Brief Report

Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database

P. ELLIOTT MILLER, MD,[1] SUMEET PAWAR, MD,[1] BENJAMIN VACCARO, MD,[1] MEGAN MCCULLOUGH, MD,[1] POOJA RAO, MBBS, PhD,[3] ROHIT GHOSH, MSc,[3] PRASHANT WARIER, PhD,[3] NIHAR R. DESAI, MD, MPH,[1,2] AND TARIQ AHMAD, MD, MPH[1]

New Haven, Connecticut; and Mumbai, India

ABSTRACT

Background: Traditional statistical approaches to prediction of outcomes have drawbacks when applied to large clinical databases. It is hypothesized that machine learning methodologies might overcome these limitations by considering higher-dimensional and nonlinear relationships among patient variables.

Methods and Results: The Unified Network for Organ Sharing (UNOS) database was queried from 1987 to 2014 for adult patients undergoing cardiac transplantation. The dataset was divided into 3 time periods corresponding to major allocation adjustments and based on geographic regions. For our outcome of 1-year survival, we used the standard statistical methods logistic regression, ridge regression, and regressions with LASSO (least absolute shrinkage and selection operator) and compared them with the machine learning methodologies neural networks, naïve-Bayes, tree-augmented naïve-Bayes, support vector machines, random forest, and stochastic gradient boosting. Receiver operating characteristic curves and C-statistics were calculated for each model. C-Statistics were used for comparison of discriminatory capacity across models in the validation sample. After identifying 56,477 patients, the major univariate predictors of 1-year survival after heart transplantation were consistent with earlier reports and included age, renal function, body mass index, liver function tests, and hemodynamics. Advanced analytic models demonstrated similarly modest discrimination capabilities compared with traditional models (C-statistic $\leq$0.66, all). The neural network model demonstrated the highest C-statistic (0.66) but this was only slightly superior to the simple logistic regression, ridge regression, and regression with LASSO models (C-statistic = 0.65, all). Discrimination did not vary significantly across the 3 historically important time periods.

Conclusions: The use of advanced analytic algorithms did not improve prediction of 1-year survival from heart transplant compared with more traditional prediction models. The prognostic abilities of machine learning techniques may be limited by quality of the clinical dataset. (*J Cardiac Fail 2019;25:479–483*)

Key Words: Advanced analytics, heart transplantation, prediction algorithms.

A key expectation of the "big data" revolution in medicine is that advanced analytic methods will unearth nonlinear relationships and higher-dimensional associations between clinical variables, leading to significant improvements in the usefulness of information collected on patients.[1,2] The clinical potential of this approach has been aggressively promoted under the umbrella of personalized or precision medicine.[3] With widespread use of electronic health records (EHRs), it is anticipated that advanced analytics will allow for clinician decision making based on distilled interpretations of large amounts of patient data.[4] So far, however, there is no tangible evidence to suggest that

From the [1]Section of Cardiovascular Medicine, Yale School of Medicine, New Haven, Connecticut; [2]Center for Outcomes Research and Evaluation, New Haven, Connecticut and [3]Qure.ai, Mumbai, India.

Manuscript received July 31, 2018; revised manuscript received January 15, 2019; revised manuscript accepted January 23, 2019.

Reprint requests: Tariq Ahmad MD, MPH, Section of Cardiovascular Medicine, Center for Outcomes Research and Evaluation (CORE), Yale University School of Medicine, New Haven, CT 06520 Tel: 203-785-7191; Fax: 203-785-2917. E-mail: tariq.ahmad@yale.edu

See page 483 for disclosure information.

1071-9164/$ - see front matter

https://doi.org/10.1016/j.cardfail.2019.01.018

simply applying more complex algorithms to patient data can remove its shortcomings and provide novel clinically important information.[5] Furthermore, it appears that success in these efforts has largely involved their application to imaging data and very highly curated databases.[6] A reason for this might be that even the most robust of statistical methodologies can not circumvent systemic issues with data granularity and quality.[7,8] To explore this question, we examined whether a host of machine learning methodologies could outperform traditional statistical approaches for prediction of 1-year mortality in a large national database of patients undergoing cardiac transplantation.[9]

Methods

Study Sample

The United Network of Organ Sharing (UNOS) is a private, nonprofit organization that manages the United States organ transplant system under contract with the federal government and maintains a robust clinical database of patients waitlisted and who underwent solid organ transplantation.[10] We included patients in the database since its inception in 1987 through 2014, limiting our analysis to adult patients ($\geq$18 years old) undergoing their first single-organ heart transplantation. We excluded patients with >20% missing data, yielding 50,453/62,621 patients for our analyses. The Ethics Committee for Clinical Research at Yale University approved the study protocol. The data were anonymized and deidentified before analysis, and the Institutional Review Board waived the need for written informed consents from the participants; the data are publicly available on request, without patient or center identifiers.

Outcome and Variables

Our primary outcome of interest was 1-year survival. Consistent with previous analyses of UNOS predictors, variables included in our traditional methods included age (both donor and recipient), creatinine, body mass index, liver function tests, aspartate transaminase, and hemodynamics.

Statistical Analysis

We applied several traditional and advanced machine-learning methods to the data set for 2006−2014 as well as across UNOS regions and 3 historically important times corresponding to major UNOS allocation adjustments, which included inception to 1996 (n = 14,770), 1996 to 2006 (n = 18,587), and 2006 to 2014 (n = 17,096). These time periods were chosen to account for changes in national allocation algorithms, which changed the severity of illness of patients being transplanted. Before 1996, there were only 2 tiers (status 1 and status 2) of medical urgency. In 1996, the Organ Procurement and Transplant Organization expanded to 3 tiers (status 1A, 1B, and 2). In 2006, changes were made to the heart allocation sequence so that patients with a higher status could be offered an organ in another zone before lower-status patients in the same zone as the donor.[11]

The traditional methodologies tested were: logistic regression (LR), ridge regression, and regressions with LASSO (least absolute shrinkage and selection operator). The advanced analytical methodologies tested were: neural networks, naïve-Bayes, tree-augmented naïve-Bayes, support vector machines (SVMs), random forest, and stochastic gradient boosting (see detailed explanations below). Receiver operating characteristic (ROC) curves for sensitivity and specificity of each model were generated and a C-statistic calculated by calculating the area under the ROC curve (AUC) to estimate fit. To test C-statistic stability and validity of the model, serial bootstrapping was applied 10 times. Models were trained on 80% of the data and then tested on the remaining 20%. Calibration curves were generated for each model to measure the agreement between predicted and observed risk. Patients were grouped into deciles according to the model's predicted risk and plotted against observed risk with the use of the Hosmer and Lemeshow goodness of fit test. Calibration was further confirmed by calculating Brier scores for each model, which encapsulate the model's uncertainty, resolution, and reliability into one value by measuring the average gap (mean squared difference) or alignment between forecasted probabilities and actual outcomes. Analyses were performed with the use of R 3.2.2 (R Development Core Team, Vienna, Austria). Two−sided $P \leq .05$ was considered to be statistically significant for all analyses.

Rationale for Choice of Various Statistical Methods

Logistic Regression

LR is a relatively simple method that aims to model the survival probability with the multiple variables as linear a function with learnable weights. Like other regression-based techniques, LR iteratively learns the weights to minimize the cross-entropy loss for predictions. Because LR assumes a linear model, this prevents the iterative learning to get stuck at local minima. Because of its inherent assumption of linear relationship between input variables and output, LR fails to capture nonlinear relationships between output and input variables without explicitly including nonlinear or interaction terms.

Support Vector Machines

SVMs are another set of machine learning algorithm that aim to find the most separating hyperplane between the survival groups in the n-dimensional space of inputs. SVMs can have linear or nonlinear kernels that are used to measure distance across input data points. SVMs are convex and therefore guarantee global minima assumptions of convex optimization. A major drawback with SVMs is the manual kernel selection, thereby choosing the apt nonlinearity to model the internal representations of data points and the choice of loss functions that can be used for SVM. Finally, SVMs, in their default form, are not able to predict probabilistically, thus rendering themselves inapplicable for comparison metrics such as AUC.

Decision Trees

Decision trees aim to predict survival of patients by modeling outcome as a sequence of decisions based on input variables. Decision trees aim to segregate patients into distinct survival clusters conditioned on decision sequence. Interpretability in the form of visualizing the decision sequence is the most important advantage in decision trees. By modeling classifications as decision steps, decision trees can capture nonlinear dependencies. However, in case of high-dimensional input, such as for survival prediction, decision trees are extremely computationally expensive. Also, decision trees tend to be unstable, eg, partially duplicating inputs can lead to completely different decision tree.

Random Forest

Random forest is an ensemble of multiple decision trees. Random trees in ensembles can rectify the instability of results because there are multiple trees contributing to the results, thus making the predictions robust to data. However, the other problems of decision trees persist with the use of random forest models.

Neural Networks

Multi-layer perceptron (MLP) is the most basic form of neural networks. MLP can be viewed as a logistic regression classifier where the input is first transformed successively with the use of a sequence of learnt nonlinear transformations. This sequence of transformations projects the input data into a space where it becomes linearly separable. This virtually makes MLP a universal approximator, implying that any function can ideally be approximated by MLP if trained appropriately. The biggest downside of MLP is that, in its default form, it does not lend itself to interpretability, eg, to draw on the relative importance of features and other attributes. MLPs are more prone to overfitting because the number of parameters tend to be lot more than in a simple logistic regression. Also, training of such algorithms is generally computationally expensive and time consuming.

Results

As shown in Fig. 1, advanced analytical models demonstrated similar discrimination compared with traditional analytical models. The neural network model demonstrated the highest C-statistic (0.66) but this was only slightly superior to the simple LR, ridge regression, and regressions with LASSO models (C-statistic =0.65, all). All other advanced models produced inferior C-statistics, ranging from SVMs (0.52) to tree-augmented naïve-Bayes (0.62) and random forest (0.63). The performance of these prediction algorithms was similar to the most commonly cited risk score for mortality after heart transplantation—the IMPACT (Index for Mortality Prediction After Cardiac Transplantation) score, which has a C-statistic of 0.65.[9]

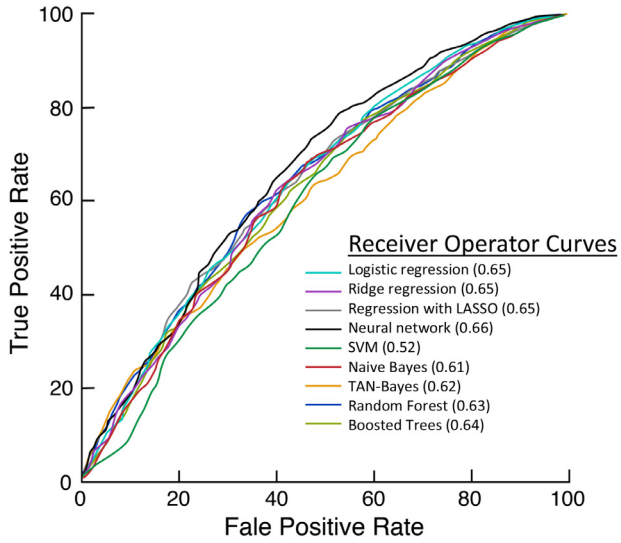

Fig. 1. Model discrimination according to traditional and novel analytical models. LASSO, least absolute shrinkage and selection operator; SVM, support vector machines; TAN, tree-augmented naïve.

We applied the neural network models to each UNOS region and across the 3 time historically important periods (Table 1). All regions across all 3 time periods demonstrated only modest discrimination. Cross-validation to assess the stability of the models demonstrated the tightest correlation with the neural network model (variation 0.004, C-statistic range 0.66−0.67). All other models demonstrated an acceptable level of correlation (all <0.02 variation across 10 validation tests) except for the SVMs model (variation 0.12, C-statistic range 0.47−0.59). Hosmer-Lemeshow testing depicted good calibration with the 3 traditional models as well as random forest and stochastic gradient boosting models. Inferior calibration was observed with the remaining advanced models. Quantitative assessment of calibration with the use of Brier scores confirmed these patterns (data not presented).

The major predictors of 1-year survival after heart transplantation are shown in Fig. 2, which also depicts the change in AUC as the number of variables in the model increases. Key predictors included donor and recipient age, bilirubin, creatinine clearance, hemodynamics, donor blood pH, and candidate diagnosis, closely reflecting previously published reports.

Discussion

We found that the prediction of 1-year outcomes after cardiac transplantation was similar between machine learning and traditional statistical methods in the central repository of patients undergoing heart transplantation in the United States. All of the models developed in this study showed similar and very modest discrimination, with C-statistics consistently ~0.65 regardless of their complexity. A traditional statistical approach consisting of multivariate logistic regression has been previously used on the UNOS

Table 1. C-Statistics for Neural Networks Across United Network of Organ Sharing Regions and Time Periods*

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6	Region 7	Region 8	Region 9	Region 10	Region 11	Average or Total
AUC												
Period 1	0.548	0.589	0.575	0.604	0.582	0.71	0.589	0.628	0.619	0.607	0.66	0.613
Period 2	0.702	0.671	0.639	0.68	0.658	0.698	0.647	0.512	0.715	0.62	0.663	0.659
Period 3	0.605	0.692	0.671	0.645	0.673	0.658	0.678	0.62	0.695	0.626	0.676	0.663
Patient numbers												
Period 1	611	1656	1726	1649	2398	462	1391	1085	723	1435	1634	14,770
Period 2	686	2392	2383	1897	2700	583	1779	933	1179	1893	2182	18,587
Period 3	704	2398	1999	1845	2626	474	1626	1045	1091	1402	1886	17,096

AUC, area under the receiver operating characteristic curve.
*Period 1: inception to 1996; period 2: 1996 to 2006; period 3: 2006 to 2014.

database for predicting 1-year mortality with a C-statistic of 0.65, consistent with our results. Specifically, our findings replicate the predictive capabilities of the IMPACT score.[9] Although machine learning has been touted as a path to unearthing nonlinear relationships and higher-dimensional associations between variables in medicine, it remains to live up to its expectations.[12] Our results raise the notion that large clinical datasets might lack the accuracy and granularity needed for machine learning methodologies to uncover unique associations.

Similar limitations in the application of machine learning methods to large datasets of patient information have been demonstrated for prediction of heart failure readmissions.[2,13] However, these methods have performed extremely well when applied to imaging information, as noted in recent reports involving head computerized tomography and echocardiography, with C-statistics >0.90.[8,14] We think that our results, when considered in conjunction with these previous findings, provide an insight into the inability of machine learning methods to overcome key systemic limitations of clinical datasets. In this case, whereas UNOS is a robust clinical registry, it is based on administrative data, which can blunt phenotyping of complex patients and attenuate the predictive ability of both traditional methods as well as machine learning techniques; indeed, this has been noted several times in efforts aimed at extracting meaningful information from the EHR.[15] Our analysis illustrates this limitation as we see the AUC for neural networks plateau after inclusion of ~25 variables in the model.

Several limitations of this conclusion must be considered. The patient journey after heart transplantation is very complex, and 1-year outcomes are likely to be determined by events after transplantation. It is very likely that detailed patient information after the surgery would significantly improve our ability to predict outcomes. However, pretransplantation prognostication is given an inordinate amount of attention during decision making for listing and transplanting patients.[16]

To our knowledge, the present study is the first to compare different predictive models in patients undergoing cardiac transplantation. We demonstrated that the neural networks model demonstrated the highest discrimination and the most reliable C-statistic when validated. However, its calibration was inferior to all of the traditional models. This is a common problem with advanced methodologies, wherein algorithms can become unstable due to multicollinear predictors or overfitting due to random correlations. Future work is required to understand the best use of specific statistical methodologies to apply to clinical datasets of different characteristics, and consensus is needed on how to validate the resultant findings.[17]

Conclusion

We found that machine learning performed similarly to traditional statistic methods for prediction of 1-year survival after cardiac transplantation. All statistical methodologies tested

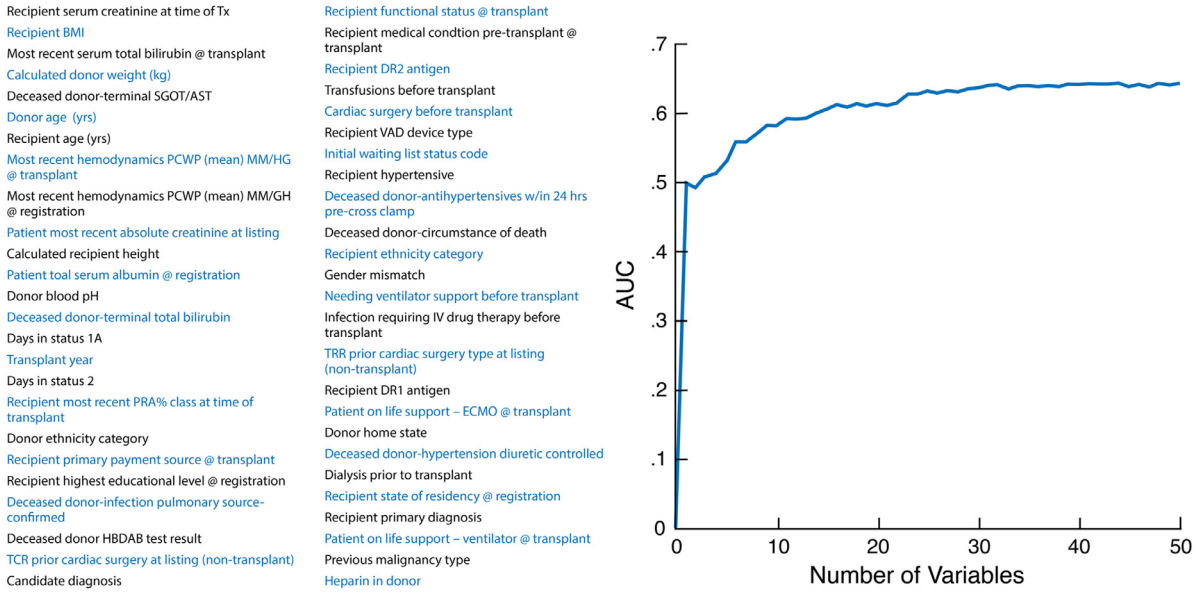

Fig. 2. Predictors of 1-year mortality after cardiac transplantation with the use of logistic regression.

offered only modest predictive power, with C-statistics consistently ≤0.66. Before widespread application of machine learning methodologies to clinical datasets, a focus on improving the quality of available data might be required.

Disclosures

Dr Desai reports being a recipient of a research agreement from Johnson & Johnson, through Yale University, to develop methods of clinical trial data sharing. Dr Ahmad is supported by grant K12 HS023000-04 from the Agency for Healthcare Research and Quality. All of the other authors report no potential conflicts of interest or financial relationships.

References

1. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33:1163–70.
2. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, Negahban SN, Krumholz HM. Analysis of machine learning techniques for heart failure readmissions. Circ Cardiovasc Qual Outcomes 2016;9:629–40.
3. Parikh RB, Schwartz JS, Navathe AS. Beyond genes and molecules—a precision delivery initiative for precision medicine. N Engl J Med 2017;376:1609–12.
4. Srinivas TR, Taber DJ, Su Z, Zhang J, Mour G, Northrup D, et al. Big data, predictive analytics, and quality improvement in kidney transplantation: a proof of concept. Am J Transplant 2017;17:671–81.
5. Joyner MJ, Paneth N, Ioannidis JP. What happens when underperforming big ideas in research become entrenched? JAMA 2016;316:1355–6.
6. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. J Am Heart Assoc 2018 Apr 12;7(8). pii: e008081.
7. Ahmad T, Testani JM, Desai NR. Can big data simplify the complexity of modern medicine?: prediction of right ventricular failure after left ventricular assist device support as a test case. JACC Heart Fail 2016;4:722–5.
8. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018.
9. Weiss ES, Allen JG, Arnaoutakis GJ, George TJ, Russell SD, Shah AS, Conte JV. Creation of a quantitative recipient risk index for mortality prediction after cardiac transplantation (IMPACT). Ann Thorac Surg 2011;92:914–21. discussion 921–2.
10. Ravi Y, Lella SK, Copeland LA, Zolfaghari K, Grady K, Emani S, Sai-Sudhakar CB. Does recipient work status pre-transplant affect post-heart transplant survival? A United Network for Organ Sharing database review. J Heart Lung Transplant 2018;37:604–10.
11. Colvin-Adams M, Valapour M, Hertz M, Heubner B, Paulson K, Dhungel V, et al. Lung and heart allocation in the United States. Am J Transplant 2012;12:3213–34.
12. Shortliffe EH, Sepulveda MJ. Clinical decision support in the era of artificial intelligence. JAMA 2018 Dec 4;320(21):2199–200.
13. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. JAMA Cardiol 2017;2:204–9.
14. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation 2018;138:1623–35.
15. Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. Int J Med Inform 2015;84:1094–8.
16. Colvin-Adams M, Smith JM, Heubner BM, Skeans MA, Edwards LB, Waller CD, et al. OPTN/SRTR 2013 annual data report: heart. Am J Transplant 2015;15(Suppl 2):1–28.
17. Medved D, Ohlsson M, Hoglund P, Andersson B, Nugues P, Nilsson J. Improving prediction of heart transplantation outcome using deep learning techniques. Sci Rep 2018 Feb 26;8(1):3613.