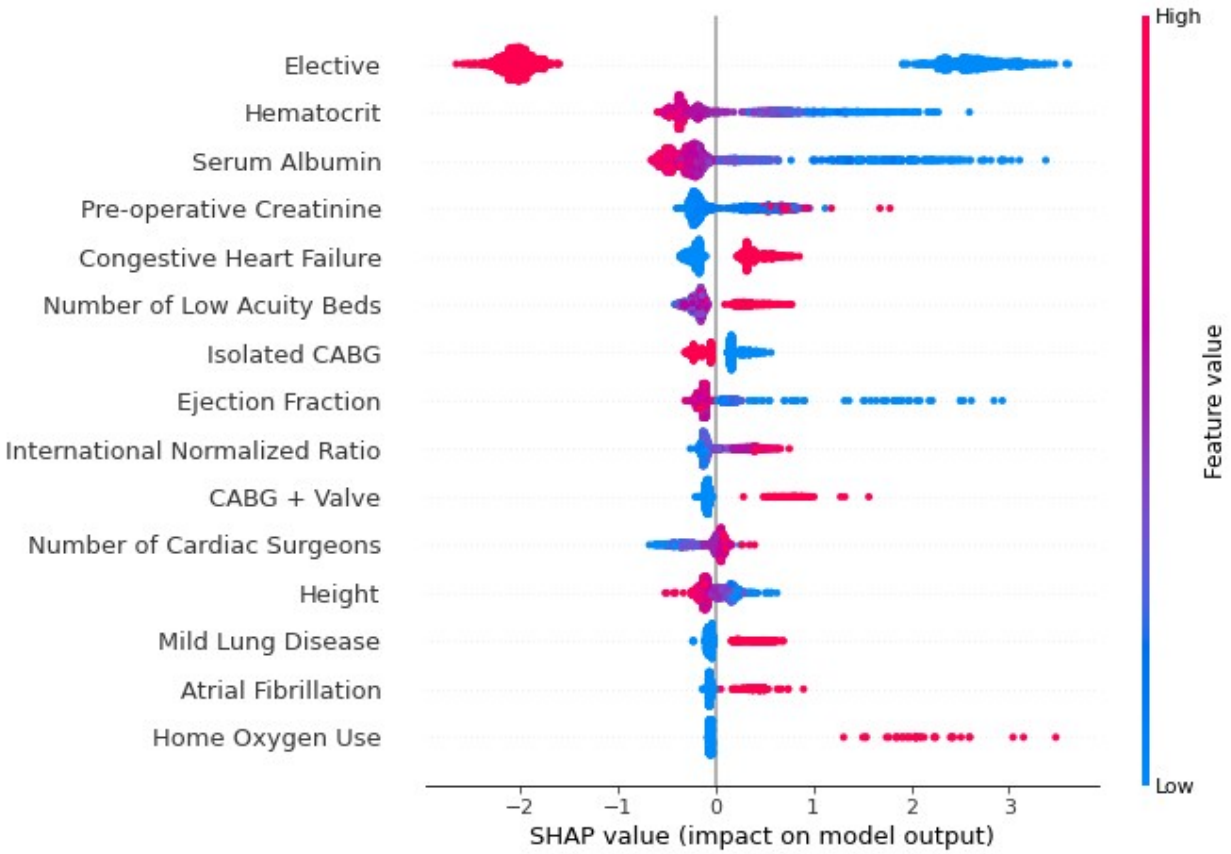
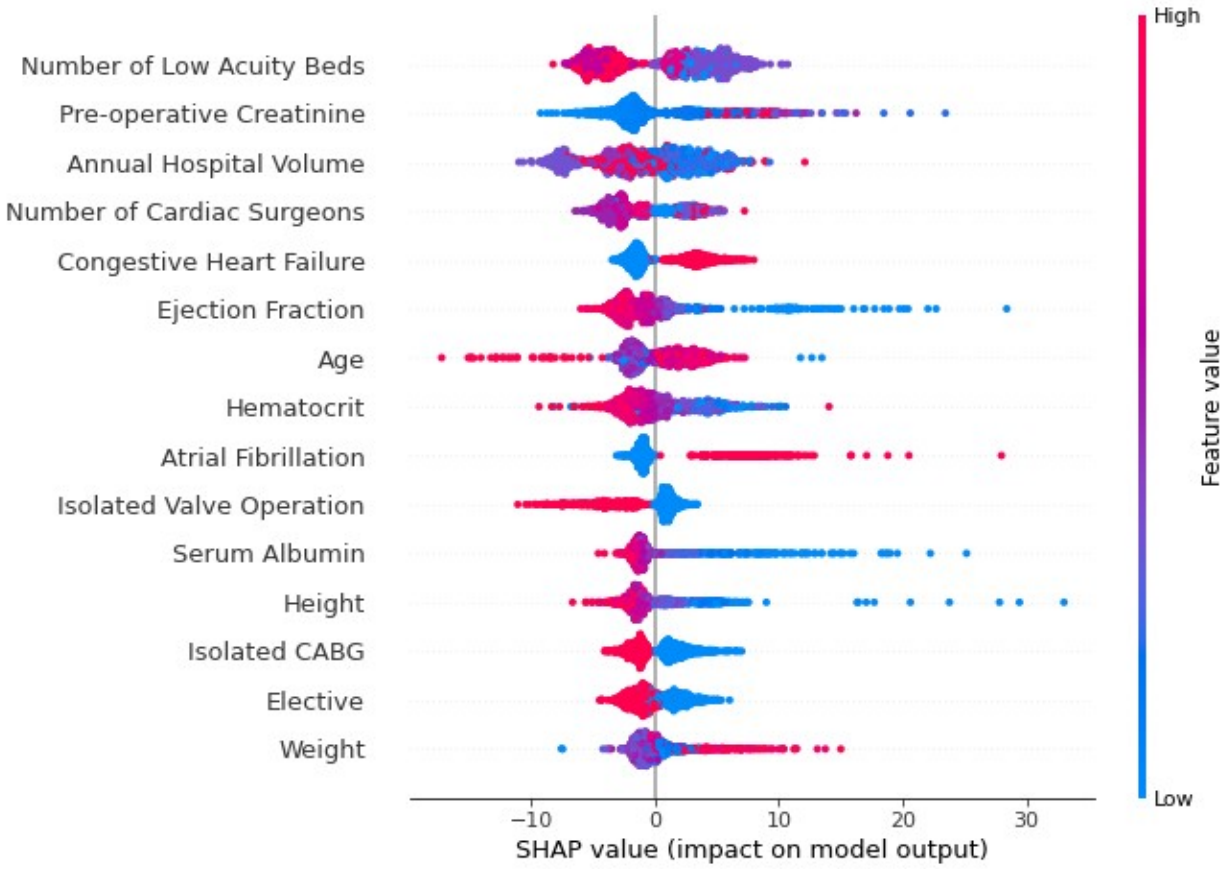
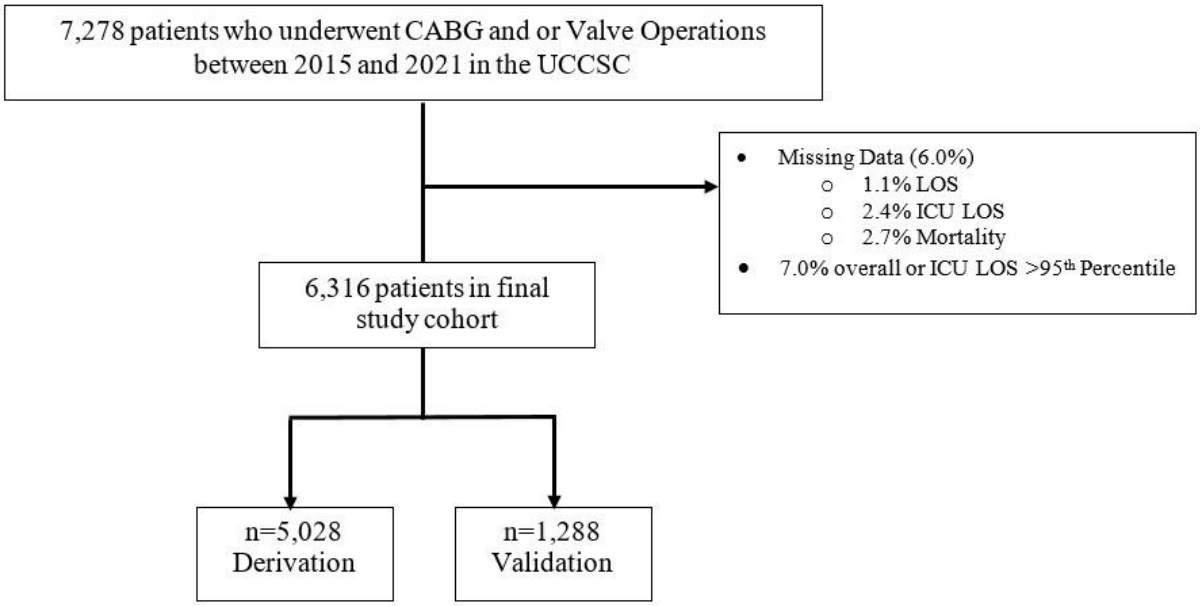


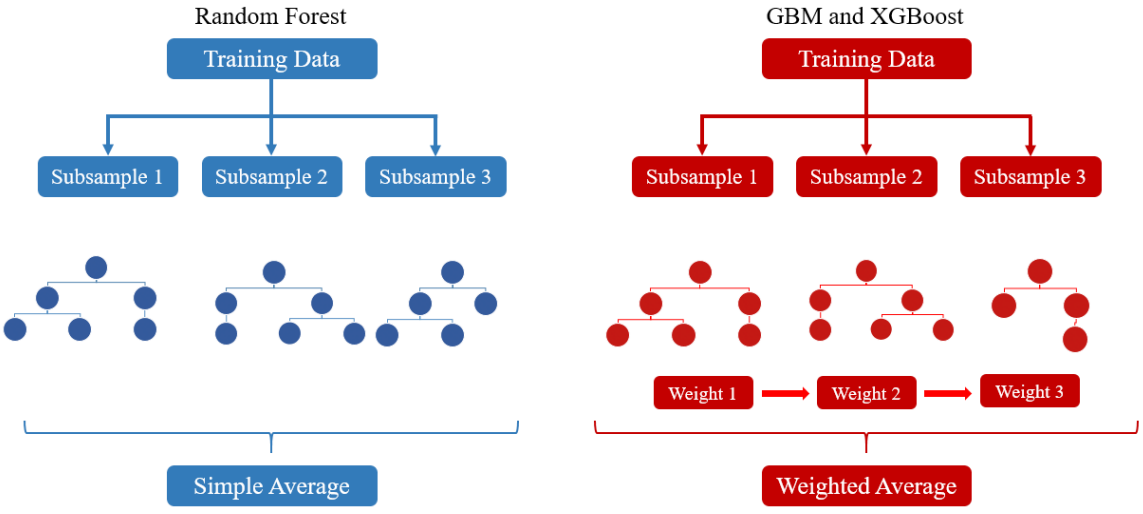
Figure 3: Interpretation of GBM-based model for prediction of intensive care unit length of stay (ICU LOS, hours) using SHapley summary plots. Y axis is ordered by increasing feature importance, and the x-axis is the marginal effect of each parameter on predicted ICU LOS. Red dots show the impact of high feature values on predicted ICU LOS, while blue dots show the impact of low feature values.



Supplemental Figure 1: Study CONSORT diagram. *CABG: Coronary Artery Bypass Grafting.*



Supplemental Figure 2: Schematic representing the algorithmic design of random forest, gradient boosted machines (GBM) and extreme gradient boosting (XGBoost).



Supplemental Figure 3: Calibration plot of observed versus predicted length of stay in days. *R^2: Coefficient of Determination; GBM: Gradient Boosted Machine; LR: Linear Regression.*