Transplant International

ORIGINAL ARTICLE

Machine learning to predict transplant outcomes: helpful or hype? A national cohort study

Sunjae Bae[1,2,3], Allan B. Massie[1,2], Brian S. Caffo[3], Kyle R. Jackson[2] & Dorry L. Segev[1,2]

1 Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD, USA
2 Department of Surgery, Johns Hopkins School of Medicine, Baltimore, MD, USA
3 Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD, USA

Correspondence
Dorry L. Segev MD, PhD, Department of Surgery, Johns Hopkins Medical Institutions, 2000 E Monument St., Baltimore, MD 21205, USA.
Tel.: 410-502-6115;
fax: 410-614-2079;
e-mail: dorry@jhmi.edu

SUMMARY

An increasing number of studies claim machine learning (ML) predicts transplant outcomes more accurately. However, these claims were possibly confounded by other factors, namely, supplying new variables to ML models. To better understand the prospects of ML in transplantation, we compared ML to conventional regression in a “common” analytic task: predicting kidney transplant outcomes using national registry data. We studied 133 431 adult deceased-donor kidney transplant recipients between 2005 and 2017. Transplant centers were randomly divided into 70% training set (190 centers/97 787 recipients) and 30% validation set (82 centers/35 644 recipients). Using the training set, we performed regression and ML procedures [gradient boosting (GB) and random forests (RF)] to predict delayed graft function, one-year acute rejection, death-censored graft failure C, all-cause graft failure, and death. Their performances were compared on the validation set using -statistics. In predicting rejection, regression ($C = {}_{0.601}0.611_{0.621}$) actually outperformed GB ($C = {}_{0.581}0.591_{0.601}$) and RF ($C = {}_{0.569}0.579_{0.589}$). For all other outcomes, the C-statistics were nearly identical across methods (delayed graft function, 0.717–0.723; death-censored graft failure, 0.637–0.642; all-cause graft failure, 0.633–0.635; and death, 0.705–0.708). Given its shortcomings in model interpretability and hypothesis testing, ML is advantageous only when it clearly outperforms conventional regression; in the case of transplant outcomes prediction, ML seems more hype than helpful.

Transplant International 2020; 33: 1472–1480

Key words
kidney transplantation, machine learning, prediction, regression

Received: 2 October 2019; Revision requested: 30 November 2019; Accepted: 29 June 2020;
Published online: 28 September 2020

Introduction

Machine learning (ML) algorithms have emerged as alternatives to conventional regression modeling, largely due to their ability to analyze nontabular (e.g., image or natural language) and high-dimensional (typically >10 000 variables) data (Table 1) [1]. It seems natural that transplantation researchers are drawn to these methods, especially considering the frequent use of large registry data analyses in transplantation [2]. Indeed, this is a growing area of investigation [3–16], with several recent studies that reported superior predictive performance of ML in predicting delayed graft function, graft survival, and mortality after kidney or liver transplantation [3,4,6,16].

However, the reported improvements in predictive performance may not be fully attributable to ML, because those studies often supplied more clinical

doi:10.1111/tri.13695

Table 1. A brief comparison of regression and machine learning.

	Regression	Machine learning
Mathematical assumptions	Several	Usually fewer
Analyzing high-dimensional data (e.g., >10 000 variables)	Possible, but labor-intensive	Capable
Analyzing nontabular data (e.g., images, clinical notes)	Limited	Capable, but often requires extensive labor/resources
Model interpretability	Fully transparent and human-readable	Limited or absent
Ability to incorporate prior clinical/biological knowledge	Capable (e.g., assisted variable selection)	Limited or absent
Hypothesis testing	Built-in	Limited or absent

information to the ML models than to the conventional regression-based models. In other words, had the ML models been developed on the same set of variables as the regression-based models, we might have observed minimal or no gain in predictive performance. From a theoretical standpoint, the predictive performance of a regression model should at least match that of a ML algorithm under ideal conditions [17], and even violations of these conditions can mostly be addressed using statistical techniques. As such, a properly developed regression model is expected to perform similarly to ML.

Furthermore, a key limitation of most ML algorithms is that they deliver "black-box" predictions, whereas regression provides interpretable models, allows face validity checking, and enables biological hypothesis testing. These black-box predictions can sometimes be driven by senseless associations. For example, it was discovered that a ML algorithm, trained to determine malignancy from images of skin lesions, diagnosed lesions as malignant when a ruler was pictured near the lesion, because, in the training data, a ruler was drawn when the pathologist suspected a malignancy; identifying this harmful quirk was difficult because the ML was a black box that did not show how it was evaluating the images [18,19]. Since ML is entirely data-driven and does not reveal its mechanism so that face validity can be checked, these approaches are not risk-free.

To better understand the possible role of ML in transplantation, we aimed to evaluate the performance of ML algorithms in a "common" study setup relevant to a wide gamut of transplantation research. Thus, we conducted a head-to-head comparison of ML algorithms versus regression in predicting various kidney transplant (KT) outcomes using the same populations and the same set of variables abstracted from the U.S. national registry data.

Materials and methods

Data source

This study used data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donors, waitlisted candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), US Department of Health and Human Services, provides oversight to the activities of the OPTN and SRTR contractors. A detailed description of the data has been provided elsewhere [2]. This study used de-identified registry data and was exempted by the Johns Hopkins Medicine Institutional Review Boards (NA_00042871).

Study population

Our study included 133 431 adult (18 or older) deceased-donor KT recipients at 272 KT centers from January 1, 2005, to December 31, 2017. The dataset was randomly divided at center level into a 70% training set (190 centers; 97 787 recipients) and a 30% validation set (82 centers; 35 644 recipients).

Outcomes

We studied five outcomes: delayed graft function (DGF), one-year acute rejection (AR), death-censored graft failure (DCGF), death, and all-cause graft failure (ACGF). DGF was defined as the need for dialysis within the first week after transplant. AR included all acute rejection episodes reported up to one-year follow-up. Since the exact dates of the episodes are not available on OPTN/SRTR data, AR was treated as a binary outcome, as opposed to a time-to-event outcome.

DCGF was defined as the time from KT to graft failure (re-initiation of dialysis or re-KT), censoring for death. ACGF was defined as the time from KT to graft failure or death. Graft failure and death were collected by OPTN from multiple sources, including follow-up reports from transplant centers, Centers for Medicare & Medicaid Services ESRD Death Notification Form (CMS 2746), and the Social Security Death Master File. All recipients were censored at the end of study on December 31, 2017.

Model development

We developed prediction models on the 70% training set using generalized linear regression and two ML techniques: gradient boosting (GB) and random forests (RF).

For regression, we conducted logistic regressions on DGF and AR, and Cox regressions on DCGF, death, and ACGF. Missing values of the covariables were handled using multiple imputation with 10 iterations. To address any nonlinear associations of continuous variables and clinical outcomes, we included linear spline terms into the models. Knots were determined based on previous literature and exploratory data analyses, which involved comparing the fit of univariable ACGF models using different sets of knots. The regression models were finalized using these knots (Table S1).

Gradient boosting was performed using the R package "XGBoost" [20]. We used the logistic objective function for DGF and AR, and Cox proportional hazard objective function for DCGF, death, and ACGF. Missing values of the covariables were imputed during training in a way that is analogous to multiple imputation [20]. The tuning parameters were chosen via cross-validation on the 70% training dataset.

Random forests was performed using the R package "rfsrc" [21]. We used the Gini splitting rule for DGF and AR, and the log-rank splitting rule for DCGF, death, and ACGF [21]. Similar to GB, missing values of the covariables were imputed during training. The tuning parameters were chosen via cross-validation on the 70% training dataset.

All prediction models included the same set of covariables, including donor variables (age, race, sex, ABO blood type, height, weight, stroke as the cause of donor death, terminal serum creatinine, cytomegalovirus (CMV), hepatitis C, diabetes, hypertension, donation after and cardiac death), recipient variables (age, sex, race, primary cause of end-stage renal disease, ABO blood type, primary insurer, body mass index (BMI), human immunodeficiency virus (HIV), CMV, hepatitis B, hepatitis C, Epstein–Barr virus, previous transplant, pre-emptive transplant, time on dialysis, panel reactive antibody (PRA), diabetes, hypertension, previous malignancy, symptomatic peripheral vascular disease, total serum albumin, and education level), and transplant variables (HLA-A/B/DR mismatches and cold ischemic time).

Evaluation of predictive performance: *C*-statistic

We used the 30% validation set to evaluate the predictive performance of the models. Our primary measure of predictive performance was the *C*-statistic. The *C*-statistic is a measure of discrimination, that is, whether the model correctly assigns higher predicted risk to those who actually develop the outcome versus those who do not. Specifically, the *C*-statistic was derived using the area under the receiver operating characteristic curve (AUROC) for binary outcomes (DGF and AR), and Harrell's concordance for time-to-event outcomes (DCGF, death, and ACGF) [22]. In addition, we conducted a sensitivity analysis in which the *C*-statistics for the time-to-event outcomes were estimated again using a novel method proposed by Uno and colleagues [23]. Unlike the conventional method, Uno's *C*-statistic is independent from the censoring distribution of the study population.

We first estimated the *C*-statistic over the entire validation set, and then in 16 subgroups stratified by quartiles of Kidney Donor Profile Index (KDPI) and Estimated Post-Transplant Survival (EPTS) to identify whether the predictive performance of the models vary by donor and recipient risk level [24]. In 3965 (3.0%) recipients, KDPI and EPTS values could not be calculated due to missing values. These recipients were excluded from the stratified analysis.

Evaluation of predictive performance: brier score

Our secondary measure of predictive performance was the Brier score [25]. The Brier score is a measure of calibration, that is, how close the predicted risk is to the actual risk. A lower Brier score indicates a smaller difference between the predicted and actual risk, hence superior calibration. It is important to assess both discrimination and calibration as models with superior discrimination may have inferior calibration, with over- or under-predicted risk [26]. For time-to-event outcomes (DCGF, death, and ACGF), we used the integrated Brier score, an extension of the Brier score for time-to-event outcomes [27]. In addition, we created calibration plots to visualize the calibration of the prediction methods across the spectrum of predicted risk. We stratified the validation set into 20 equally distanced bins by the predicted risk of the outcome and estimated the observed risk within each bin. The observed risk was defined as the incidence ratio for binary outcomes and as the cumulative incidence at 5 years post-KT for time-to-event outcomes.

Statistical analysis

All analyses were performed using R version 3.5.0. We used subscripts to indicate 95% confidence intervals as per the Louis and Zeger style [28].

Results

Study population

Overall, the training set and the validation set showed similar characteristics. Median recipient age was 54 years in both sets. Median donor age was 41 years in the training set and 39 in the validation set. The training set included 39.7% female recipients, 32.6% African American recipients, 39.8% female donors, and 14.1% African American donors. The validation set included 39.7% female recipients, 34.7% African American recipients, 39.6% female donors, and 14.3% African American donors (Table 2).

In the training set, 27.4% of the recipients developed DGF and 11.1% developed AR. The 5-year cumulative incidences were 13.8% for DCGF, 15.4% for death, and 24.5% for ACGF. The median follow-up was 4.8 years for death and 4.2 years for DCGF and ACGF.

Predictive performance: *C*-statistic

In our comparison of predictive performance in the validation set, ML algorithms did not show superior discrimination over regression in any of the five outcomes we studied (Fig. 1). Of note, regression actually showed higher *C*-statistic ($C = {}_{0.601}0.611_{0.621}$) than both ML algorithms, GB ($C = {}_{0.581}0.591_{0.601}$) and RF ($C = {}_{0.569}0.579_{0.589}$), in predicting one-year AR. For all other outcomes, the three methods showed nearly identical performance. For DGF, the *C*-statistics were ${}_{0.714}0.721_{0.727}$ for regression, ${}_{0.717}0.723_{0.729}$ for GB, and ${}_{0.711}0.717_{0.723}$ for RF. For DCGF, the *C*-statistics were ${}_{0.629}0.637_{0.646}$ for regression, ${}_{0.633}0.642_{0.650}$ for GB, and ${}_{0.629}0.638_{0.646}$ for RF. For death, the *C*-statistics were ${}_{0.701}0.708_{0.715}$ for regression, ${}_{0.698}0.705_{0.712}$ for GB, and ${}_{0.698}0.705_{0.713}$ for RF. For ACGF, the *C*-statistics were ${}_{0.628}0.634_{0.640}$ for regression, ${}_{0.629}0.635_{0.641}$ for GB, and ${}_{0.627}0.633_{0.639}$ for RF. Across the 16 subgroups of the 30% validation set stratified by the quartiles of KDPI and EPTS, regression, GB, and RF showed very similar predictive performance in all five outcomes (Fig. S1).

We found similar trends in our sensitivity analysis where the *C*-statistics were estimated using Uno's method for time-to-event outcomes. For DCGF, Uno's *C*-statistics were 0.623 for regression, 0.624 for GB, and 0.611 for RF. For death, Uno's *C*-statistics were 0.707 for regression, and 0.703 for both GB and RF. For ACGF, Uno's *C*-statistics were 0.635 for both regression and GB, and 0.632 for RF.

Predictive performance: brier score

All methods showed similar calibration in predicting binary outcomes (DGF and AR), whereas ML algorithms showed inferior calibration than regression in predicting time-to-event outcomes (DCGF, death, and ACGF). The Brier scores were very similar between the methods for DGF (regression, 0.161; GB, 0.160; and RF, 0.161) and for AR (regression, 0.089; GB, 0.090; and RF, 0.091). In contrast, regression showed lower Brier scores (i.e., smaller prediction errors) than ML algorithms for DCGF (regression, 0.179; GB, 0.187; RF, 0.185), death (regression, 0.183; GB, 0.206; and RF, 0.197), and ACGF (regression, 0.193; GB, 0.201; and RF, 0.208; Table 3). In addition, our calibration plots suggested that all three prediction methods had comparable calibration across the spectrum of predicted risk (Fig. 2).

Discussion

In this comparison of ML algorithms versus regression in predicting KT outcomes using large national registry data, ML did not outperform regression-based models. In terms of discrimination, we observed similar *C*-statistics across regression and ML algorithms in all transplant outcomes, with the exception of one-year AR where logistic regression actually showed a higher *C*-statistic than ML algorithms. Furthermore, in terms of calibration as measured in the Brier score, regression outperformed ML algorithms in predicting time-to-event outcomes (DCGF, death, and ACGF), whereas regression and ML algorithms showed similar performance in predicting binary outcomes (DGF and AR).

Table 2. Population characteristics.

Clinical factor	Training set (*n* = 97 787)	Validation set (*n* = 35 644)
Recipient factors		
Age (year), median (IQR)	54 (44, 63)	54 (43, 63)
Female	39.7%	39.7%
Race		
White	42.5%	42.1%
African American	32.6%	34.7%
Hispanic/Latino	16.2%	15.0%
Other/multi-racial	8.6%	8.3%
Preemptive transplant	9.9%	10.7%
Time on dialysis (year), median (IQR)	3.5 (1.6, 5.7)	3.6 (1.6, 5.9)
Cause of ESRD		
Glomerulonephritis	21.9%	22.3%
Diabetes	26.9%	27.6%
Hypertension	23.6%	24.3%
Others	27.7%	25.8%
Panel reactive antibody		
0–9	54.1%	50.7%
10–79	23.3%	26.4%
80–100	16.8%	17.7%
Missing	5.7%	5.2%
BMI (kg/m^2), median (IQR)	27.6 (24.1, 31.7)	27.8 (24.2, 31.8)
Previous transplants	14.7%	14.6%
Cold ischemic time (h), median (IQR)	17.0 (11.7, 23.0)	16.4 (11.0, 22.4)
Donor factors		
Age (year), median (IQR)	41 (25, 52)	39 (25, 51)
Female	39.8%	39.6%
Race		
White	69.0%	68.1%
African American	14.1%	14.3%
Hispanic/Latino	13.6%	14.2%
Other/multi-racial	3.3%	3.4%
Terminal serum creatinine (mg/dl), median (IQR)	0.9 (0.7, 1.3)	0.9 (0.7, 1.3)
Donation after cardiac death	16.4%	14.1%

BMI, body mass index; ESRD, end-stage renal disease; IQR, interquartile range.

Predicting KT outcomes using the U.S. national registry data is perhaps one of the “generic” analytic tasks that pertain to a wide gamut of transplantation research, ranging from fine-tuning organ allocation policy [29,30] to informing clinical decision-making [14,31] to identifying independent effects by correctly adjusting for confounders [32,33]. The lack of ML’s advantage in this setting implies that, despite the recent successes, and recent claims of successes, surrounding ML in many areas of medicine, regression is a valuable, and sometimes a preferable, analytic method in transplantation research.

Our findings are consistent with a study in heart transplantation by Miller *et al.* [34] that found no meaningful difference in predicting 1-year survival between logistic regression and ML algorithms using the same set of variables, with *C*-statistics around 0.65 in most methods. We have extended this approach to kidney transplantation, to outcomes beyond 1 year, to Cox regression which is the typical method for evaluating survival, and to nonsurvival outcomes such as DGF and AR. Our findings are also consistent with studies [3–5,35] that reported only minor performance differences (e.g., *C*-statistic from 0.706 to 0.724); we extend these studies in the context of a true head-to-head comparison that shows no performance advantage of ML and, actually, some performance advantage of regression with some outcomes.

On the other hand, our findings are contrary to several recent studies that reported high predictive performance of ML algorithms. In some cases, the exact reason for the discrepancy is unclear due to the absence of a head-to-head, same-variable, same-population comparison against regression [7–10]. But, more importantly, many of these studies purposefully explored ML as a tool to incorporate additional clinical information into prediction [6,13,16], rather than testing if ML outperforms regression on equal footing. For example, Lau *et al.* [16] reported that RF and neural network outperformed traditional models such as Donor Risk Index (DRI) in predicting graft survival after liver transplantation. However, the ML models included numerous key variables that are not included in DRI, such as recipient disease category, donor serum albumin level, and geographical location. Therefore, these studies have shown that ML methods identified potentially more influential clinical factors which led to better prediction. Purely in terms of predictive performance, these studies do not indicate that ML alone can achieve a new level of predictive performance that regression cannot reach, because an equally comprehensive regression-based model could have demonstrated similar performance. In that sense, our current study is a necessary follow-up to the previous ML prediction studies.

Although our findings do not support the application of ML on simple prediction of KT outcomes using routinely collected tabular data, there are research questions in organ transplantation that might be well suited for ML. Theoretically, ML methods are capable of handling

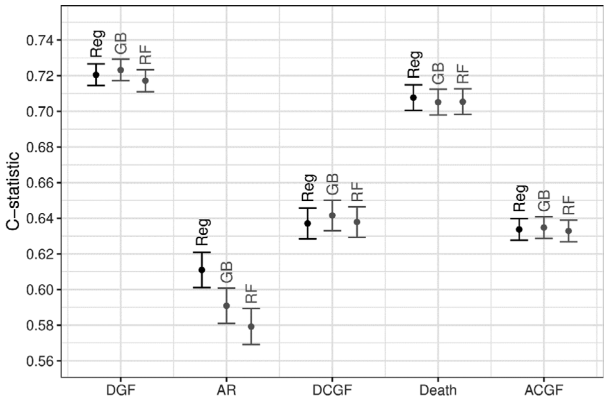


Figure 1 Predictive performance of regression (Reg), gradient boosting (GB), and random forests (RF) in predicting kidney transplant outcomes, as measured in the *C*-statistic. ACGF, all-cause graft failure; AR, one-year acute rejection; DCGF, death-censored graft failure; DGF, delayed graft function. Regression represents logistic regressions for delayed graft function and acute rejection, and Cox regressions for death-censored graft failure, death, and all-cause graft failure. *Y*-axis indicates the area under the receiver operating characteristic curve (AUROC) for delayed graft function and acute rejection, and Harrell's concordance statistic for death-censored graft failure, death, and all-cause graft failure.

Table 3. Predictive performance of regression, gradient boosting, and random forests in predicting kidney transplant outcomes, as measured in the Brier score.

Outcome	Regression	Gradient boosting	Random forests
Delayed graft function	0.161	0.160	0.161
Acute rejection, one year	0.089	0.090	0.091
Death-censored graft failure	0.179	0.187	0.185
Death	0.183	0.206	0.197
All-cause graft failure	0.193	0.201	0.208

Lower Brier score indicates superior calibration.

interactions between predictors in a flexible manner [14,36], integrating nontabular data such as clinical notes or graft biopsy images with tabular clinical data [13,37], and analyzing high-dimensional data such as genes or biomarkers [38]. Our study was not focused on evaluating these benefits, and our findings should not discourage future applications of ML on such research endeavors. However, in the context of straightforward outcome prediction, we emphasize that ML does not seem to provide a predictive advantage, yet suffers from a number of weaknesses that risk misleading modeling, limit our ability to assess face validity or test biological hypotheses, and diminish the interpretability of the models themselves.

Our study has several limitations. First, we cannot rule out the possibility that there exists a ML algorithm that outperforms the algorithms investigated in this study. However, considering that we observed nearly identical predictive performance from all three methods including regression, it is not unreasonable to assume that these performance measures are bound by the inherent variability of the data, not by the competency of the methods. Second, our findings are not generalizable to any analyses that include new types of data not present in the transplant national registry, especially nontabular clinical information. As discussed above, ML might be actually advantageous in these cases. Lastly, there could be specific subgroups in which GB or RF outperforms regression models because of their ability to handle interactions without modeling assumptions. However, such effects were not observed in our analyses.

Our findings suggest that ML does not outperform conventional regression-based approaches in predicting various KT outcomes using routinely collected tabular data. Given that regression modeling presents an interpretable model and enables hypothesis testing, the advantage of using ML over regression in simple predictions of KT outcomes is questionable. The lack of ML's advantage in our "generic," controlled analytic setting implies that, in this case, ML is more hype than helpful.

Authorship

SB, ABM, BSC, KRJ and DLS: participated in the design of the study and writing of the manuscript. SB: performed the data analysis.

Funding

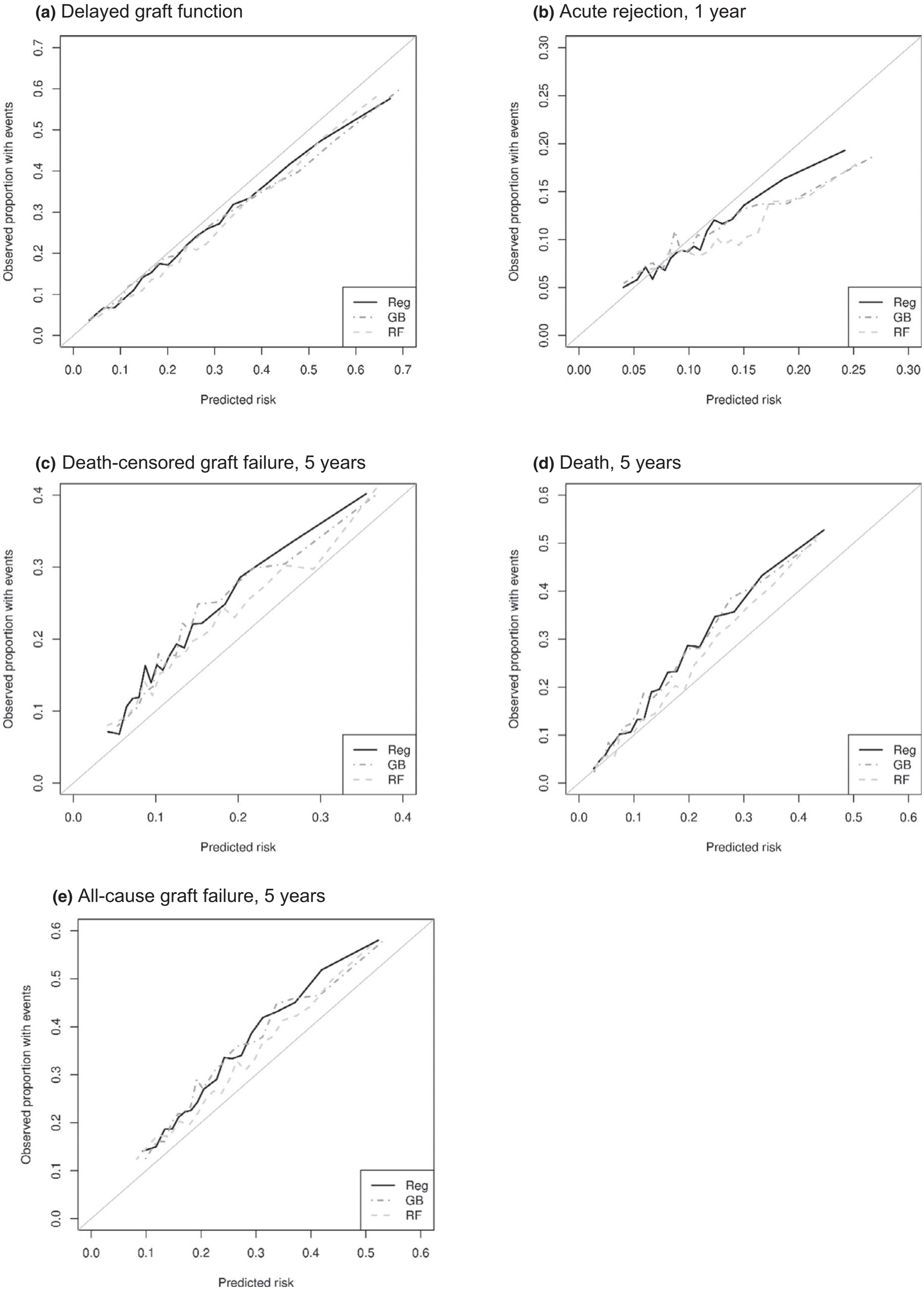
This work was supported by ASN Foundation for Kidney Research (Bae) and the National Institute of Diabetes and Digestive and Kidney Diseases (K01DK101677, Massie; F32DK113719, Jackson; and K24DK101828, Segev).

Conflict of interest

The authors have declared no conflicts of interest.

Acknowledgements

The analyses described here are the responsibility of the authors alone and do not necessarily reflect the



(a) Delayed graft function

(b) Acute rejection, 1 year

(c) Death-censored graft failure, 5 years

(d) Death, 5 years

(e) All-cause graft failure, 5 years

Figure 2 Calibration plot of regression (Reg), gradient boosting (GB), and random forests (RF) in predicting kidney transplant outcomes.

views or policies of the Department of Health and Human Services, nor do mention of trade names, commercial products or organizations imply endorsement by the U.S. Government. The data reported here have been supplied by the Hennepin Healthcare Research Institute as the contractor for the Scientific Registry of Transplant Recipients (SRTR). 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 SRTR or the U.S. Government.

SUPPORTING INFORMATION

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

Figure S1. Predictive performance of regression (Reg), gradient boosting (GB), and random forests (RF) in predicting kidney transplant outcomes in Kidney Donor Profile Index (KDPI) and Estimated Post-Transplant Survival (EPTS) strata.

Table S1. Regression coefficients.

REFERENCES

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436.
2. Massie AB, Kuricka LM, Segev DL. Big data in organ transplantation: registries and administrative claims: big data in organ transplantation. *Am J Transplant* 2014; **14**: 1723.
3. Decruyenaere A, Decruyenaere P, Peeters P, Vermassen F, Dhaene T, Couckuyt I. Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods. *BMC Med Inform Decis Mak* 2015; **15**: 83.
4. Mark E, Goldsman D, Gurbaxani B, Keskinocak P, Sokol J. Using machine learning and an ensemble of methods to predict kidney transplant survival. *PLoS One* 2019; **14**: e0209068.
5. Lin RS, Horn SD, Hurdle JF, Goldfarb-Rumyantzev AS. Single and multiple time-point prediction models in kidney transplant outcomes. *J Biomed Inform* 2008; **41**: 944.
6. Yoo KD, Noh J, Lee H, *et al.* A machine learning approach using survival statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study. *Sci Rep* 2017; **7**: 8904.
7. Shahmoradi L, Langarizadeh M, Pourmand G, Fard Z, Borhani A. Comparing three data mining methods to predict kidney transplant survival. *Acta Inform Med* 2016; **24**: 322.
8. Brown TS, Elster EA, Stevens K, *et al.* Bayesian modeling of pretransplant variables accurately predicts kidney graft survival. *Am J Nephrol* 2012; **36**: 561.
9. Greco R, Papalia T, Lofaro D, Maestripieri S, Mancuso D, Bonofiglio R. Decisional Trees in renal transplant follow-up. *Transplant Proc* 2010; **42**: 1134.
10. Topuz K, Zengul FD, Dag A, Almehmi A, Yildirim MB. Predicting graft survival among kidney transplant recipients: a Bayesian decision support model. *Decis Support Syst* 2018; **106**: 97.
11. Dag A, Oztekin A, Yucel A, Bulur S, Megahed FM. Predicting heart transplantation outcomes through data analytics. *Decis Support Syst* 2017; **94**: 42.
12. Lasserre J, Arnold S, Vingron M, Reinke P, Hinrichs C. Predicting the outcome of renal transplantation. *J Am Med Inform Assoc* 2012; **19**: 255.
13. Srinivas TR, Taber DJ, Su Z, *et al.* Big data, predictive analytics, and quality improvement in kidney transplantation: a proof of concept. *Am J Transplant* 2017; **17**: 671.
14. Bae S, Massie AB, Thomas AG, *et al.* Who can tolerate a marginal kidney? Predicting survival after deceased donor kidney transplant by donor-recipient combination. *Am J Transplant* 2019; **19**: 425.
15. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomed Signal Process Control* 2019; **52**: 456.
16. Lau L, Kankanige Y, Rubinstein B, *et al.* Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation* 2017; **101**: e125.
17. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY: Springer, 2009.
18. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. *J Invest Dermatol* 2018; **138**: 2108.
19. Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017; **542**: 115.
20. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA: ACM Press, 2016: 785–794.
21. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008; **2**: 841.
22. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med* 2004; **23**: 2109.
23. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011; **30**: 1105.
24. Organ Procurement and Transplantation Network. A Guide to Calculating and Interpreting the Estimated Post-Transplant Survival (EPTS) Score Used in the Kidney Allocation System (KAS) n.d. https://optn.transplant.hrsa.gov/media/1511/guide_to_calculating_interpreting_epts.pdf (accessed June 22, 2019).
25. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 1950; **78**: 1.
26. Alba AC, Agoritsas T, Walsh M, *et al.* Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA* 2017; **318**: 1377.
27. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999; **18**: 2529.
28. Louis TA, Zeger SL. Effective communication of standard errors and

confidence intervals. *Biostat Oxf Engl* 2009; **10**: 1.

29. Bae S, Massie AB, Luo X, Anjum S, Desai NM, Segev DL. Changes in discard rate after the introduction of the Kidney Donor Profile Index (KDPI). *Am J Transplant* 2016; **16**: 2202.
30. Zhou S, Massie AB, Luo X, *et al.* Geographic disparity in kidney transplantation under KAS. *Am J Transplant* 2018; **18**: 1415.
31. Bae S, Durand CM, Garonzik-Wang JM, *et al.* Anti-thymocyte globulin versus interleukin-2 receptor antagonist in kidney transplant recipients with hepatitis C virus. *Transplantation* 2020; **104**: 1294.
32. Kucirka LM, Durand CM, Bae S, *et al.* Induction immunosuppression and clinical outcomes in kidney transplant recipients infected with human immunodeficiency virus. *Am J Transplant* 2016; **16**: 2368.
33. Orandi BJ, Luo X, Massie AB, *et al.* Survival benefit with kidney transplants from hla-incompatible live donors. *N Engl J Med* 2016; **374**: 940.
34. Miller PE, Pawar S, Vaccaro B, *et al.* Predictive abilities of machine learning techniques may be limited by dataset characteristics: insights from the UNOS database. *J Card Fail* 2019; **25**: 479.
35. Luck M, Sylvain T, Cardinal H, Lodi A, Bengio Y. Deep learning for patient-specific kidney graft survival analysis. ArXiv170510245 Cs Stat, 2017.
36. Foster JC, Liu D, Albert PS, Liu A. Identifying subgroups of enhanced predictive accuracy from longitudinal biomarker data by using tree-based approaches: applications to fetal growth. *J R Stat Soc Ser A Stat Soc* 2017; **180**: 247.
37. Wood-Trageser MA, Lesniak AJ, Demetris AJ. Enhancing the value of histopathological assessment of allograft biopsy monitoring. *Transplantation* 2019; **103**: 1306.
38. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015; **16**: 321.