Prediction of Mortality and Allograft Loss in Pediatric Heart Transplant Patients Using Machine Learning

# METHODS

## Modeling algorithms

We define the term ‘modeling algorithm’ as the collection of steps taken to develop a prediction model. We compared modeling algorithms based on the following characteristics: (1) number of predictors used in the model, (2) method to select the specified number of predictors, (3) method to develop a prediction model using the selected predictors. We considered seven options for the number of predictors, three options for variable selection, and four options for model development (Figure 1). All possible combinations of these options were assessed, leading to the evaluation of 84 unique modeling algorithms.

**Methods to select predictors:**

We assess contribution importance, permutation importance, and stepwise importance. Contribution importance represents fractional contribution of each predictor to the model based on the total gain of this feature’s splits (1). Higher percentage means a more important predictive feature. Permutation importance measures the importance of each predictor by measuring how much the model’s prediction error increases when the values of the predictor are randomly permuted (2). Stepwise importance follows the traditional stepwise technique that includes or excludes variables one at a time until a stopping criteria is met (3). Though it is not a valid procedure for statistical inference, stepwise modeling algorithms can be useful for prediction (4). There is no numeric importance value for stepwise importance but a subset of the most important variables is easily obtained by using a pre-specified number of steps.

**Methods to develop a risk prediction model:**

Proportional hazards regression is a semi-parametric modeling technique that estimates a baseline hazard function and assumes predictors have a multiplicative effect on the hazard for an event of interest. Gradient boosting (hereafter referred to as boosting) develops an ensemble of weak prediction models (1). Each learner in the ensemble attempts to correct errors from the previous learners, and the ensemble’s prediction is the aggregate of its individual learners’ predictions. We developed boosting models using decision trees as learners (5), a technique that has been recognized in numerous settings to be state of the art for statistical prediction (6). Random forests are also ensembles of decision trees, but in contrast to boosting, the random forest comprises de-correlated decision trees that are grown using bootstrapped replicates of the original data (2, 7). We applied the standard random forest algorithm for risk prediction, which grows decision tree nodes by splitting the data based on an individual variable (8). We also applied an extension of the random forest that splits data using linear combinations of variables (9). Random forests grown using linear combinations of input variables are referred to as ‘oblique’ random forests.

## Internal validation

We applied Monte-Carlo cross-validation to identify the most effective modeling algorithm for developing a final prediction model and to estimate the performance of modeling algorithms when they are applied to new data (10). Monte-Carlo cross-validation is an extension of split-sample testing, which splits available data into training and testing sets, then develops a model using the training set and validates the model in the testing set. Monte-Carlo cross-validation replicates this procedure, using a different split of the available data for each replicate, leading to a reduction in the variance of performance estimates for the modeling algorithms assessed (11). In the current analysis, we completed a total of 500 replications of split-sample testing.

## Measures of model performance

Model performance was evaluated based on discrimination and calibration, as recommended by published guidelines (12, 13). Discrimination and calibration were measured using a time-dependent concordance (C-) statistic and modified D’Agostino-Nam test, respectively (14, 15). Point estimates and 95% confidence intervals (CIs) for these metrics were computed empirically using the distribution of results from Monte-Carlo cross-validation. Point estimates were the median value of a performance metric, while the 2.5th and 97.5th percentiles defined lower and upper bounds for 95% CIs, respectively. We did not find previously published prediction models for graft loss among pediatric patients, so we did not analyze net reclassification improvement as it requires a baseline prediction model (16).

## Missing data

Missing values were imputed after data were split into training and testing sets during each replicate of Monte-Carlo cross-validation. All information from the testing data was withheld during this process (17). Specifically, missing values were imputed in both the training and testing data using the mean and mode of continuous and categorical variables, respectively, computed using the training data. While imputation to the mean is not appropriate for statistical inference, this technique has been shown to produce prediction models with Bayes consistency when missing values are non-informative (18).

## Bayesian analysis of model performance

We applied Bayesian hierarchical models to draw inferences regarding the expected performance of modeling algorithms, accounting for correlated performance within each replicate of Monte-Carlo cross-validation (19). Specifically, we estimated the posterior probability that a given technique to fit a risk prediction model had superior discrimination or calibration compared to other techniques. All comparisons were made holding the number of predictor variables and the method to select predictor variables fixed; i.e., both models used the same predictor variables.

## Statistical analysis

We conducted the current analysis following previously published guidelines on multivariable prediction models for individual prognosis and diagnosis (20). The 1 year incidence of graft loss or morality was computed overall and by transplant year. Characteristics of patients were calculated as mean with standard deviation or percent in the overall population and stratified by transplant year. Using the results from internal validation, we visually assessed the discrimination and calibration of candidate modeling algorithms to develop a final prediction model. We used posterior predicted probability to perform inference on pairwise comparisons between select candidate modeling algorithms. Based on the results, we selected a final prediction modeling algorithm and applied it to all the available data. We tabulated a summary of the variables included into the final prediction model - including the numeric importance value and rank along with the count and percent of how many missing values were present. We used partial dependence to estimate multivariable adjusted predicted risk as a function of each variable included in the final prediction model, separately (21).

Analyses were conducted using SAS version 9.4, R version 4.0.3, and a number of open-source R packages (22–26). All R code for the current analysis is publicly available at <https://github.com/bcjaeger/graft-loss>. Data for the current analysis are available by request from **FILL IN**.

# RESULTS

Overall, the 1-year incidence of graft loss or mortality after transplant was 0.08 (0.07, 0.09). The incidence was highest during 2014 and lowest during 2016, with incidence rates (95% CI) of 0.10 (0.07, 0.14) and 0.06 (0.04, 0.09), respectively. The overall sample was 55.0% male, 65.1% white, and had mean (standard deviation) age of 6.50 (6.20) years at the time of transplant (**Table 1**).

## Internal validation

Internally validated estimates of the C-statistic showed a progressive increase in model discrimination as more predictors were included, except when proportional hazards regression was applied to develop the risk prediction model (**Figure 2**). Additionally, when stepwise variable importance was used to select predictors, model discrimination was lower compared to permutation and contribution importance. The oblique random survival forest obtained the highest C-statistic in 15 of 21 comparisons, and obtained the highest overall C-statistic of 0.78 (95% CI 0.76, 0.80) using 75 predictors selected by permutation importance. As measuring 75 predictors may be tedious for clinical practice, we focus on results when 20 predictors were included in the prediction model.

Modeling algorithms using permutation importance to select predictors obtained superior calibration compared to their counterparts using contribution importance (**Figure 3**). For example, when 20 predictors were selected using permutation and contribution importance prior to fitting an oblique random survival forest, the median (95% CI) p-values for mis-calibration were 0.41 (0.16, 0.66) and 0.33 (0.11, 0.60), respectively.

**Bayesian analysis of model performance**: As discrimination and calibration were generally higher when permutation importance was applied to select 20 predictor variables, we focused the Bayesian analysis of model performance to this setting (**Figure 4**). Oblique random survival forests obtained the highest C-statistic, 0.763 and p-value for miscalibration, 0.422. The posterior probability that oblique random survival forests obtained superior discrimination versus other modeling algorithms ranged from 0.7 (versus boosting) to 0.93 (versus proportion hazards). The posterior probability that oblique random survival forests obtained superior calibration versus other modeling algorithms ranged from 0.61 (versus standard random forests) to 0.85 (versus boosting).

**Selection of the final modeling algorithm**: Under this setting, using permutation importance to select predictors and oblique random survival forests to fit a risk prediction model led to the best discrimination (C = 0.76 and calibration (P-value for mis-calibration = 0.41). Therefore, we fit our final prediction model using the oblique random survival forest after selecting 20 predictors using permutation importance.

## Model summary

**Table 1:** Characteristics of patients included in the current analysis.

| **Variable** | **Overall** | **Transplant year** | | |
| --- | --- | --- | --- | --- |
| **Before 2014** | **2014 through 2016** | **After 2017** |
| F1 Recipient Age (years) at Listing1 | 6.50 (6.20) | 6.36 (6.20) | 6.44 (6.19) | 6.73 (6.21) |
| F0 Male recipient | 55.0% | 52.9% | 56.1% | 56.6% |
| F0 Race | | | | |
| black | 17.1% | 18.5% | 15.9% | 16.5% |
| other | 17.9% | 14.9% | 18.0% | 21.5% |
| white | 65.1% | 66.6% | 66.1% | 61.9% |
| F0 Recipient Hispanic or Latino | 20.1% | 15.4% | 22.0% | 24.2% |
| F0 Primary Etiology | | | | |
| other | 2.48% | 3.64% | 1.56% | 2.09% |
| cardiomyopathy | 49.6% | 52.0% | 49.8% | 46.2% |
| congenital\_hd | 48.0% | 44.4% | 48.7% | 51.7% |
| F1 Surgeries Prior to Listing | 46.9% | 44.4% | 47.6% | 49.4% |
| F1 Medical History at Listing | 83.8% | 99.6% | 83.9% | 68.6% |
| F1T ECMO at Transplant | 4.30% | 4.50% | 4.74% | 3.54% |
| F1T No MSCD at Transplant | 69.7% | 72.1% | 71.5% | 64.5% |
| 1Table values are mean (standard deviation) and percent for continuous and categorical variables, respectively. | | | | |

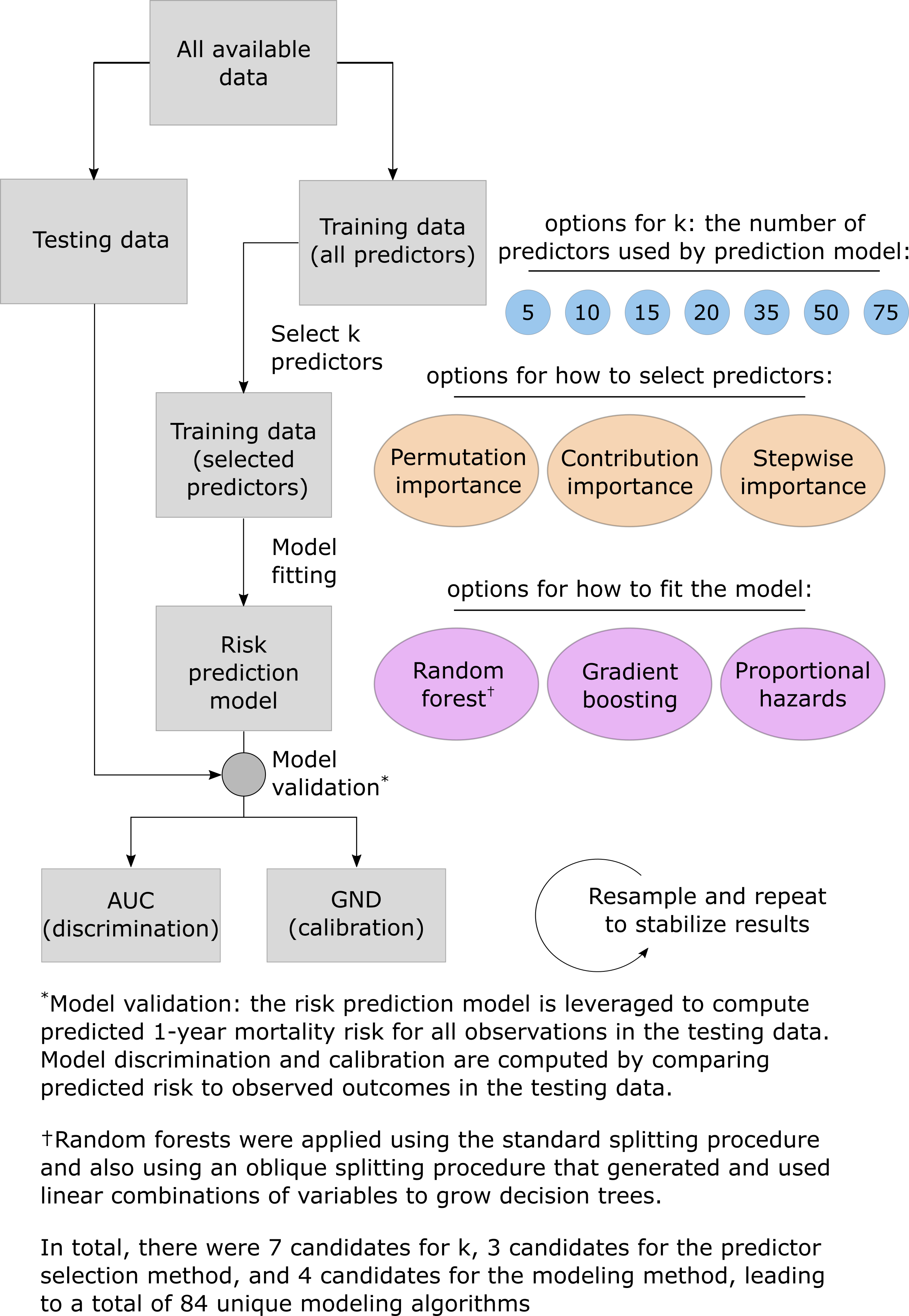
**Table 2:** Variable missingness and importance for predictors used in the final prediction model.

| **Predictor variable** | **Missing values** | | | **Permutation Importance** | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Number** | **Percent** | **Imputed to** | **Change in C-statistic** | **Scaled** | **Rank** |
| F1T Cardiopulmonary bypass time (minutes) | 158 | 4.17 | 161 | 0.00865 | 1.00 | 1 |
| F0 Primary Etiology | 0 | 0.00 | cardiomyopathy | 0.00754 | 0.87 | 2 |
| F1T ECMO at Transplant | 0 | 0.00 | no | 0.00601 | 0.69 | 3 |
| F1 Surgeries Prior to Listing | 8 | 0.21 | no | 0.00409 | 0.47 | 4 |
| F1T No MSCD at Transplant | 0 | 0.00 | yes | 0.00343 | 0.40 | 5 |
| F0 CHD: Single Ventricle | 156 | 4.12 | no | 0.00337 | 0.39 | 6 |
| F1T Transplant Lab: Serum Albumin g/dL | 422 | 11.1 | 3.70 | 0.00176 | 0.20 | 7 |
| F1 Medical History at Listing | 533 | 14.1 | yes | 0.00175 | 0.20 | 8 |
| F1 Recipient Height(inches) at Listing | 83 | 2.19 | 39.0 | 0.00170 | 0.20 | 9 |
| F1 PRA Max List | 326 | 8.61 | 0.00 | 0.00169 | 0.20 | 10 |
| F1 Listing Lab: BUN mg/dL | 176 | 4.65 | 16.0 | 0.00165 | 0.19 | 11 |
| F0 CHD: Hypoplastic Left Heart | 860 | 22.7 | no | 0.00160 | 0.19 | 12 |
| F1T Recipient Height (inches) at Transplant | 64 | 1.69 | 40.2 | 0.00155 | 0.18 | 13 |
| F1 Surgeries Prior to Listing: Norwood Stage I: BT Shunt | 0 | 0.00 | no | 0.00134 | 0.16 | 14 |
| F1T Recipient Weight (pounds) at Transplant | 7 | 0.18 | 35.3 | 0.00118 | 0.14 | 15 |
| F1T BSA (m2) at Transplant | 65 | 1.72 | 0.68 | 0.00116 | 0.13 | 16 |
| F1T Transplant Lab: BUN mg/dL | 102 | 2.69 | 16.0 | 0.00108 | 0.13 | 17 |
| F1 Recipient Age (years) at Listing | 0 | 0.00 | 4.31 | 0.00101 | 0.12 | 18 |
| F1 BSA (m2) at Listing | 83 | 2.19 | 0.64 | 0.00087 | 0.10 | 19 |

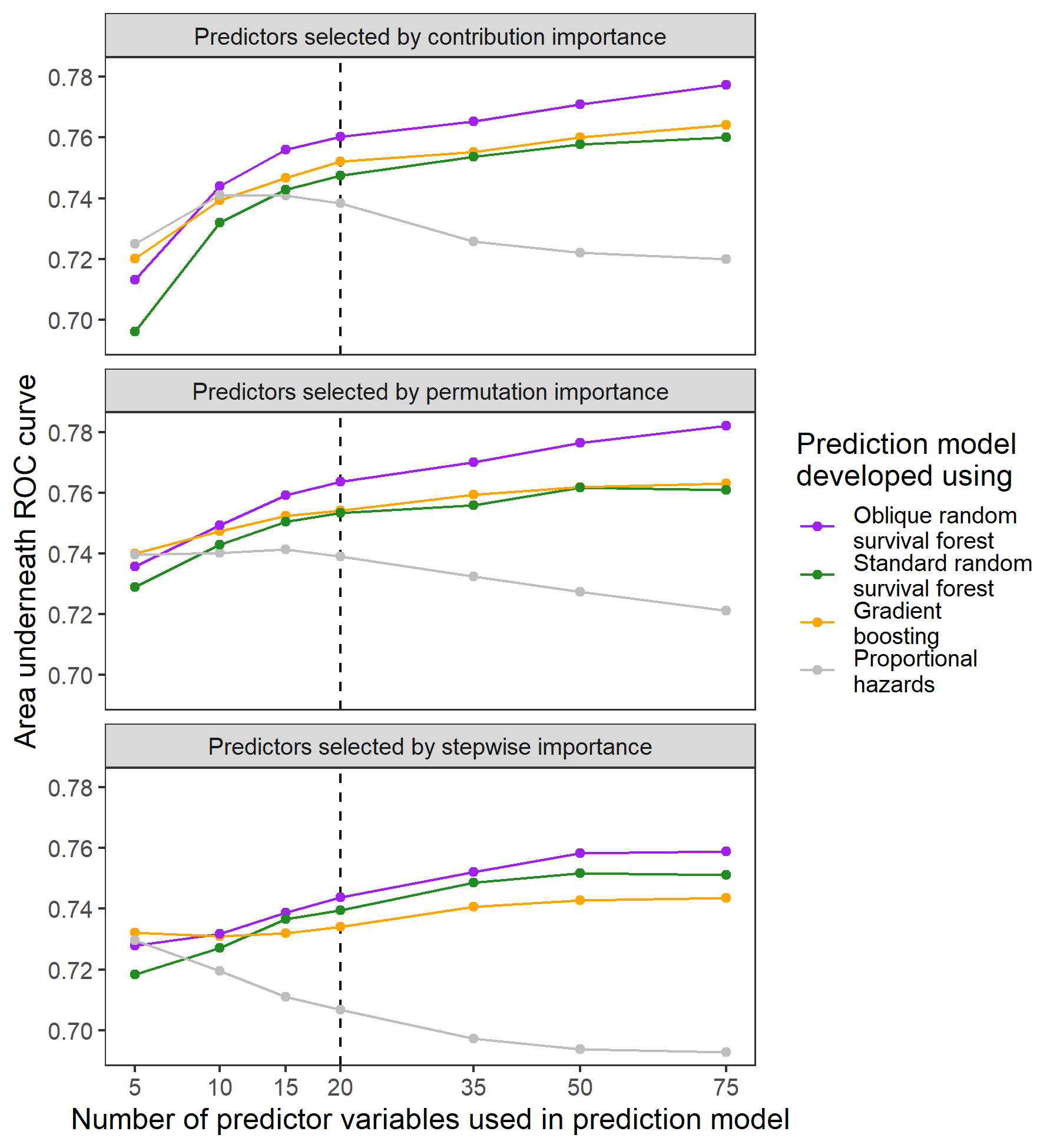
**Table 3:** Predicted risk for graft loss or mortality at 1 year following transplant according to the final prediction model from the current analysis.

|  | **1-year predicted risk for graft loss or mortality** | | |
| --- | --- | --- | --- |
| **Characteristic** | **Risk (95% CI)** | **Ratio (95% CI)** | **Difference (95% CI)** |
| F1T Cardiopulmonary bypass time (minutes) | | | |
| 25th percentile: 125 | 4.12 (3.98, 4.26) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 161 | 5.40 (5.13, 5.57) | 1.31 (1.27, 1.35) | 1.29 (1.10, 1.45) |
| 75th percentile: 209 | 5.85 (5.47, 6.15) | 1.42 (1.37, 1.45) | 1.73 (1.52, 1.90) |
| F0 Primary Etiology | | | |
| Other | 7.64 (7.43, 7.98) | 1 (Reference) | 0 (Reference) |
| Cardiomyopathy | 4.36 (4.10, 4.53) | 0.57 (0.55, 0.58) | -3.28 (-3.44, -3.12) |
| Congenital heart disease | 7.47 (7.27, 7.61) | 0.98 (0.96, 1.00) | -0.17 (-0.32, 0.04) |
| F1T ECMO at Transplant | | | |
| No | 4.68 (4.44, 5.04) | 1 (Reference) | 0 (Reference) |
| Yes | 17.3 (17.1, 17.7) | 3.70 (3.59, 3.84) | 12.7 (12.5, 13.0) |
| F1 Surgeries Prior to Listing | | | |
| No | 4.90 (4.62, 5.20) | 1 (Reference) | 0 (Reference) |
| Yes | 6.13 (6.02, 6.31) | 1.25 (1.20, 1.28) | 1.23 (1.02, 1.35) |
| F1T No MSCD at Transplant | | | |
| No | 7.34 (6.99, 7.57) | 1 (Reference) | 0 (Reference) |
| Yes | 4.21 (3.97, 4.44) | 0.57 (0.55, 0.59) | -3.14 (-3.35, -2.95) |
| F0 CHD: Single Ventricle | | | |
| No | 4.89 (4.69, 5.15) | 1 (Reference) | 0 (Reference) |
| Yes | 5.83 (5.62, 6.05) | 1.19 (1.17, 1.21) | 0.94 (0.83, 1.02) |
| F1T Transplant Lab: Serum Albumin g/dL | | | |
| 25th percentile: 3.30 | 5.09 (4.82, 5.25) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 3.70 | 4.74 (4.56, 4.98) | 0.93 (0.92, 0.95) | -0.35 (-0.44, -0.25) |
| 75th percentile: 4.10 | 4.65 (4.44, 4.82) | 0.91 (0.89, 0.93) | -0.44 (-0.57, -0.33) |
| F1 Medical History at Listing | | | |
| No | 3.78 (3.64, 3.91) | 1 (Reference) | 0 (Reference) |
| Yes | 5.10 (4.87, 5.31) | 1.35 (1.32, 1.38) | 1.32 (1.20, 1.42) |
| F1 Recipient Height(inches) at Listing | | | |
| 25th percentile: 25.2 | 6.12 (5.79, 6.41) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 39.0 | 5.37 (5.19, 5.57) | 0.88 (0.87, 0.89) | -0.76 (-0.84, -0.68) |
| 75th percentile: 58.7 | 5.03 (4.86, 5.15) | 0.82 (0.80, 0.84) | -1.10 (-1.22, -0.98) |
| F1 PRA Max List | | | |
| 25th percentile: 0.00 | 4.55 (4.41, 4.78) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 0.00 | 4.55 (4.29, 4.75) | 1.00 (1.00, 1.00) | 0.00 (0.00, 0.00) |
| 75th percentile: 10.0 | 5.09 (4.89, 5.37) | 1.12 (1.10, 1.14) | 0.54 (0.46, 0.63) |
| F1 Listing Lab: BUN mg/dL | | | |
| 25th percentile: 12.0 | 4.72 (4.44, 4.94) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 16.0 | 4.58 (4.32, 4.80) | 0.97 (0.96, 0.98) | -0.14 (-0.20, -0.07) |
| 75th percentile: 22.0 | 4.50 (4.34, 4.77) | 0.95 (0.94, 0.97) | -0.22 (-0.31, -0.12) |
| F0 CHD: Hypoplastic Left Heart | | | |
| No | 4.81 (4.54, 5.02) | 1 (Reference) | 0 (Reference) |
| Yes | 5.41 (5.13, 5.62) | 1.13 (1.11, 1.16) | 0.60 (0.53, 0.71) |
| F1T Recipient Height (inches) at Transplant | | | |
| 25th percentile: 26.8 | 6.22 (5.96, 6.44) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 40.2 | 5.42 (5.04, 5.56) | 0.87 (0.85, 0.89) | -0.80 (-0.92, -0.69) |
| 75th percentile: 59.1 | 5.03 (4.81, 5.20) | 0.81 (0.79, 0.83) | -1.19 (-1.28, -1.00) |
| F1 Surgeries Prior to Listing: Norwood Stage I: BT Shunt | | | |
| No | 4.62 (4.42, 4.81) | 1 (Reference) | 0 (Reference) |
| Yes | 7.51 (7.30, 7.73) | 1.62 (1.59, 1.68) | 2.89 (2.85, 3.04) |
| F1T Recipient Weight (pounds) at Transplant | | | |
| 25th percentile: 16.7 | 5.60 (5.42, 5.83) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 35.3 | 5.27 (5.00, 5.45) | 0.94 (0.92, 0.96) | -0.33 (-0.43, -0.24) |
| 75th percentile: 94.1 | 5.22 (4.99, 5.39) | 0.93 (0.91, 0.96) | -0.38 (-0.49, -0.24) |
| F1T BSA (m2) at Transplant | | | |
| 25th percentile: 0.38 | 5.66 (5.40, 5.96) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 0.68 | 5.32 (5.08, 5.57) | 0.94 (0.92, 0.95) | -0.34 (-0.43, -0.27) |
| 75th percentile: 1.33 | 5.23 (4.96, 5.38) | 0.92 (0.90, 0.95) | -0.44 (-0.57, -0.30) |
| F1T Transplant Lab: BUN mg/dL | | | |
| 25th percentile: 12.0 | 4.31 (4.14, 4.48) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 16.0 | 4.74 (4.52, 4.97) | 1.10 (1.08, 1.12) | 0.43 (0.34, 0.50) |
| 75th percentile: 22.0 | 5.35 (5.07, 5.54) | 1.24 (1.22, 1.26) | 1.04 (0.96, 1.11) |
| F1 Recipient Age (years) at Listing | | | |
| 25th percentile: 0.50 | 5.82 (5.63, 6.09) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 4.31 | 5.40 (5.12, 5.61) | 0.93 (0.90, 0.94) | -0.42 (-0.57, -0.37) |
| 75th percentile: 12.8 | 5.01 (4.82, 5.26) | 0.86 (0.84, 0.88) | -0.81 (-0.96, -0.70) |
| F1 BSA (m2) at Listing | | | |
| 25th percentile: 0.34 | 5.27 (5.08, 5.41) | 1 (Reference) | 0 (Reference) |
| 50th percentile: 0.64 | 5.14 (5.01, 5.40) | 0.98 (0.97, 0.99) | -0.13 (-0.18, -0.03) |
| 75th percentile: 1.30 | 5.55 (5.36, 5.76) | 1.05 (1.04, 1.07) | 0.27 (0.21, 0.38) |

**Figure 1:** Model development algorithm considered in the current analysis.



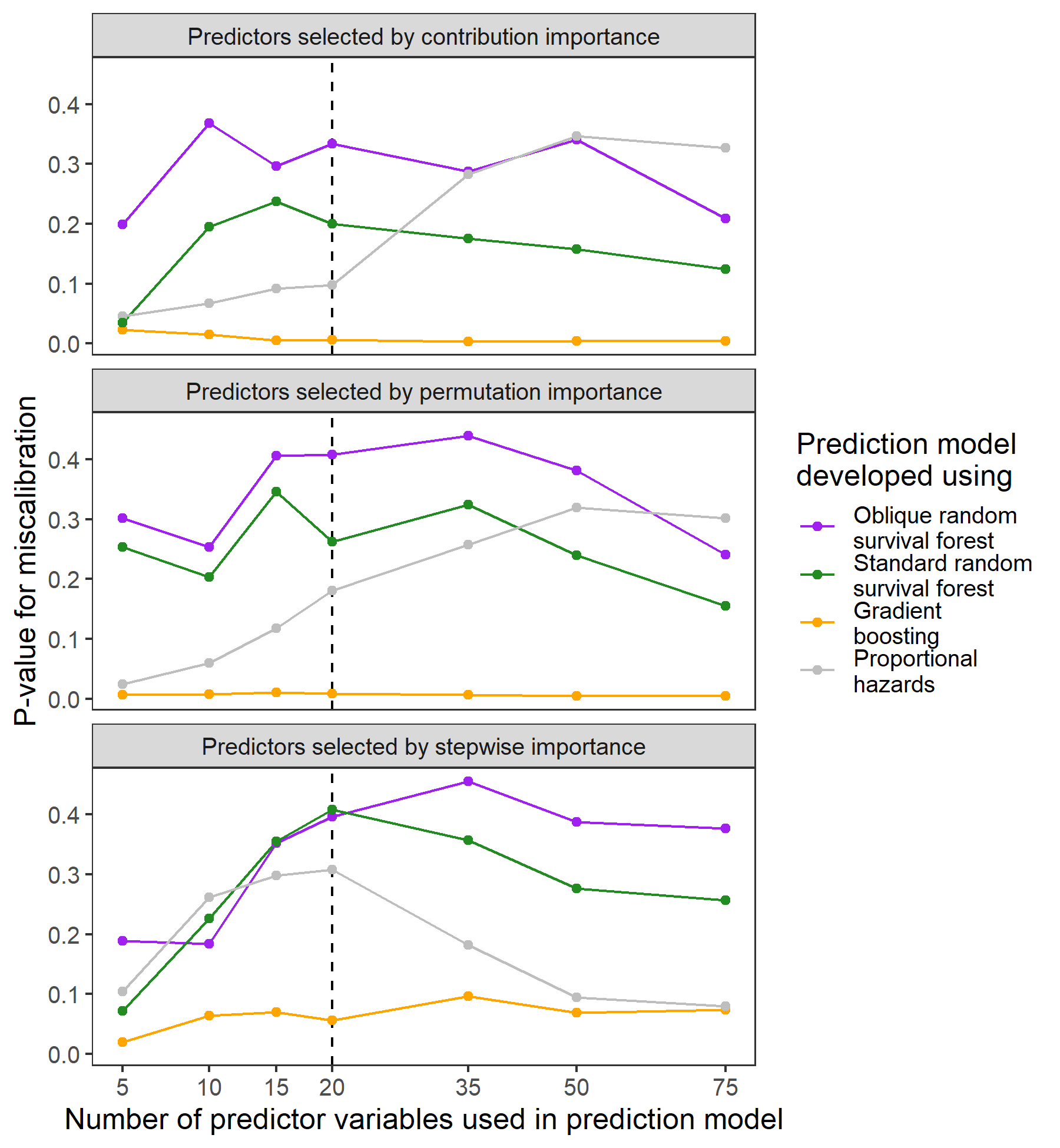
**Figure 2:** Internally validated estimates of model discrimination as a function of the number of predictor variables included in the prediction model.



Each point in the figure is the median value of model performance aggregated across 500 replicates of Monte-Carlo cross validation.

Discrimination was assessed at 1 year after transplant using **FILL IN**.

**Figure 3:** Internally validated estimates of model calibration as a function of the number of predictor variables included in the prediction model.

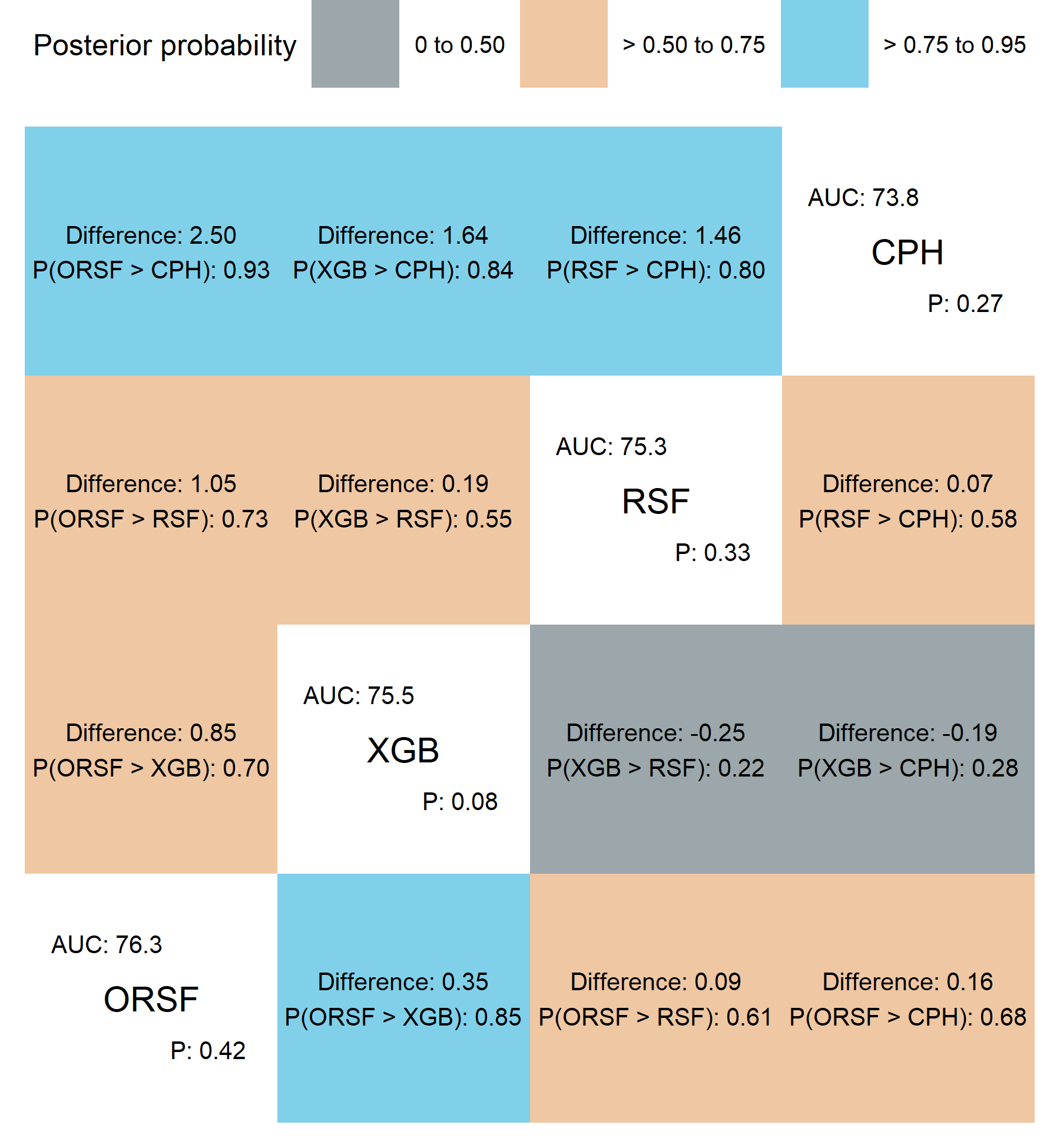


Model performance was defined using the p-value from a statistical test for miscalibration (REF).

Each point in the figure is the median value of model performance aggregated across 500 replicates of Monte-Carlo cross validation.

Calibration was assessed at 1 year after transplant using **FILL IN**.

**Figure 4:** Bayesian estimates of differences in model performance when 20 variables were selected using permutation importance.



# REFERENCES

1. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of statistics 2001:1189–1232.

2. Breiman L. Random forests. Machine learning 2001;45:5–32.

3. Zhang Z. Variable selection with stepwise and best subset approaches. Annals of translational medicine 2016;4.

4. Freedman L, Pee D, Midthune D. The problem of underestimating the residual error variance in forward stepwise regression. Journal of the Royal Statistical Society: Series D (The Statistician) 1992;41:405–412.

5. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

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

7. Breiman L. Bagging predictors. Machine learning 1996;24:123–140.

8. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS, others. Random survival forests. The annals of applied statistics 2008;2:841–860.

9. Jaeger BC, Long DL, Long DM, et al. Oblique random survival forests. The Annals of Applied Statistics 2019;13:1847–1883.

10. Xu Q-S, Liang Y-Z. Monte carlo cross validation. Chemometrics and Intelligent Laboratory Systems 2001;56:1–11.

11. Steyerberg EW, Harrell Jr FE, Borsboom GJ, Eijkemans M, Vergouwe Y, Habbema JDF. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. Journal of clinical epidemiology 2001;54:774–781.

12. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.) 2010;21:128.

13. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. European heart journal 2014;35:1925–1931.

14. Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Statistics in medicine 2013;32:5381–5397.

15. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. Statistics in medicine 2015;34:1659–1680.

16. Pencina MJ, D’Agostino Sr RB, D’Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in medicine 2008;27:157–172.

17. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Springer; 2013.

18. Josse J, Prost N, Scornet E, Varoquaux G. On the consistency of supervised learning with missing values. arXiv preprint arXiv:1902.06931 2019.

19. Benavoli A, Corani G, Demšar J, Zaffalon M. Time for a change: A tutorial for comparing multiple classifiers through bayesian analysis. The Journal of Machine Learning Research 2017;18:2653–2688.

20. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. Annals of internal medicine 2015;162:W1–W73.

21. Molnar C. Interpretable machine learning. Lulu. com; 2020.

22. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available at: <https://www.R-project.org/>.

23. Jaeger B. Table.glue: Make and apply customized rounding specifications for tables. Available at: <https://github.com/bcjaeger/table.glue>.

24. Landau WM. The drake R package: A pipeline toolkit for reproducibility and high-performance computing. Journal of Open Source Software 2018;3. Available at: <https://doi.org/10.21105/joss.00550>.

25. Kuhn M, Wickham H. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.; 2020. Available at: <https://www.tidymodels.org>.

26. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. Journal of Open Source Software 2019;4:1686.