Received: 30 August 2018 | Revised: 19 December 2018 | Accepted: 4 January 2019

DOI: 10.1111/petr.13360

ORIGINAL ARTICLE

WILEY

Prediction of mortality following pediatric heart transplant using machine learning algorithms

Rebecca Miller[1] | Dmitry Tumin[2] | Jennifer Cooper[3,4] | Don Hayes Jr[5,6] | Joseph D. Tobias[1,7]

[1]Department of Anesthesiology and Pain Medicine, Nationwide Children's Hospital, Columbus, Ohio

[2]Department of Pediatrics, Brody School of Medicine, East Carolina University, Greenville, North Carolina

[3]The Research Institute, Nationwide Children's Hospital, Columbus, Ohio

[4]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio

[5]Section of Pulmonary Medicine, Nationwide Children's Hospital, Columbus, Ohio

[6]Department of Pulmonary and Critical Care Medicine, The Ohio State University College of Medicine, Columbus, Ohio

[7]Department of Anesthesiology and Pain Medicine, The Ohio State University College of Medicine, Columbus, Ohio

Correspondence
Rebecca Miller, Department of Anesthesiology and Pain Medicine, Nationwide Children's Hospital, Columbus, OH.
Email: Rebecca.Miller@nationwidechildrens.org

Abstract

Background: Optimizing transplant candidates' priority for donor organs depends on the accurate assessment of post-transplant outcomes. Due to the complexity of transplantation and the wide range of possible serious complications, recipient outcomes are difficult to predict accurately using conventional multivariable regression. Therefore, we evaluated the utility of 3 ML algorithms for predicting mortality after pediatric HTx.

Methods: We identified patients <18 years of age receiving HTx in 2006-2015 in the UNOS Registry database. Mortality within 1, 3, or 5 years was predicted using classification and regression trees, RFs, and ANN. Each model was trained using cross-validation, then validated in a separate testing set. Model performance was primarily evaluated by the area under the receiver operating characteristic (AUC) curve.

Results: The training set included 2802 patients, whereas 700 were included in the testing set. RF achieved the best fit to the training data with AUCs of 0.74, 0.68, and 0.64 for 1-, 3-, and 5-year mortality, respectively, and performed best in the testing data, with AUCs of 0.72, 0.61, and 0.60, respectively. Nevertheless, sensitivity was poor across models (training: 0.22-0.58; testing: 0.07-0.49).

Discussion: ML algorithms demonstrated fair predictive utility in both training and testing data, but the sensitivity of these algorithms was generally poor. With the registry missing data on many determinants of long-term survival, the ability of ML methods to predict mortality after pediatric HTx may be fundamentally limited.

KEYWORDS
heart transplantation, machine learning, united network for organ sharing, UNOS

1 | INTRODUCTION

Because of the limited supply of donor organs for transplantation, decisions regarding transplant candidacy and donor organ allocation are influenced by expectation of post-transplant survival.[1] Maximizing utility of transplantation and optimizing donor-recipient matching depend on accurate assessment of recipients' post-transplant mortality risk. In the US, accurate prediction of post-transplant survival is also important for fair evaluation of transplant center performance. Specifically, transplant centers can

Abbreviations: ANN, artificial neural networks; AUC, area under the receiver operating characteristic curve; BMI, body mass index; CART, classification trees; CHD, congenital heart disease; CM, cardiomyopathy; CMV, cytomegalovirus; ECMO, extracorporeal membrane oxygenation; FiO2, fraction of inspired oxygen; HTx, heart transplant; ICU, intensive care unit; IMPACT, index for mortality prediction after cardiac transplantation; LVAD, left ventricular assist device; ML, machine learning; pCO2, partial pressure of carbon dioxide; pO2, partial pressure of oxygen; RF, random forest; SGOT, serum glutamic oxaloacetic transaminase; SGPT, serum glutamic pyruvic transaminase; SRTR, scientific registry of transplant recipient; UNOS, united network for organ sharing; US, United States.

be penalized by the Centers for Medicare & Medicaid Services if their performance does not meet expectations based on risk-adjusted outcomes.[2,3] Although these issues have placed statistical prediction of recipient survival at the forefront of the debate over donor organ allocation, the complexity of organ transplantation and the wide range of possible serious complications mean that recipient outcomes are difficult to predict accurately. This problem is especially pronounced in pediatric transplantation, where analysis of small cohorts can introduce error into predictive models. In the case of HTx, existing clinical risk scores exhibit AUC (AUC, a global measure of model fit) of 0.48-0.77, indicating predictive values that range from acceptable (AUC > 0.7) to no better than chance (AUC = 0.5).[4-7] In studies of pediatric HTx recipients, reported AUCs of multivariable regression models have ranged between 0.67 and 0.78, similarly demonstrating limited accuracy in predicting post-transplant outcomes.[8,9] In effect, published models are only able to discriminate between low- and high-risk patients 67%-78% of the time, indicating that survival cannot be reliably predicted.

Traditionally, risk scores for post-transplant outcomes, such as the IMPACT, have been generated using multivariable logistic or Cox proportional hazards regression.[7] The SRTR, which reports transplant outcomes in the US, uses models that describe graft and patient survival using Cox proportional hazards regression.[10] These models are specific to age-group (pediatric vs. adult) and organ type, and include a selection of recipient and donor variables from the UNOS registry that were found to produce the best-fitting model.[10] The models are refit in each Program-Specific Report cycle to capture changing predictors of transplant outcomes.[10]

In contrast to regression-based approaches, ML is rapidly emerging as a valuable tool for predicting surgical outcomes. For example, ML algorithms have been reported to significantly enhance prediction of outcomes following neurosurgery, when compared to logistic regression.[11] ML has also been proposed to improve prediction of transplant outcomes. Recently, several studies have attempted to improve HTx outcome prediction using ML techniques in adults or combined pediatric and adult populations.[12-18] Compared to regression-based modeling approaches, ML algorithms can capture more complex interactions between characteristics, which may result in improved predictions of transplantation outcomes. A variety of ML techniques, including ANN, classification and regression trees (CART), RF, support vector machines, and naïve Bayes classifiers, have been used to predict outcomes of organ transplantation.

Despite the conceptual appeal of ML algorithms, prior studies have demonstrated variable predictive utility of ML approaches in this setting (AUC: 0.54-0.84).[12,13,15,16,18] Understanding the performance of ML algorithms for predicting pediatric HTx outcomes could inform future decisions as to whether ML-based algorithms should be used in candidate selection and allocation of donor organs. Therefore, we compared the performance of 3 ML algorithms for predicting all-cause mortality after pediatric HTx. We hypothesized that all ML methods would demonstrate high predictive utility for mortality after pediatric HTx.

2 | METHODS

This study was considered exempt from review by the Institutional Review Board at Nationwide Children's Hospital due to the deidentified nature of the UNOS Registry database. We identified patients aged <18 years in the UNOS Registry database who underwent first-time HTx between the years 2006 and 2015, excluding patients who underwent concurrent lung transplantation. The primary outcome was all-cause mortality within 1 year (for transplants performed in 2006-2015), with secondary outcomes including mortality within 3 or 5 years (for transplants performed in 2006-2013 or 2006-2011, respectively). Patients were eligible for analysis of a given outcome if their mortality status was known at that time point. Vital status was ascertained by transplant centers and mandatorily reported to UNOS. Patients were excluded from analysis if they were re-transplanted or lost to follow-up before a given time point.

We tested three ML algorithms that have been recently used in HTx outcomes research: ANN, CART, and RF.[12,14-16,18] ANNs are designed to mimic biological neural processing and consist of weighted connections between neurons. As inputs flow through the system, neurons respond by firing at certain thresholds, producing the final output. In CARTs, a type of decision tree, a sample is divided into branches based on input characteristics until a final output is reached. RFs are an ensemble classifier composed of a collection of independent decision trees.[19] Variables for each algorithm were selected from recipient and donor data available at transplantation. In line with prior studies, variables with >10% missing values were excluded from consideration,[17,20] and data were split randomly as 80% training and 20% testing (stratified on patient mortality at last known follow-up, to maintain approximately even mortality rates between training and testing sets).[21] Using the training data, categorical measures were divided into binary variables to retain clinically relevant distinctions between categories, while eliminating potentially inaccurate ordinal relationships between categories. For highly collinear variable pairs, the variable having the largest mean absolute correlation with other variables in the training set was excluded from the set.[22] Variables with near-zero variance were also removed from the training set. Missing values were imputed separately in training and testing data using single imputation by predictive mean matching. For ANNs, continuous variables were normalized to have a mean of 0 and a variance of one within the training data for a given outcome, so that all predictors would initially be given equal importance. Testing data for each outcome were then normalized using mean and standard deviation estimates from the training data set. To reduce the risk of overprediction of survival, deaths in the training data were synthetically oversampled as previously described.[12,23]

We followed a twofold approach for variable selection, with the intent to include both relevant variables that risk scores may not traditionally incorporate, as well as variables known to be clinically significant. Each model was trained with an initial feature set that included all available covariates, using 10-fold cross-validation to select optimal tuning parameters. For each ML model, up to the 15 most important variables for outcome prediction were retained for inclusion in the final feature set.[24] For CART, importance was defined by how much a variable's inclusion improved model accuracy. For RF, importance was defined similarly, but averaged across the individual trees in the model and normalized by the standard error. For ANN, importance was evaluated by calculating the AUC when each variable was used as the only predictor.[22]

To ensure the inclusion of clinically significant variables, this final feature set also included recipient variables identified a priori as important predictors of mortality following HTx based on literature review: gender, age at transplant, race/ethnicity (white, black, Hispanic, other), weight at transplant, ABO blood type, diagnosis category (congenital heart disease, CM, other), payor type at transplant (private insurance, public insurance, other), most recent creatinine level prior to transplant, total days on waitlist, medical condition prior to transplantation (in ICU, hospitalized but not in ICU, not hospitalized), ending waitlist status (1A vs. others), implantable defibrillator at listing, LVAD at transplant, ECMO at transplant, mechanical ventilation at transplant, and use of inotropic agents at transplant.[13,16-18,25,26] Donor characteristics identified on literature review included gender, age, weight, ABO blood type, and recipient blood match level (identical, compatible, incompatible).[13,17] Other available characteristics were included if they met the criteria described above during model training.

Models were trained on the final feature set using 10-fold cross-validation to select optimal tuning parameters. For comparability to previous studies that used cross-fold validation, model performance on the training data was evaluated by the mean AUC for the 10 folds. The final trained model was then validated on the separate testing set, and testing performance was evaluated by the AUC. The DeLong test was used to compare AUC (representing model performance) between training and testing data.[27] To further assess model performance, we calculated sensitivity and specificity, where sensitivity describes the proportion of deceased patients that the model correctly classifies as deceased, and specificity describes the proportion of surviving patients that the model correctly classifies as surviving. To characterize model calibration and fit, we also calculated the calibration slope and intercept. Calibration slopes differing from one suggest model overfitting to the training data set, while calibration intercepts differing from 0 suggest systematic bias toward under- or overpredicting the risk of mortality.[28] Analysis was performed using Stata/IC 14.2 (College Station, TX: StataCorp, LP) and R version 3.4.3 (R Foundation for Statistical Computing, Vienna, Austria) with packages pROC, caret, cvAUC, DMwR, MICE, and rms.[27,29-33]

3 | RESULTS

We initially included 2802 patients in the training set and 700 patients in the testing set. Exclusions due to retransplant and loss to follow-up are summarized for each end-point in Table 1. Among patients retained for analysis, the overall mortality rate was 9% at 1 year, 15% at 3 years, and 23% at 5 years. The most important features identified when models were trained on all covariates are listed in Table 2. Primary diagnoses of CHD or CM, the use of mechanical ventilation at transplant, and donor B1 antigen levels were important for all RF and CART models. ECMO support at transplant was an important feature for all 1-year models but for none of the 5-year models. Similarly, the recipient's pre-transplant serum bilirubin level was important for all 1-year models, most 3-year models, and only the RF model of 5-year mortality. Recipient gender was an important feature in all 5-year models, while donor gender was important for most models of 1- and 3-year outcomes. The most important features identified in the final models, ranked by descending importance, are listed in Table S1.

Performances of the various models in predicting survival (AUC, sensitivity, and specificity) are summarized for training and testing data in Table 3. RF achieved the best fit to the training data with AUCs of 0.74, 0.68, and 0.64 for 1-, 3-, and 5-year mortality, respectively. RF also achieved the best overall performance on the testing data, with AUCs of 0.72, 0.61, and 0.60, for 1-year, 3-year, and 5-year mortality. However, sensitivity was poor in all models. ANNs had the highest overall sensitivity (0.55-0.58) but the lowest specificity (0.63-0.75) on the training data. RFs had the highest overall specificity (0.82-0.94) but the poorest sensitivity (0.22-0.34). In the testing data, sensitivity declined for RFs and ANNs, but improved slightly for CARTs. RF had particularly poor sensitivity (0.07-0.34), while CARTs achieved the highest sensitivity (0.44-0.49).

Model calibration is described by calibration slope and intercept in Table 4. The calibration line shows the linear relationship between the actual mortality rate and the predicted mortality risk, where an

TABLE 1 Number of pediatric HTx recipients retained for analysis, according to study outcome

		Pediatric HTx recipients (N)			
Mortality	Data set	Transplanted during time period	Excluded due to reported survival with last known follow-up preceding time point	Excluded due to retransplant	Final population
1-year (2006-2015)	Training	2802	244	13	2545
	Testing	700	60	5	635
3-year (2006-2013)	Training	2160	270	34	1856
	Testing	525	55	11	459
5-year (2006-2011)	Training	1557	227	45	1285
	Testing	399	67	12	320

TABLE 2 Variables added to prediction models with each method, ranked by descending importance

1-year mortality			3-year mortality			5-year mortality		
RF	ANN	CART	RF	ANN	CART	RF	ANN	CART
Primary diagnosis of CHD[a]	ECMO at transplant[a]	Primary diagnosis of CHD[a]	Primary diagnosis of CHD[a]	Recipient weight at transplant[a]	Donor B1 antigen levels	Primary diagnosis of CHD[a]	Recipient gender[a]	Primary diagnosis of CHD[a]
Primary diagnosis of CM[a]	Recipient BMI at transplant	Mechanical ventilation at transplant[a]	Primary diagnosis of CM[a]	Donor blood pH	Primary diagnosis of CHD[a]	Donor CMV status	Nationally allocated organ	Donor B1 antigen levels
ECMO at transplant[a]	ECMO at listing	ECMO at transplant[a]	Most recent serum total bilirubin prior to transplant	Donor gender[a]	Mechanical ventilation at transplant[a]	Donor B1 antigen levels	Ending waitlist status 1B	Mechanical ventilation at transplant[a]
Mechanical ventilation at transplant[a]	Mechanical ventilation at transplant[a]	Donor gender[a]	Donor pO2	ECMO at listing	Donor gender[a]	Most recent serum total bilirubin prior to transplant	Initial waitlist status 1B	Number of transfusions donor received during terminal hospitalization
Donor gender[a]	Nationally allocated organ	Recipient in ICU prior to transplant[a]	Donor pO2 on FiO2	Donor SGOT	Primary diagnosis of CM[a]	Donor SGOT	Recipient Epstein-Barr virus status	Identical ABO match[a]
Donor B1 antigen levels	Recipient in ICU prior to transplant[a]	Primary diagnosis of CM[a]	Donor B1 antigen levels	Donor A2 antigen levels	Donor A1 antigen levels	Mechanical ventilation at transplant[a]	Primary diagnosis of CHD[a]	Donor pCO2
Donor height	Donor blood type AB[a]	Donor pO2 on FiO2	Donor A1 antigen levels	Donor blood type O	Donor CMV status	Primary diagnosis of CM[a]	Donor A1 antigen levels	Recipient gender[a]
Most recent serum total bilirubin prior to transplant	Death mechanism donor: natural causes	Donor B1 antigen levels	Donor height	Death mechanism donor: gunshot wound	Recipient black race	Recipient gender[a]	Donor height	Public insurance at transplant[a]
Donor A1 antigen levels	Recipient not hospitalized prior to transplant[a]	Donor height	Number of transfusions donor received during terminal hospitalization	Death mechanism donor: drug intoxication	ECMO at transplant	Recipient black race[a]	Donor Hepatitis B surface antigen	Primary diagnosis of CM[a]
Donor pO2 on FiO2	Donor Hepatitis B surface antigen	Recipient not hospitalized prior to transplant[a]	Donor creatinine levels	Life support at listing	Donor pO2 on FiO2	Donor pCO2	Donor pO2	Donor SGOT
Donor pO2	Primary diagnosis of CHD[a]	Most recent serum total bilirubin prior to transplant	Most recent creatinine levels prior to transplant[a]	Mechanical ventilation at listing	Recipient gender[a]	Number of transfusions donor received during terminal hospitalization	Death mechanism of donor: Sudden Infant Death Syndrome	Recipient black race[a]

(*Continues*)

ideal model would have a slope of 1 and an intercept of 0. RFs had the best overall calibration on both training and testing data. On the training data, RFs had slopes of 1.07, 0.94, and 0.88 and intercepts of −1.21, −1.04, and −0.87 for 1-year, 3-year, and 5-year outcomes, respectively. On the testing data, RFs had slopes of 1.25, 0.60, and 0.86 and intercepts of −0.92, −1.38, and −1.20 for 1-year, 3-year, and 5-year outcomes, respectively.

4 | DISCUSSION

Existing HTx risk scores demonstrate limited accuracy for the prediction of post-transplant outcomes. This limitation has contributed to skepticism about the use of a continuous risk scoring system for donor heart allocation.[34] Recently, a number of studies have attempted to improve the prediction of HTx outcomes using ML algorithms. However, these studies have demonstrated variable performance of ML algorithms and rarely focused on a pediatric population, where the limitations of conventional multivariable regression are more acute. To address this, we assessed the performance of 3 ML algorithms specifically in children undergoing HTx. All ML algorithms demonstrated fair predictive utility for 1-year mortality, with RF achieving the best performance. However, sensitivity was consistently poor for all algorithms.

All ML models performed best when predicting 1-year outcomes, with predictive utility declining for later outcomes. There are two possible explanations for this decline in performance. First, recipient death at later time points is more likely to be caused by factors not adequately captured in the UNOS Registry. Second, analysis of long-term outcomes included a smaller population of recipients, which may have limited the identification of common risk factors in a diverse patient cohort. Across all three time points, RF was the best-performing ML model and demonstrated fair predictive utility (based on AUC) in both the training and testing sets. However, this algorithm had especially poor sensitivity for predicting mortality after HTx. Indeed, all algorithms evaluated in the study had unacceptably low sensitivity, suggesting inherent limitations to predicting survival after pediatric HTx using ML methods.

Poor sensitivity is often exacerbated by data sets with rare outcomes, as ML algorithms tend to be biased toward the more common outcome, in this case, survival.[35] To reduce this bias, data sets can be balanced using oversampling of the rare outcome, undersampling of the common outcome, or a combination of the two.[12] Synthetic oversampling, as we used in this study, selects deceased patients and generates new examples based on deceased patients with similar characteristics.[23] However, this approach may not capture the varied causes of death in a small but heterogeneous population. Graft failure is the most common cause of death following pediatric HTx, while other common causes include cardiovascular and cerebrovascular disease, infection, and respiratory disease.[36] Variation in causes of death may be difficult to adequately describe with a ML model, especially when focusing on the limited population of pediatric HTx recipients. Furthermore, the UNOS Registry offers

TABLE 2 (Continued)

1-year mortality			3-year mortality			5-year mortality		
RF	ANN	CART	RF	ANN	CART	RF	ANN	CART
Total serum albumin at listing	Donor height	Donor weight[a]	Donor weight[a]	Number of transfusions donor received during terminal hospitalization	Most recent serum total bilirubin prior to transplant	Donor pO2	Recipient Hispanic ethnicity[a]	Recipient weight at transplant[a]
Recipient in ICU prior to transplant[a]	Most recent serum total bilirubin prior to transplant	Recipient age[a]	Recipient BMI at transplant	Most recent creatinine levels prior to transplant[a]	Identical ABO match[a]	Recipient BMI at listing	Death mechanism of donor: cardiovascular	Recipient BMI at listing
Days on waitlist[a]	Recipient blood type B[a]	Orthotopic traditional procedure	Mechanical ventilation at transplant[a]	Donor blood urea nitrogen	Donor pO2	Donor SGPT	LVAD at transplant[a]	Recipient BMI at transplant
Most recent creatinine levels prior to transplant[a]	Regionally-allocated organ	Recipient blood type O[a]	Recipient weight at transplant[a]	Orthotopic traditional procedure	Recipient in ICU prior to transplant[a]	Recipient BMI at transplant	Donor age[a]	Private insurance at transplant[a]

ANN, artificial neural network; BMI, body mass index; CART, classification and regression tree; CHD, congenital heart disease; CM, cardiomyopathy; CMV, cytomegalovirus; ECMO, extracorporeal membrane oxygenation; FiO2, fraction of inspired oxygen; RF, random forest; ICU, intensive care unit; LVAD, left ventricular assist device; pCO2, partial pressure of carbon dioxide; pO2, partial pressure of oxygen; RF, random forest; SGOT, serum glutamic oxaloacetic transaminase; SGPT, serum glutamic pyruvic transaminase.

[a] Variable was also included based on literature search.

TABLE 3 Model performance for prediction of 1-year, 3-year, and 5-year all-cause mortality following pediatric HTx

		Sensitivity		Specificity		AUC		
Mortality[a]	Model	Training[b]	Testing	Training[b]	Testing	Training[b]	Testing	*P*[c]
1-y	RF	0.22	0.07	0.94	0.98	0.74	0.72	0.549
Training N = 2545	ANN	0.58	0.23	0.75	0.94	0.73	0.65	0.075
Testing N = 635	CART	0.39	0.44	0.83	0.81	0.68	0.67	0.967
3-y	RF	0.25	0.13	0.91	0.94	0.68	0.61	0.092
Training N = 1856	ANN	0.55	0.44	0.69	0.69	0.67	0.57	0.038
Testing N = 459	CART	0.41	0.49	0.75	0.67	0.61	0.61	0.998
5-y	RF	0.34	0.34	0.82	0.80	0.64	0.60	0.407
Training N = 1285	ANN	0.55	0.48	0.63	0.63	0.63	0.54	0.058
Testing N = 320	CART	0.43	0.48	0.67	0.64	0.58	0.58	0.904

ANN, artificial neural network; CART, classification and regression tree.

[a] Censored if transplant performed too late for survival to be reported at time point or if last known follow-up preceded time point for living patients.

[b] Reported values are the mean for the 10 cross-validated folds in the final model.

[c] Comparison between training and testing AUCs.

TABLE 4 Model calibration for prediction of 1-year, 3-year, and 5-year all-cause mortality following pediatric HTx

		Intercept		Slope	
Mortality[a]	Model	Training	Testing	Training	Testing
1 y	RF	−1.21	−0.92	1.07	1.25
Training N = 2545	ANN	−2.07	−1.21	0.60	0.73
Testing N = 635	CART	−1.95	−1.98	0.44	0.46
3 y	RF	−1.04	−1.38	0.94	0.60
Training N = 1856	ANN	−1.55	−1.73	0.43	0.26
Testing N = 459	CART	−1.54	−1.73	0.39	0.38
5 y	RF	−0.87	−1.20	0.88	0.86
Training N = 1285	ANN	−1.12	−1.43	0.50	0.20
Testing N = 320	CART	−1.09	−1.38	0.28	0.33

ANN, artificial neural network; CART, classification and regression tree.

[a] Censored if transplant performed too late for survival to be reported at time point or if last known follow-up preceded time point for living patients.

limited information on surgical history, comorbidities in other organ systems, social determinants of health, and other factors contributing to post-transplant patient survival. Lack of information on these characteristics may fundamentally limit predictive utility of ML algorithms trained on the registry data alone. A complex set of factors affect patient outcomes, but models can only make predictions using the data available. Regardless of model sophistication, if a database does not adequately capture factors affecting patient outcomes, the model cannot capture these relationships. Additionally, although models such as ANNs can describe complex relationships among variables, increasingly complex models generally require a larger amount of data to produce accurate predictions. Capturing these relationships may be a challenge in the relatively small pediatric HTx population.

Previous studies applying ML algorithms to HTx outcomes in adults or combined pediatric and adult populations have reported AUCs for short-term mortality comparable to the results of our analysis. Studies examining 1-year outcomes have reported AUCs of 0.59-0.66 for the best-performing models, while studies examining 5-year outcomes have reported AUCs of 0.60-0.67.[12,15,16,18] Studies examining longer-term outcomes have reported a wider range of AUCs, ranging from 0.54 to 0.84 for 9- or 10-year outcomes, although this time range was beyond the scope of our study.[12,13,15,18] While the majority of these studies also used UNOS Registry data, they frequently included a broader time range, beginning with transplants performed as early as 1985.[12-14,17,18] This may have biased results by not sufficiently accounting for changes in wait-listing criteria, evolution in mechanical circulatory support technology, or changes in post-transplant management during this time. Furthermore, these studies were either limited to adults or included both adults and children, restricting specific inference for the pediatric cohort.[12-18]

In contrast to our study, most previous studies reported performance based on a cross-validated data set, without reporting performance on a separate test set.[12-15,17,18] In *k*-fold cross-validation, the data set is split randomly into *k* parts, trained on all but one fold, and tested on the last fold. Cross-validation is useful for minimizing the biases that can result from dividing data into training and testing sets that contain relatively few cases of the event to be predicted.[37] However, validation with a separate test set, ideally an external data set, may be necessary to accurately assess the predictive utility of these models for new transplant recipients.[38] Although we could not validate our models with an external data set, we used a held-out test set for internal validation. The decline in performance of the ANNs when validated with the test set suggests these models may have overfit the training data, even with the use of cross-validation to select model parameters and the use of synthetic oversampling to reduce bias toward the more common outcome of survival.

This study has several limitations. First, random selection of patients into training and testing data sets does not correspond to prediction of outcomes in "future" cohorts. However, this approach may have reduced bias compared to division by year of transplant. Additionally, there were a large number of censored patients. The most common reason for censoring was loss to follow-up, while censoring due to retransplant was relatively rare. Binary classification is not robust to censoring like models designed to predict survival time, so censoring limited the patients that could be included in our study. However, these models were selected to be comparable to most previous studies. We also considered ML algorithms separately, to correspond to methods in recent literature. However, a super-learning approach, which combines multiple ML algorithms to create a single model, may have improved predictive performance.[39] Outside of our control, some limitations are inherent to the use of the UNOS Registry database. Because some variables in the registry had missing data, we excluded variables that were ≥10% missing and used single imputation for other missing values. Our decision to exclude variables with high missing data rates is similar to prior ML studies, but differs from the SRTR approach of treating missing values for certain variables as a separate category. Some UNOS variables are also not initially grouped in parsimonious categories for analysis. Our results are therefore sensitive to our chosen grouping of categorical variables.

As shown by the limited predictive utility and especially the low sensitivity of ML algorithms in our study, there are inherent difficulties with applying ML models in clinical settings. Interpretability of model output typically must be traded for (expected) improvement in predictive value. When examining the outputs of models such as RFs or ANNs, observers cannot determine the rationale that causes a model to assign a certain risk score. For physicians, this can limit the clinical usefulness of these predictions, especially due to the inability to clearly predict risk according to individual patients' unique circumstances. More broadly, despite the availability of increasingly complex analytic approaches for mining large databases, traditional analysis triangulating findings across multiple independently collected data sets can more definitively support a clinical consensus about relevant predictors of surgical outcomes.[40]

Our findings also suggest limitations inherent to evaluating center performance based on predictive modeling (eg, comparing observed to expected center-specific survival). Although center performance is currently evaluated using predictions derived from the UNOS registry, our results suggest that even complex ML approaches cannot accurately predict outcomes with the available data. This raises the concern that centers could be improperly penalized due to inaccurate predictions. For example, institutions with a high proportion of patients for whom survival is overpredicted by the model might be improperly penalized for adverse patient outcomes.

Although our evaluation of ML algorithms demonstrated fair predictive utility for 1-year HTx outcomes, this performance did not represent significant improvement over previously published risk scoring systems. Therefore, improved prediction of post-transplant mortality risk remains essential to credibly incorporate such predictive modeling into donor heart allocation algorithms.

ACKNOWLEDGMENTS

The data reported here have been supplied by the UNOS as the contractor for the Organ Procurement and Transplantation Network. The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government.

AUTHORS' CONTRIBUTIONS

All authors: Conceived and designed the study; RM and DH: Contributed to acquisition of the data; all authors: Contributed to analysis and interpretation of the data; RM: Drafted the manuscript; DT, JC, DH, and JDT: Critically revised the manuscript; all authors: Approved the final version to be published.

ORCID

Rebecca Miller https://orcid.org/0000-0001-9629-4848
Don Hayes https://orcid.org/0000-0002-6734-6052

REFERENCES

1. About transplantation. Organ Procurement and Transplantation Network. Available at: https://optn.transplant.hrsa.gov/learn/about-transplantation.
2. Howard RJ, Cornell DL, Schold JD. CMS oversight, OPOs and transplant centers and the law of unintended consequences. *Clin Transplant*. 2009;23:778-783.
3. VanWagner LB, Skaro AI. Program-specific reports: Implications and impact on program behavior. *Curr Opin Org Transplant*. 2013;18:210-215.
4. Nguyen LS, Coutance G, Ouldamar S, et al. Performance of existing risk scores around heart transplantation: validation study in a 4-year cohort. *Transpl Int*. 2018;5:520-530.
5. Schulze PC, Jiang J, Yang J, et al. Preoperative assessment of high-risk candidates to predict survival after heart transplantation. *Circ Heart Fail*. 2013;6:527-534.
6. Segovia J, Cosío MD, Barceló JM, et al. RADIAL: a novel primary graft failure risk score in heart transplantation. *J Heart Lung Transplant*. 2011;30:644-651.

7. Weiss ES, Allen JG, Arnaoutakis GJ, et al. Creation of a quantitative recipient risk index for mortality prediction after cardiac transplantation (IMPACT). *Ann Thorac Surg*. 2011;92:914-922.
8. Almond CS, Gauvreau K, Canter C, et al. A risk-prediction model for in-hospital mortality after heart transplantation in US children. *Am J Transplant*. 2012;12:1240-1248.
9. Davies RR, Russo MJ, Mital S, et al. Predicting survival among high-risk pediatric transplant recipients: an analysis of the United Network for Organ Sharing database. *J Thoracic Cardiovasc Surg*. 2008;135:147-155.
10. SRTR risk adjustment model documentation: Posttransplant outcomes. Scientific Registry of Transplant Recipients. Available at: https://www.srtr.org/reports-tools/risk-adjustment-models-posttransplant-outcomes.
11. Senders JT, Staples PC, Karhade AV, et al. Machine learning and neurosurgical outcome prediction: A systematic review. *World Neurosurg*. 2018;109:476-486.
12. Dag A, Oztekin A, Yucel A, et al. Predicting heart transplantation outcomes through data analytics. *Decis Supp Syst*. 2017;94:42-52.
13. Dag A, Topuz K, Oztekin A, et al. A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival. *Decis Support Syst*. 2016;86:1-12.
14. Delen D, Oztekin A, Kong ZJ. A machine learning-based approach to prognostic analysis of thoracic transplantations. *Artif Intell Med*. 2010;49:33-42.
15. Medved D, Nusques P, Nilsson J, et al. Selection of an optimal feature set to predict heart transplantation outcomes. *Conf Proc IEEE Eng Med Biol Soc*. 2016;2016:3290-3293.
16. Nilsson J, Ohlsson M, Höglund P, et al. The International Heart Transplant Survival Algorithm (IHTSA): A new model to improve organ sharing and survival. *PLoS ONE*. 2015;10:e0118644.
17. Oztekin A, Delen D, Kong ZJ. Predicting the graft survival for heart-lung transplantation patients: an integrated data mining methodology. *Int J Med Inform*. 2009;78:e84-e96.
18. Yoon J, Zame WR, Banerjee A, et al. Personalized survival predictions for cardiac transplantation via tree predictors. *PLoS ONE*. 2017; 13: arXiv.
19. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Berlin: Springer-Verlag; 2009.
20. Ayllón MD, Ciria R, Cruz-Ramírez M, et al. Validation of artificial neural networks as a methodology for donor-recipient matching for liver transplantation. *Liver Transpl*. 2017;24(2):192-203.
21. Lee CK, Hofer I, Gabel E, et al. Development and validation of a deep neural network model for prediction of postoperative in-hospital mortality. *Anesthesiology*. 2017;129(4):649-662.
22. Kuhn M. The caret package (2017) Available at: https://topepo.github.io/caret/.
23. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
24. Lau L, Kankanige Y, Rubinstein B, et al. Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation*. 2017;101:e125-e132.
25. Rossano JW, Dipchand AI, Edwards LB, et al. The Registry of the International Society for Heart and Lung Transplantation: Nineteenth pediatric heart transplantation report-2016; Focus Theme: Primary diagnostic indications for transplant. *J Heart Lung Transplant*. 2016;35:1185-1195.
26. Hong KN, Iribarne A, Worku B, et al. Who is the high-risk recipient? Predicting mortality after heart transplant using pretransplant donor and recipient risk factors. *Ann Thorac Surg*. 2011;92:520-527.
27. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
28. Fenlon C, O'Grady L, Doherty ML, et al. A discussion of calibration techniques for evaluating binary and categorical predictive models. *Prev Vet Med*. 2018;149:107-114.
29. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28:1-26.
30. LeDell E, Petersen M, van der Laan M. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron J Stat*. 2015;9:1583-1607.
31. Torgo L. (2013) Package DMwR. Comprehensive R archive network. Available at: http://cran r-project org/web/packages/DMwR/DMwR.pdf.
32. Van Buuren S, Groothuis-Oudshoon K. MICE: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45:1-67.
33. Harrell FE. (2018) rms: Regression Modeling Strategies, R package version 5.1-2. Available at: https://cran.r-project.org/web/packages/rms/.
34. Frigerio M. Optimal and Equitable Allocation of Donor Hearts: Which Principles Are We Translating Into Practices? *Transplant Direct*. 2017;3:e197.
35. Kotsiantis S, Kanellopoulos D, Pintelas P, et al. Handling imbalanced datasets: a review. *Comput Sci Eng*. 2006;30:25-36.
36. Colvin M, Smith JM, Hadley N, et al. OPTN/SRTR 2016 Annual data report: heart. *Am J Transplant Suppl*. 2018;1:291-362.
37. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. 1995; IJCAI: 1137–1145.
38. Siontis GC, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. 2015;68:25–34.
39. Cooper JN, Minneci PC, Deans KJ. Postoperative neonatal mortality prediction using superlearning. *J Surg Res*. 2018;221:311–319.
40. Karamlou T, Velez DA, Nigro JJ. Encrypted prediction: a hacker's perspective. *J Thorac Cardiovascular Surg*. 2017;154:2038–2040.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Miller R, Tumin D, Cooper J, Hayes D Jr, Tobias JD. Prediction of mortality following pediatric heart transplant using machine learning algorithms. *Pediatr Transplantation*. 2019;23:e13360. https://doi.org/10.1111/petr.13360