Predicting deceased donor kidney transplant outcomes: Comparing KDRI/KDPI with machine learning

Eric Pahl (✉ eric-pahl@uiowa.edu)

University of Iowa <https://orcid.org/0000-0002-8635-3012>

Nick Street

University of Iowa

Hans Johnson

University of Iowa

Alan Reed

University of Iowa

Research article

Keywords: Kidney transplantation, end-stage renal disease, KDRI/KDPI, machine learning

Posted Date: November 6th, 2019

DOI: <https://doi.org/10.21203/rs.2.16892/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Kidney transplantation is a cost-effective treatment for end-stage renal failure patients that provides a significant survival benefit and improves their quality of life compared to other forms of renal replacement. The predominant method used for donor kidney quality assessment is the Cox regression-based, piecewise linear kidney donor risk index (KDRI). A machine learning method (random forest) was compared to KDRI for predicting graft failure at 12, 24, and 36 months after transplantation.

Methods: Random forest was trained and evaluated with the same deceased donor kidney transplant data (n=70242) initially used to develop KDRI (1995-2005) and included four readily available recipient variables from the estimated post-transplant survival score. **Results:** When comparing type II error rates of 10%, random forests predicted an additional 2,148 successful grafts at 36 months after transplant (126%) than KDRI. Many high-KDRI kidneys, at risk for discard, were correctly predicted for successful transplantation with random forests. Random forest performed significantly better than KDRI for graft Kaplan-Meier survival analysis from 0-240 months (log-rank test $p < 0.00$). **Conclusions:** Machine learning methods can provide a significant improvement over KDRI for the assessment of kidney offers. This work lays the foundation for the use of machine learning methodologies in transplantation and describes the steps to measure, analyze, and validate future models.

Background

There is little research into the methodology regarding how organ transplant stakeholders make decisions, predictions, and assessments of viability and matching for deceased donor kidney transplantation at the time of organ offer and acceptance. Clinical decisions have traditionally relied on the time-tested orthodoxy of data derived from Cox regression-based models to provide statistical relevance to decision making. Recent advances in machine learning methods (MLM) provide the opportunity to create highly nonlinear models with complex interactions among variables that may provide superior predictive power. MLM usage in medical domains is increasing as demonstrated by successful applications for predicting better utilization of perioperative antibiotics, predicting hospital lengths of stay, and indeed even recently in formulating alternative models for organ distribution and post-transplant implied utilization.¹⁻⁵

Kidney transplantation is a cost-effective treatment for end-stage renal disease (ESRD) patients that provides a significant survival benefit and improves their quality of life compared to other forms of renal replacement; patients gain an estimated six quality-adjusted life years beyond other treatment methods. More than 600,000 ESRD patients in the US are on chronic dialysis therapy and the waitlist for life-saving kidney transplants numbers over 100,000 due to a shortage of identified, sufficiently low-risk, transplantable kidneys.⁶ Unfortunately, approximately 10,000 waiting patients die each year (~22 per day).^{7,8} The numbers of patients removed from the waitlist due to death and being too sick to undergo

transplantation were 4,830 and 4,411 respectively in 2016.⁶ The unacceptably high number of deaths from the waitlist could be prevented by increasing deceased donor kidney transplantation (DDKT).

Despite evidence-based clinical and cost advantages of DDKT, nearly 1 in 5 viable deceased-donor kidneys recovered for transplantation are discarded (~4,000 per year).⁶ Many studies have demonstrated a significant survival benefit for wait-listed patients to accept (or for centers to accept on their behalf) any kidney available for transplant, regardless of the current acceptance metrics.⁹⁻¹² The reasons for these discards are multifactorial and include, but are not limited to, recipient variability, donor variability, governmental regulations, local health provider biases, and financial pressures.¹³ A comprehensive discussion of how these factors impact the health provider decision making at the time of organ offer and acceptance is beyond the scope of this manuscript.

Healthcare professionals can better understand the risk of transplantation with models that capture an individual's health state more entirely in the context of a specific prospective organ. Our paper uses MLM to recreate the Kidney Donor Risk Index (KDRI) to determine if MLM can lead to a better predictive model than the Cox regression initially used. The Cox regression employed to develop the KDRI resulted in a piecewise linear formula used both in donor allocation and distribution (recipient acceptance criteria). We also question why only donor variables are considered at this stage for allocation purposes? The MLM may provide additional modeling opportunities that incorporate more variables and answer new questions.

The KDRI is commonly adjusted annually and implemented as the derivative metric, the Kidney Donor Profile Index (KDPI). The KDRI was developed in 2009 and has been the industry and regulatory standard for kidney quality since 2011.¹⁴ Despite the need and the industry's enthusiasm for the adoption of this metric, the KDRI has many limitations. The KDRI has a c-statistic of 0.600, a measure of predictive quality equivalent to the area under the receiver operating characteristic (ROC) curve (AUC). The KDRI was developed using a Cox proportional hazard regression method. The final KDRI model included 15 variables measured from the donor, resulting in a piecewise linear model that suffers biases when arbitrarily categorized variables are present. For example, in the KDRI model the risk coefficient change based on age over or under 18 and 50; weight over or under 80; creatinine over or under 1.5; etc.¹⁴ By design, the KDRI model incorporates only variables measured from the donor to assess the risk of transplantation for a recipient.

The identified limitations of KDRI lead to several questions that this work explores: Can more successful transplants be predicted using MLM? Might more accurate predictions of graft failure (GF) at 12, 24, and 36 months post-transplantation encourage the utilization of kidneys that are otherwise at risk for discard for transplantation in patients? Further, it would seem logical that predictive measures of transplant outcomes should incorporate recipient variables, like those present in the estimated post-transplant survival score (EPTS) to further improve the modeling. The EPTS was also developed as a piecewise linear model (age over or under 25), using four recipient variables, and is used separately in allocation algorithms. Because EPTS and KDRI were constructed independently of one another, the simultaneous use of these separate models, as in the current allocation system, cannot capture interactions among the donor and recipient variables. A combined model including readily available variables from the donor and recipient, utilizing machine learning, may improve the predictive capabilities for longer-term graft survival.

Machine learning methods are routinely used to capture non-linear relationships in the presence of outliers and noise without compromising predictive quality, adding significant variance, or overtraining.^{15–17} Complicated relationships among many variables, potentially involving independently insignificant variables, cannot be modeled adequately with linear methods. Variables with known collection challenges (i.e., apparent outliers, or high noise) are routinely excluded from linear modeling but may provide enhanced knowledge to MLM allowing better predictive performance. Retrospective transplant data are sparse with many missing and miscoded values. MLM have demonstrated success in these dirty data conditions in other areas of medicine.^{18–20}

The specific aim and purpose of this retrospective quantitative research study was to build an MLM using national DDKT and patient follow-up data from 1995-2005, the same data and variables used to develop KDRI, and compare predictive performance using 10-fold cross-validation for GF at 12, 24, and 36 months after transplant, both with and without recipient information.

Methods

Data

We obtained Standard Transplant Analysis and Research (STAR) data from the United Network for Organ Sharing (UNOS) inclusive of September 29, 1987, through March 31, 2016. This experiment complied with the data use agreement in accordance with applicable regulating bodies. The UNOS STAR data contained many common free-text entry errors that were re-coded from unknown, blank spaces, empty strings, "NA", "N/A", "na", 999, 998, (for example) to a consistent value of NA. We re-coded predictive variables as categorical or numeric values based on our interpretation of the STAR data documentation. The UNOS STAR data are incomplete, mutable, corrected, and backfilled as updates became available. The minor

discrepancies between the data used by Rao and in the present study are due to this feature and are presented in Table 1.

We reconstructed, as closely as possible, the original data set used by Rao for the development of KDRI.¹⁴ First, the data were filtered by transplantation date within the acceptable date range (01/01/1995 to 12/31/2005) for only DDKT, Initial Data n_i (see data reduction in the first row of Table 1). Excluded in sequence from the analysis were: recipients aged less than 18 years, recipients with a previous transplant, multi-organ transplant recipients, and ABO-incompatible patients, keeping consistent with Rao. We also removed observations with invalid and/or missing data: donor height (<50cm, >213cm), weight (<10kg, >175kg), and creatinine (<0.1mg/dL, >8.0mg/dL). Finally, we removed observations without a valid entry for the KDRI_RAO variable. Table 1 compares the number of removed observations from Rao's study and our study for each missing or invalid variable.¹⁴

The clinical setting for our experiment was to predict kidney transplant outcomes with the data present at the time of organ allocation; the same data a clinician would have or could approximate (such as an estimate of the expected cold ischemia time from the current data) at the time of organ offer. Donor variables were those used in the final model of KDRI: age, race, history of hypertension, history of diabetes, serum creatinine, cerebrovascular cause of death, height, weight, donation after cardiac death, hepatitis C virus status, HLA-B and HLA-DR mismatching, en-bloc transplant, and double kidney transplant indicators (two for one), and known cold ischemia time at time of offer. Additional transplant recipient variables from EPTS score were used in our MLM; recipient age, diabetes, and time on dialysis - notably excluding recipients with prior transplant and multiorgan transplants based on the filter criteria from Rao, et al. 2009.¹⁴

Machine Learning Models

Machine learning models are predictive models that are induced semi-automatically from labeled data. The input consists of instances (e.g., organ characteristics, or donor/recipient pair characteristics) with ground-truth labels (e.g., success or failure). The supervised learning algorithm (e.g., random forests) optimizes a mapping from inputs to outputs, such that future unlabeled cases can be predicted correctly with high probability. Such techniques are often designed to handle large datasets with nonlinear effects, multiple dependencies, missing values, etc. This flexibility requires careful parameter tuning to avoid overly-complex models that generalize poorly to new, unseen cases. Oftentimes, MLM involve simple algorithms that are repeated thousands of times and making small changes as it converges to an optimal performance value (e.g. minimal Type II error rate). The general limitations for these processes are

computing power and parallel processing capabilities – two resources that continue to grow exponentially (e.g. Moore's Law).

In contrast to Rao's Cox proportional hazard regression method, we explored a supervised random forest (RF) MLM classification model. The RF algorithm constructs and combines the predictions of thousands of machine-generated decision trees to model the probability of graft failure. We created response variables based on patient follow-up data after transplant. This study considered three binary outcome predictions: graft failure (GF) at 12, 24, and 36 months after DDKT were abbreviated as GF12, GF24, and GF36 respectively. For example, the positive (+) outcome for GF12 meant that we observed graft loss within 12 months after DDKT and the negative (-) outcome for GF12 meant that graft was successful at 12-month follow-up. The three models tested for these outcomes were: 1. RF using only donor variables from KDRI (RFD), 2. RF using donor variables from KDRI and recipient variables from EPTS (RFDR), and 3. Rao's KDRI.

The RF algorithm involved the automated creation of thousands of decision trees trained with cases with known outcomes.²¹ The algorithm created each decision tree using a subset of the training data called a bootstrap sample (BSS). Each BSS balances the training sample by resampling the minority outcome observations (positives) and under sampling the majority outcome cases (negatives) such that the resulting ratio was 1:1. Each tree consists of multiple decision nodes constructed by randomly selecting a subset of the predictive variables and choosing the one that maximizes the Gini index, a measure of information gained from the use of each variable. Training continued until "exhaustion": each tree completely fits the training BSS. When classifying new examples, all the trees make a prediction, and the output of the RF is the percentage of votes for each outcome, one vote per tree, and the predicted outcome is assigned based on majority vote.

The choice of the RF algorithm allowed for mixed data types (binary, categorical, and numerical) without scaling or significant data modification. The RF training is computationally efficient, is robust to outliers and co-linearities, contains simple tuning parameters, and has demonstrated success for a variety of healthcare data applications.^{21,22} RFs have also demonstrated utility in predicting deceased donor organ transplantation success and offer acceptances in simulated organ allocation models.^{3,4}

The contribution of each tree in an RF is similar to getting thousands of opinions based on a professional colleague's background. The clinical implementation of the RF algorithm works similarly to secondary and tertiary opinions among professionals in multiple specialties convening for the treatment of a

complicated case. A random subset of all available variables and clinical observations informed the construction of every decision tree and resulted in a unique perspective represented by each tree. The consensus among the interdisciplinary experts, or decision trees, is the final treatment plan for the complicated case with thorough consideration among many different perspectives.

Figure 1 demonstrates the different numbers of trees used for RFD and RFDR models including stratifying and balancing to obtain BSS and the resulting AUC. Increasing the number of trees improved the AUC attained by RFD and RFDR, and both converged between 1000 and 1500 trees. The number of trees is a significant hyperparameter for the RF algorithm. KDRI results do not depend on the number of trees.

Figure 1. AUC for RF models created with different numbers of trees. KDRI model's AUC performance does not depend on the number of trees.

Analyses

We evaluated the predictive performance of the RF models with an industry standard 10-fold cross-validation approach.²³ The completion of the cross-validation process yielded a ranked list of predicted transplant outcome probabilities. In 10-fold cross-validation, the models are evaluated on test data that are never used in training. Thousands of decision trees are generated in the RF algorithm and each tree is individually overfitted on a subsection of the data. The RF model, as a combination of thousands of trees, avoids overfitting which is demonstrated by the testing and validation steps and held-out data. The KDRI model yielded a score that was already present in the data as a variable: KDRI_RAO. We used the KDRI_RAO variable to rank the DDKT observations to compare them with the predicted ranked lists from the RF models. We evaluated the models, using the predicted ranked lists, by generating the ROC curve and calculating the AUC or c-statistic as standard measures of predictive quality.

Kaplan-Meier (KM) curves provided the basis for survival analysis and comparison of the models' predicted failure vs. success at each GF interval (12, 24 and 36 months). We chose a prediction cut-off at a 10% false negative rate (FNR: Type II error, the rate at which the model predicted graft success, but the graft failed) to divide the test cases into two groups. The log-rank test was used to determine statistical significance among the KDRI and MLM models' respective predicted DDKT success and failure groups.

Table 2 contains descriptions and definitions of predictions; true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The models' predictions are labeled "Predicted GF" and the

observed transplant outcomes are “Actual GF.” When the Predicted GF and Actual GF are the same, the model predicted correctly. There are two types of model prediction errors; False Positives (Type I errors) happen when a model predicts that the DDKT would fail but the observed data show success, this is a missed opportunity; False Negatives (Type II errors) happen when a model predicts that the DDKT would succeed but the observed data show failure. False Negatives are the worst prediction errors from a clinical perspective; predicting a success, transplanting, and the graft failing or patient dying.

Results

RFDR performed better than RFD in all scenarios: adding recipient variables improved the predictive quality of the RFDR model. This is particularly true when predicting longer graft survival outcomes where RFDR performed increasingly better with longer time periods; GF12 (AUC = 0.636), GF24 (AUC = 0.638), GF36 (AUC = 0.644). RFD and KDRI, both models built without recipient variables, did not predict as well as the RFDR model did at the longer graft outcomes.

Figure 2 shows ROC curves for each predictive model under different graft failure considerations. In Figure 2 (a), KDRI (AUC 0.630) performed the same as RFD (AUC 0.627, $p = 0.317$) and RFDR (AUC 0.636, $p = 0.052$) at GF12. In Figure 2 (b), KDRI (AUC 0.626) performed the same as RFD (AUC 0.622, $p = 0.187$) and significantly worse than RFDR (AUC 0.638, $p < 0.000$) at GF24. In Figure 2 (c), KDRI (AUC 0.622) performed the same as RFD (AUC 0.618, $p = 0.096$) and significantly worse than RFDR (AUC 0.644, $p < 0.000$) at GF36.

Figure 2 (a-c). ROC curves of three predictive models for different graft failure criteria. P-values were calculated from DeLong's comparison method between ROC curves for RFD and RFDR respectively vs. KDRI are shown. The vertical axis, Sensitivity, at 0.9 is equivalent to the 10% FNR cut-off used for comparing the models in Table 3. The diagonal line represents a ROC curve with an AUC of 0.500.

Table 3 (a) describes the predictions for each model fixed for a 10% FNR at GF 36 months for purposes of direct comparison (Type II error; model predicts success, but graft fails). The FNR cut off maintains that TP and FN are the same for each model. The percentages shown in Table 3 are relative to KDRI. RFD and RFDR performed significantly better than KDRI when predicting which kidney transplant matches would succeed (TN). RFD identified 154 (2%) more successful kidney transplants than KDRI. The RFDR, with additional recipient criteria, identified 2148 (26%) more successful kidney transplant matches than KDRI. Table 3 (b) shows the effect of using different FNR comparison cut-offs for the comparison of KDRI and RFDR at 36GF prediction. The TN, Delta TN, and other counts of additional correct predictions for successful DDKT were accumulated through the entirety of the 10-year study period, 1995-2005.

Figure 3 (a) - (f) shows the KM survival curves for each model in monthly intervals. Predicted failure (+) and predicted success (-) groups were split based on FN rate (FNR) at 10%. Results show the percentage of patients surviving in each group on the vertical axes and up to 250 months following transplant on the horizontal axes. Figure 3 (a) - (c) show KDRI and RFD directly compared at GF12, GF24, and GF36. Figure 3 (d) - (f) show KDRI and RFDR directly compared at GF12, GF24, and GF36. In all comparisons, FNR was held constant at 10% making all the predicted failure groups (+) statistically similar. RFD survival groups (-) were statistically similar when compared with KDRI survival groups (-) in GF12, GF24, and GF36. RFDR survival groups (-) were statistically significantly better than KDRI survival groups (-) in all GF classifications.

Figure 3 (a-f). Kaplan-Meier (KM) survival analysis comparison for KDRI and RF models using different graft failure criteria to split predicted survival groups for each model. Legends include the label for each line, the population size for that group at time zero, and the p-value calculated from the log-rank test between respective predicted outcome groups (i.e., - / +).

Discussion And Conclusions

Our objective was to develop MLM prediction tools to inform decisions for specific DDKT matches that were better than the current methods available. The RFD model performed slightly better than the existing KDRI model. The RFDR model demonstrated significantly higher performance than KDRI, especially for GF36 predicting thousands more successful DDKTs with the same prediction failure rate as KDRI. Safer clinical decisions informed by MLM could empower clinicians to transplant more organs into patients resulting in better outcomes. The demonstrated results of the RFDR experiments also indicate that using recipient data improves the performance of predicting longer DDKT graft survival. We believe it was appropriate to compare these models directly over the observation time (1995-2005) because KDRI was not implemented clinically or administratively until 2011.

Models with only donor variables, KDRI and RFD, showed decreased longevity predictions whereas RFDR (both donor and recipient variables) better predicted outcomes at GF24 and GF36. The inclusion of recipient data improved AUC and longevity prediction; this also makes clinical sense and can likely be refined by adding additional data points in the future. Demonstrated improvements in long-term graft survival are associated with increased patient quality-adjusted life years, lower patient health costs, and increased value for all stakeholders.¹⁰ These are the types of decisions at the time of organ offer that can drive real value for the system as opposed to those that optimize only measures of short term success. Studies such as these are extremely timely as professional societies, regulatory agencies and others seek metrics and strategies to drive overall system performance. The survival observations that trend from the

RFDR GF36 (-) group extend past 250 months after transplant and are significantly better than the KDRI (-) group; this trend continues to grow over time. RFDR makes more successful predictions than KDRI and significantly more of the predicted DDKT recipient grafts survive longer in aggregate.

One limitation of our study was that we used the final KDRI model instead of recalculating the Cox regression. This limitation exists because we used the same coefficients as the final KDRI developed by Rao, et al. 2009, with these same data, instead of recalculating the coefficients for each validation fold. Ultimately, the test data used to evaluate KDRI were the very same data that were used to develop the KDRI originally. Alternatively, the MLM were evaluated correctly because the MLM were trained only on training data and tested only on held out testing data during the 10-fold cross-validation protocol. Despite the wrongly inflated performance of KDRI, the MLM still performed statistically significantly better.

These results are sufficient to justify future exploration of MLM and leveraging big data applied to the processes of organ transplantation. Despite a small statistically significant improvement in AUC, the margin of improvement by our MLM amounted to 2148 additional DDKT over 10 years (about 200 per year) that were correctly classified as successful transplants at the same (10% FNR) misclassification risk as KDRI. These results are clinically significant because the tools used by transplant teams will influence life-changing decisions. A clinical team with KDRI and an acceptable 10% graft failure rate at 36 months may be influenced to be more conservative for kidney offers that would have been successful. The same clinical team with MLM may be influenced to be more aggressive for the same kidney offers and capture an additional 26% of successful kidney transplants. Deploying MLM in a live clinical setting may be an avenue to increase the number of DDKT without sacrificing patient outcomes.

Our study design was purposely confined the same data and variables available for KDRI by Rao, et al. 2009, and did not allow us to leverage the full capabilities of MLM. Our future work will not have these design constraints and we hypothesize that with the benefit of the additional variables, more recent data, and missing data interpolation, the performance will be greater than what we have demonstrated here. Our future work intends to expand this research by pursuing more aggressive strategies for optimizing the predictive quality of MLM with the inclusion of additional data sources with the ultimate goal of providing real-time decision support to clinicians at the time of organ offer. If one of the stated goals of the OPTN is to increase the number of transplants, it has to start with clinicians being able to make the best use of available data at the point of organ offer. This, of course, will need to be coupled with other larger changes in system dynamics and policy to achieve overall success. MLM will allow us to use variables from the recipient, donor, transplant center (administrative, logistical-temporal turnaround data) and perhaps even behavioral data to predict the transplant outcomes with higher accuracy and clinically relevant predictive quality.

Abbreviations

AUC, area under the ROC curve, BSS, bootstrap sample, DDKT, deceased donor kidney transplantation, ESRD, end-stage renal disease, FN, false negative, FNR, false negative rate, FP, false positive, GF, graft failure, KDPI, kidney donor profile index, KDRI, kidney donor risk index, KM, Kaplan-Meier, MLM, machine learning methods, OPTN, organ procurement and transplantation network, RF, random forests, RFD, random forests with donor-only variables, RFDR, random forests with donor and recipient variables, ROC, receiver operating characteristic, STAR, standard transplant analysis and research, TN, true negative, TP, true positive, UNOS, united network of organ sharing.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The datasets analyzed during the current study are available in the UNOS STAR Files repository, [<https://optn.transplant.hrsa.gov/data/request-data/>].

Competing interests: Eric Pahl is the co-founder of OmniLife (formerly HealthTech Solutions), a company developing real-time software tools to assist clinicians in organ transplantation and procurement. Prof. Nick Street, Prof. Hans Johnson, and Dr. Alan Reed are advisors to OmniLife. The authors have no other vested interest, directly or indirectly, related to this work other than academic achievement.

Funding: Research reported in this publication was supported by the National Library of Medicine of the NIH, award number: R43LM012575. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Additionally, Eric Pahl received funding to support this research from a fellowship sponsored by the Interdisciplinary Graduate Program in Informatics at the University of Iowa.

Authors' contributions: All authors participated in the intellectual content, conception, and design of the paper, took public responsibility for it and have agreed to have their name listed as contributors. Eric Pahl, Hans Johnson, and Nick Street contributed to the research design, writing, performance of the research, analytic tools, and data analysis. Alan Reed contributed to the research design, writing, and analysis of this work.

Acknowledgements: Not applicable.

References

1. Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med informatics*. 2016;4(3):e28. doi:10.2196/medinform.5909

2. Barnes S, Hamrock E, Toerper M, Siddiqui S, Levin S. Real-time prediction of inpatient length of stay for discharge prioritization. *J Am Med Inform Assoc.* 2016;23(e1):e2-e10. doi:10.1093/jamia/ocv106
3. Kim SP, Gupta D, Israni AK, Kasiske BL. Accept/decline decision module for the liver simulated allocation model. *Health Care Manag Sci.* 2015;18:35-57. doi:10.1007/s10729-014-9295-x
4. Reinaldo F, Rahman MA, Alves CF, Malucelli A, Camacho R. Machine learning support for kidney transplantation decision making. In: *International Symposium on Biocomputing - ISB.* ; 2010. doi:10.1145/1722024.1722079
5. Yoon J, Zame WR, Banerjee A, Cadeiras M, Alaa AM, van der Schaar M. Personalized survival predictions via Trees of Predictors: An application to cardiac transplantation. Liu N, ed. *PLoS One.* 2018;13(3):e0194985. doi:10.1371/journal.pone.0194985
6. Hart A, Smith JM, Skeans MA, et al. OPTN/SRTR 2016 annual data report: Kidney. *Am J Transplant.* 2018;18:18-113. doi:10.1111/ajt.14557
7. Heilman RL, Mathur A, Smith ML, Kaplan B, Reddy KS. Increasing the use of kidneys from unconventional and high-risk deceased donors. *Am J Transplant.* 2016. doi:10.1111/ajt.13867
8. Wey A, Salkowski N, Kremers WK, et al. A kidney offer acceptance decision tool to inform the decision to accept an offer or wait for a better kidney. *Am J Transplant.* 2018;18(4):897-906. doi:10.1111/ajt.14506
9. Massie AB, Luo X, Chow EKH, Alejo JL, Desai NM, Segev DL. Survival benefit of primary deceased donor transplantation with high-KDPI kidneys. *Am J Transplant.* 2014;14:2310-2316. doi:10.1111/ajt.12830
10. Axelrod DA, Schnitzler MA, Xiao H, et al. An economic assessment of contemporary kidney transplant practice. *Am J Transplant.* 2018;18:1168-1176. doi:10.1111/ajt.14702
11. Reese PP, Harhay MN, Abt PL, Levine MH, Halpern SD. New solutions to reduce discard of kidneys donated for transplantation. *J Am Soc Nephrol.* 2016;27:973-980. doi:10.1681/ASN.2015010023
12. Mittal S, Adamusiak A, Horsfield C, et al. A re-evaluation of discarded deceased donor kidneys in the UK: Are usable organs still being discarded? *Transplantation.* 2017;101(7). doi:10.1097/TP.0000000000001542
13. Kasiske BL, McBride MA, Cornell DL, et al. Report of a consensus conference on transplant program quality and surveillance. *Am J Transplant.* 2012;12(8):1988-1996. doi:10.1111/j.1600-6143.2012.04130.x
14. Rao PS, Schaubel DE, Guidinger MK, et al. A comprehensive risk quantification score for deceased donor kidneys: The kidney donor risk index. *Transplantation.* 2009;88:231-236. doi:10.1097/TP.0b013e3181ac620b
15. Domingos P. A few useful things to know about machine learning. *Commun ACM.* 2012;55:78. doi:10.1145/2347736.2347755
16. Chi CL, Street WN, Ward MM. Building a hospital referral expert system with a prediction and optimization-based decision support system algorithm. *J Biomed Inform.* 2008;41:371-386. doi:10.1016/j.jbi.2007.10.002

17. Chi CL, Nick Street W, Robinson JG, Crawford MA. Individualized patient-centered lifestyle recommendations: An expert system for communicating patient specific cardiovascular risk information and prioritizing lifestyle options. *J Biomed Inform.* 2012;45:1164-1174. doi:10.1016/j.jbi.2012.07.011
18. Akusok A, Eirola E, Björk K-M, Miche Y, Johnson H, Lendasse A. Brute-force missing data extreme learning machine for predicting Huntington’s disease. In: *PErvasive Technologies Related to Assistive Environments-PETRA*. New York, New York, USA: ACM Press; 2017:189-192. doi:10.1145/3056540.3064945
19. Eirola E, Akusok A, Björk K-M, Johnson H, Lendasse A. Predicting Huntington’s disease: Extreme learning machine with missing values. In: *Extreme Learning Machines-ELM*. Springer, Cham; 2016:195-206. doi:10.1007/978-3-319-57421-9_16
20. Plis SM, Hjelm DR, Salakhutdinov R, et al. Deep learning for neuroimaging: a validation study. *Front Neurosci.* 2014;8:229. doi:10.3389/fnins.2014.00229
21. Breiman L. Random forests. *Mach Learn.* 2001;45:5-32. doi:10.1023/A:1010933404324
22. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. *Int Conf Mach Learn.* 2006;23:161-168. doi:10.1145/1143844.1143865
23. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int Jt Conf Artif Intell.* 1995;(March 2001).

Tables

Table 1.
Removed Observations

	KDRI	Observed
Initial Data, n_i	92102	91996
Pediatric Transplants	3733	3724
Previous Transplants	13122	12390
Multi-Organ Transplants	1556	1850
ABO Incompatible Transplants	211	176
Invalid/Missing Donor Height	2481	1009
Invalid/Missing Donor Weight	667	475
Invalid/Missing Donor Creatinine	892	949
Without “KDRI_RAO”	0	1181
Final Study, n_f	69440	70242

Table 1. Observed preprocessing of the UNOS STAR data compared to KDRI development data from kidney transplants 1995-2005.

Table 2.	Predicted GF	Actual GF	Description
True Positive* (TP)	Failed (1)	Failed (1)	Model Correctly Predicted a Bad Kidney (Avoided a bad kidney, bigger is better)
True Negative* (TN)	Not Failed (0)	Not Failed (0)	Model Correctly Predicted a Good Kidney (bigger is better)
False Positive# (FP, Type I error)	Failed (1)	Not Failed (0)	Model Error missed a Good Kidney (Missed Opportunity, smaller is better)
False Negative! (FN, Type II error)	Not Failed (0)	Failed (1)	Model Error missed a Bad Kidney (Graft Failure, smaller is better)

Table 2. Definitions of prediction outcomes when models were used to predict graft failure within the months following a kidney transplant. Superscript symbols were used to highlight good predictions (*), bad predictions leading to waste (#), and bad predictions leading to graft failure (!).

Table 3.

	FNR at 10%	KDRI	RFD	RFDR
GF = 36				
True Positive* (TP)	100%	100%	100%	
	(6661)	(6659)	(6658)	
True Negative* (TN)	100%	102%	126%	
	(8190)	(8344)	(10338)	
False Positive [#] (FP)	100%	100%	94%	
	(37313)	(37159)	(35165)	
False Negative [!] (FN)	100%	100%	100%	
	(740)	(742)	(743)	

(a)

FNR	KDRI TNR	RFDR TNR	Delta TNR	Delta TN
(%)	(%)	(%)	(%)	(count)
5.0	9.8	13.3	136	1595
7.5	14.0	18.3	131	1950
10.0**	18.0**	22.7**	126**	2148**
12.5	22.6	26.8	119	1942
15.0	26.4	30.5	116	1858
17.5	30.0	33.8	113	1711
20.0	33.5	37.4	112	1764
50.0	67.4	69.9	104	1134

(b)

Table 3 (a-b). (a) Comparison of predictions for GF36 with KDRI, RFD, and RFDR models held at 10% FNR cut off. Superscript symbols were used to highlight good predictions (*), bad predictions leading to waste (#), and bad predictions leading to graft failure (!). (b) Comparison of predictions for GF36 between KDRI and RFDR models at various FNR. Highlighting the 10% FNR cutoff (**).

Figures

Figure 1. Models' AUC for given GF

Figure 1

AUC for RF models created with different numbers of trees. KDRI model's AUC performance does not depend on the number of trees.

Figure 2.

Figure 2

(a-c). ROC curves of three predictive models for different graft failure criteria. P-values were calculated from DeLong's comparison method between ROC curves for RFD and RFDR respectively vs. KDRI are shown. The vertical axis, Sensitivity, at 0.9 is equivalent to the 10% FNR cut-off used for comparing the models in Table 3. The diagonal line represents a ROC curve with an AUC of 0.500.

Figure 3.

Figure 3

(a-f). Kaplan-Meier (KM) survival analysis comparison for KDRI and RF models using different graft failure criteria to split predicted survival groups for each model. Legends include the label for each line, the population size for that group at time zero, and the p-value calculated from the log-rank test between respective predicted outcome groups (i.e., - / +).