STUDY PROTOCOL

REVISED Using machine learning techniques to develop risk prediction models to predict graft failure following kidney transplantation: protocol for a retrospective cohort study [version 2; peer review: 2 approved, 1 approved with reservations]

Sameera Senanayake[1], Adrian Barnett[1], Nicholas Graves[1], Helen Healy[2,3], Keshwar Baboolal[2,3], Sanjeewa Kularatna[1]

[1]Australian Center for Health Service Innovation, Queensland University of Technology, Kelvin Grove, QLD, 4059, Australia
[2]Royal Brisbane Hospital for Women, Brisbane, QLD, 4001, Australia
[3]School of Medicine, University of Queensland, Brisbane, QLD, 4001, Australia

v2 **First published:** 29 Oct 2019, **8**:1810
https://doi.org/10.12688/f1000research.20661.1

Latest published: 09 Mar 2020, **8**:1810
https://doi.org/10.12688/f1000research.20661.2

Abstract

Background: A mechanism to predict graft failure before the actual kidney transplantation occurs is crucial to clinical management of chronic kidney disease patients. Several kidney graft outcome prediction models, developed using machine learning methods, are available in the literature. However, most of those models used small datasets and none of the machine learning-based prediction models available in the medical literature modelled time-to-event (survival) information, but instead used the binary outcome of failure or not. The objective of this study is to develop two separate machine learning-based predictive models to predict graft failure following live and deceased donor kidney transplant, using time-to-event data in a large national dataset from Australia.

Methods: The dataset provided by the Australia and New Zealand Dialysis and Transplant Registry will be used for the analysis. This retrospective dataset contains the cohort of patients who underwent a kidney transplant in Australia from January 1 st, 2007, to December 31 st, 2017. This included 3,758 live donor transplants and 7,365 deceased donor transplants. Three machine learning methods (survival tree, random survival forest and survival support vector machine) and one traditional regression method, Cox proportional regression, will be used to develop the two predictive models (for live donor and deceased donor transplants). The best predictive model will be selected based on the model's performance.

Discussion: This protocol describes the development of two separate machine learning-based predictive models to predict graft failure following live and deceased donor kidney transplant, using a large national dataset from Australia. Furthermore, these two models will be the most comprehensive kidney graft failure predictive models that have used survival data to model using machine learning techniques. Thus, these

Open Peer Review

Reviewer Status ✓ ? ✓

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 09 Mar 2020	✓ report	? report	✓ report
version 1 29 Oct 2019	? report		

1 **Ramkiran Gouripeddi**, University of Utah, Salt Lake City, USA

2 **Chenxi Huang**, Yale-New Haven Hospital, New Haven, USA

3 **Slade Matthews**, The University of Sydney, Sydney, Australia

Any reports and responses or comments on the article can be found at the end of the article.

models are expected to provide valuable insight into the complex interactions between graft failure and donor and recipient characteristics.

Keywords
Machine learning, Risk prediction models, Kidney transplant, Graft failure

Corresponding author: Sameera Senanayake (sameerajayan.senanayake@hdr.qut.edu.au)

Author roles: Senanayake S: Conceptualization, Data Curation, Formal Analysis, Writing – Original Draft Preparation; **Barnett A**: Conceptualization, Formal Analysis, Software, Supervision, Writing – Review & Editing; **Graves N**: Conceptualization, Formal Analysis, Supervision, Writing – Review & Editing; **Healy H**: Conceptualization, Formal Analysis, Supervision, Writing – Review & Editing; **Baboolal K**: Conceptualization, Formal Analysis, Supervision, Writing – Review & Editing; **Kularatna S**: Conceptualization, Formal Analysis, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Sameera Senanayake is a recipient of Australian Government Research Training Program (RTP) for Postgraduate Research (PhD) Scholarship and Queensland University of Technology International Postgraduate Research (PhD) Scholarship (2018 -2021).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Senanayake S, Barnett A, Graves N *et al.* **Using machine learning techniques to develop risk prediction models to predict graft failure following kidney transplantation: protocol for a retrospective cohort study [version 2; peer review: 2 approved, 1 approved with reservations]** F1000Research 2020, **8**:1810 https://doi.org/10.12688/f1000research.20661.2

First published: 29 Oct 2019, **8**:1810 https://doi.org/10.12688/f1000research.20661.1

REVISED **Amendments from Version 1**

The revised version of the manuscript includes changes as per the reviewers' recommendation. The rationale of using survival models and how it is different from conventional regression or classification-based machine learning methods have been included in this version. Additionally, as per the reviewer recommendation, a supplementary file has been included, indicating all the input variables that will be used as independent variables in the models.

Any further responses from the reviewers can be found at the end of the article

Introduction

The prevalence of chronic kidney disease is increasing globally. Along with this increment, the number of patients in end-stage of renal disease (ESRD) and the demand for kidney transplantation, along with other renal replacement therapies, have increased over recent years[1,2]. Compared with available renal replacement therapies, renal transplantation has dramatically improved the quality of life and the survival rate of patients with ESRD. However, evidence indicate that the health systems around the world, have failed to meet the increasing demand for kidney grafts. This is evident from the growing prevalence of ESRD in the world[3]. Further, kidney transplants pose a significant cost burden to health systems compared with other treatment modalities, as they consume a large amount of resources.

It is important that the donor kidneys are transplanted to the most suitable recipients in order to minimise the number of graft failures, and thus minimise the number of patients returning to the already-burdened waiting list[4]. However, evidence indicates that the incidence of graft failure following kidney transplantation has increased over the years, possibly owing to increased transplantation of kidneys from expanded-criteria donors and donors after cardiac death, who are more prone to graft failure[5]. Graft failure is associated with prolonged hospital stay and higher health system costs[6,7]. A mechanism to predict graft failure before the actual transplantation occurs is crucial in this regard. Similar predictive models have been increasingly used in the recent past, and these have assisted clinicians with evidence-based medical decision-making[4,8–10]. Numerous conventional predictive models are available in the literature to predict the graft loss among kidney transplant patients[11–14]. Interestingly, novel techniques based on machine learning methods provide the potential to produce more favourable results[15].

Machine learning techniques have been used to develop kidney graft outcome-prediction models[4,8–10,16]. With the exception of the prediction models developed in the United States[4,10,17–19], most of the other prediction models have been developed using datasets with less than 1,000 records. However, evidence indicates that large sample sizes lead to better prediction accuracy in machine learning-based prediction models[20]. The model developed by Akl *et al*. (2008)[8], using 1,900 live donor transplant records from a single urology centre in Egypt, is the only machine learning predictive model available that is based exclusively on live donor transplants, while most of the other models are based either exclusively on deceased donor transplants, or both deceased and living donor transplants. However, evidence indicates that the graft failure rate and the predictors of graft failure significantly differ between live and deceased donor transplants[5]. Therefore, from a clinical perspective, two separate valid and reliable prediction models, i.e. live and deceased donor transplants, would give superior clinical utility.

Time-to-event (survival) information had not been modelled in any of the machine learning based prediction models available in the medical literature. Instead, most have used the binary outcome of failure or not as the outcome variable. However, presence of censored observations makes predictions done using this type of prediction models less accurate. Therefore, incorporating the timing of the event to the prediction model, could lead to better prediction models[21].

In this background. the objective of this study is to develop two separate machine learning-based predictive models to predict graft failure following live and deceased donor kidney transplant, using time-to-event (survival) data in a large national dataset from Australia.

Protocol

This study will evaluate different machine learning methods in predicting kidney graft failure.

Study cohort

The dataset was provided by the Australia and New Zealand Dialysis and Transplant Registry (ANZDATA). ANZDATA collects and reports the incidence, prevalence and outcome of dialysis treatment and kidney transplantation for patients with end-stage kidney disease across Australia and New Zealand. The retrospective dataset contains the cohort of patients who underwent a kidney transplant in Australia from January 1st, 2007, to December 31st, 2017. This included 3,758 live donor transplants and 7,365 deceased donor transplants.

Two separate predictive models will be developed for live donor and deceased donor transplants using separate datasets for live and deceased donor transplants.

Outcome

Graft survival of the most recent kidney transplants will be converted to a binary variable and will be the primary outcome. Patients who died with a functioning graft will not be considered positive for graft failure, but will be included in all models and censored at their death date. The time to the graft failure will be calculated in days from the date of transplantation. If the outcome of interest has not happened within the time period the data is available (2007 to 2017), it will be considered as right-censored with a time from the date of transplantation to the censoring date. In total, n=65 (0.9%) patients in the deceased donor dataset (n=7,365) and n=73 (1.9%) patients in the living donor dataset (n=3,758) have been lost to follow-up. Their records will be right-censored from the last date where the follow-up information is available.

Independent variables

The data consist of de-identified recipient and donor characteristics of the transplants. In all, 83 possible variables were identified as potential risk factors for graft failure (Supplementary material)[22].

Model development

Three machine learning methods and one traditional regression method, Cox proportional regression, will be used to develop the two separate predictive models, i.e. one for live donor and one for deceased donor transplants. The machine learning methods that will be used are: survival tree[23], random survival forest[24] and survival support vector machine[25]. Thus, each prediction model will be developed using four methods, and the best predictive model will be selected based on the model's performance, as described in a later section.

Model development is a systematic process which involves five steps, as indicated in Figure 1.

Data preparation. Data preparation involves several steps, such as data cleaning, handling text and categorical attributes, and feature scaling.

Retrospective datasets have the inherent property of having missing values, and most machine learning algorithms do not work well with missing values. Depending on the extent the missing values in each of the variables, the decision will be made to either exclude the particular variable, categorise the missing values as a separate category, or use an imputation method to impute the missing values.

Machine learning algorithms work well with numerical arrays compared with text (e.g. the donor's cause of death: traumatic brain injury, cerebral infarct and intracranial haemorrhage). Thus, text variables will be categorised appropriately; each category will be assigned a numerical variable (e.g., traumatic brain injury will be assigned a '1', cerebral infarct a '2', and intracranial haemorrhage a '3'). However, when text variables are converted to numeric categories, the machine learning algorithms assume that two nearby values are more similar than two distant values (e.g. '1' is more similar to '2' than '3'). To overcome this, the categorical variables will be dummy coded into nominal categories.

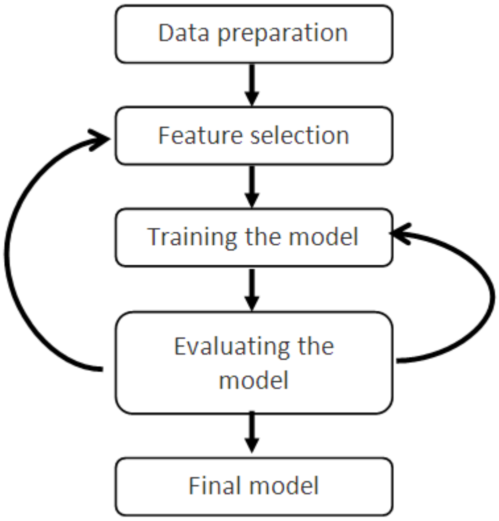

Figure 1. Model development process.

With a few exceptions, machine learning algorithms do not perform well when numerical input variables are not in the same scale (e.g., donor age ranges from 9 to 87, while the donor serum creatinine value ranges from 3 to 857). The numerical input variables will be standardised to convert them all to the same scale before applying the different machine learning methods[26]. Parameter estimates and plots will be transformed back to the original scale as this will be the most useful scale for clinicians.

Collinearity in the independent variables will be assessed using variance inflation factor (VIF)[27]. VIF of more than four will be considered as presence of collinearity and one of the correlated variables will be excluded from the analysis and the models re-run and re-validated.

Feature selection. Feature (variable) selection is the process of selecting the most relevant variables that should be included in the model. We will carefully select a potentially large set variables to be used by the feature selection methods in discussion with clinical colleagues. We will use variable selection methods that aim to create a parsimonious set of predictor variables from the larger set using cross-validation. We will reflect on the variables selected with our clinical colleagues to verify that the model makes clinical sense. Since the predictive models might potentially be used in pre-transplant decision making, only variables available before transplantation will be used in developing prediction models.

Three methods will be used for feature selection:

1. *Medical literature and expert opinion.* Studies done on kidney transplant graft survival will be reviewed to identify the significant predictors of graft survival. The identified list will be validated by clinical experts. The variables identified as potentially important in this step will be included in the predictive models.

2. *Principal component analysis*[28]. This method of feature selection will be performed using exploratory factor analysis using principal component analysis. Bartlett's test of sphericity and the Kaiser-Meyer-Olkin measure will be performed to assess the factorability of the data. Factor structure and factor loadings after varimax rotation will be assessed. The selection of factors will be done, depending on the eigenvalues. The factors are considered relevant if the eigenvalues are more than one. The input variables will be observed for their factor

coefficients, and more than 0.4 will be considered as well loaded and will be used for the model development. However, variables that have factor coefficients of less than 0.4 for all the factors will be excluded from the model development.

3. *Elastic net.* Elastic net uses both Lasso and Ridge regression to select features[29]. These two regularisation methods minimise the sum of squared residuals using L1 and L2 norms to limit the size of coefficients in the model[30]. The ideal size of the penalty will be selected using cross-validation. A stronger penalty is a tighter "lasso" that means fewer independent variables are selected. We will examine the plot of variable estimates against the penalty term to understand how the independent variables interact.

Possible combinations of the three sets of selected features from the three different feature selection mechanisms (i.e., medical literature and expert opinion, principal component analysis and elastic net) will be considered as input variables for the four methods of predictive models and the four methods of machine learning algorithms. Seven possible combinations are indicated in Table 1. The best set of input variables for each of the predictive models will be selected based on the model's performance.

Model training. During model training the dataset is randomly divided in to two parts: a training dataset and validation dataset. This prevents over-fitting and provides models that are more robust and give more realistic predictions of their prediction accuracy. Several spilt proportions have been used in models in relevant literature, such as 90:10% and 80:20%, with 70:30% being the most common[31]. Thus, in the present study dataset will be split in to two parts, with 70% of the data to train the model and 30% to validate the developed models. Given our large sample size we expect that this approach would produce similar results to multiple cross-validations. However, in live donor transplant sample of around 3,758 we will use use cross-validation to estimate the variability in our model evaluation statistics, and if the variability is large (more than 10% of the mean accuracy) then we will use cross-validation for this sample.

Table 1. Possible combinations of input variable groups.

Combination No.	Selected input variable group
Combination 1	ML & EO
Combination 2	PCA
Combination 3	EN
Combination 4	ML & EO and PCA
Combination 5	ML & EO and EN
Combination 6	PCA and EN
Combination 7	ML & EO, PCA and EN

ML & EO: Medical literature and expert opinion; PCA : Principal component analysis; EN : Elastic net

Since the outcome of interest is a survival function, the training dataset will be fitted to following models; Cox proportional regression method, survival tree, random survival forest and survival support vector machine. The R programming package, specifically the packages Survivalsvm, Ranger, Survival and LTRCtrees, will be used to develop all the predictive models[32].

1. Cox proportional regression[33]

   Cox proportional regression method is a semi-parametric model which is often used to explore the relationship between time-to-event data and several explanatory variables. This method assumes that effects of the different variables on survival are constant over time and are additive in a particular scale.

2. Survival Tree[23]

   A survival tree is a tree-like structure, where leaves represent outcome variables, i.e. graft failure (1) or no graft failure (0), and branches represent conjunctions of input variables that produced the outcome. Based on the chosen split criterion (survival statistic), a survival tree divides the data (parent node) into two groups (child nodes). The two resulting groups become the new parent nodes and are subsequently divided further into two child nodes based on the characteristics of the input variables.

   Hyper-parameters will be regularised until the optimal decision tree is created. The hyper-parameters include maximum depth of the decision tree, minimum number of samples a node must have before it can be split, and the minimum number of samples a node must have.

   Trees are often useful for identifying important interactions between independent variables. If strong interactions are found by the tree, then these may be added as additional independent variables to the other approaches as this could increase the models' predictive ability.

3. Random survival forest[24]

   Random forest is an ensemble method in machine learning where multiple unpruned survival trees are constructed via bootstrap aggregation[34,35]. Each tree predicts a classification independently and the final prediction is made based on the class (i.e. graft failure versus no graft failure) that gets the most "votes"[36,37]. This method of aggregation of multiple survival trees has several advantages: the prediction is resistant to outliers, less noisy and suitable for small datasets[38].

   The following hyper-parameters will be regularised until the optimal prediction is made: number of survival tree classifiers and maximum number of nodes.

4. Survival support vector machine[25]

   Survival support vector machine is a well-suited method to classify complex but small or medium-sized datasets.

Survival support vector machine uses hyperplanes to classify different classes and achieves high predictive accuracy when the data is linearly separable (linear kernel function). However, kernel functions (i.e. Gaussian, sigmoid and polynomial) can be used to separate even the non-linearly separable data, linearly[39,40].

Initially, the algorithm will be applied using a linear kernel function, and model performance will be assessed using other kernel functions (i.e. Gaussian, sigmoid and polynomial). Depending on the model's performance, the best kernel function will be selected. Depending on the kernel function selected, the following hyper-parameters will be regularised until the optimal prediction is made: 'C' value, Gamma value, degree and coefficient.

Evaluating the model. Performance of each model will be evaluated using model diagnostics, and the best model will be recommended depending on the results. The trained models, as described in the previous step, will be applied to the validation dataset (30% of the data). The prediction performance of each of the models will be assessed using three methods:

1. Concordance index[41]. The concordance index, or C-index, measures the discriminative ability of a survival model. It is defined as the fraction of pairs of patients that the patient who has a longer survival time is also predicted with lower risk score by the model. The range of concordance is between zero and one, with a larger value indicating better performance (and 0.5 indicating discrimination by chance).
2. Discriminative ability using the C-statistics for the censored function. This is the area under the receiver operating characteristics curve (ROC). The ROC curve is plotted with a sensitivity against 1–specificity, where sensitivity is on the y-axis and 1–specificity on the x-axis. The AUC ranges from 0 to 1, and a higher AUC indicates that the model is capable of distinguishing the cases (i.e. graft failure) with non-case (i.e. no graft failure).

The best performing model will be selected based on the results of the above-mentioned indicators. In an event of a discrepancy between the performance indicators, the results of concordance index will be considered as the main evaluator. We will also use other model checks such as residuals plots and testing for influential values, which may help to guide decision making about the "best" model.

Furthermore, the outputs of different machine leaning predictive models will be compared with Kidney Donor Risk Index (KDPI), a commonly used index which quantify graft failure risk before transplantation.

Data that will be used to develop the predictive models will be made available under restricted access with the permission from ANZDATA.

Ethics

Activities of the ANZDATA registry have been granted full ethics approval by the Royal Adelaide Hospital Human Research Ethics Committee (reference number: HREC/17/RAH/408 R20170927, approval date: 28/11/2017). Even though the data is at the individual-level in the registry, only de-identified records are requested for the analysis. All electronic data will be saved with password protection on Queensland University of Technology's secure server in encrypted folders only accessible to the nominated research staff.

Discussion

This protocol describes the development of two separate machine learning-based predictive models to predict graft failure following live and deceased donor kidney transplant, using a large national dataset from Australia. The live donor risk prediction model will be the first machine learning based predictive model developed using a large national dataset, and the deceased donor risk prediction model will be the only machine learning based predictive model that used more than 7,000 patient records outside the United States. Furthermore, these two models will be the only two predictive models which used post kidney transplant graft survival data to model using machine learning techniques. Thus, the two predictive models are expected to provide valuable insight into the complex interactions between graft failure and donor and recipient characteristics.

The dataset necessary for the study was provided by ANZDATA. ANZDATA collects and reports the incidence, prevalence and outcome of dialysis treatment and kidney transplantation for patients with end-stage kidney disease across Australia and New Zealand. This registry started in 1977 and since then all the kidney transplant activities in Australia and New Zealand have been captured in the registry, including the transplants in the private sector. The inclusion of all kidney transplants in Australia and New Zealand, and the availability and longevity of follow-up information have been the strength of this registry[42]. Thus, this registry has been the source of information for numerous high-impact publications[43–45].

The current study proposes to use four methods, namely: Cox proportional regression, survival support vector machine, random survival forest and survival tree. The best machine learning technique available to develop a predictive model is currently being discussed widely[46]. Most are of the view that no single technique fits all datasets, and it depends on the complexity of the data[47]. Thus, it is imperative that different machine learning methods are used to develop predictive models on a single dataset, so that the best could be chosen using validation parameters.

This project will have some limitations. According to medical literature, there are an abundance of risk factors for graft failure following a kidney transplant. However, the proposed predictive models will be only based on the variables which have already collected by ANZDATA, thus a complete risk factor profile may not be captured. The graft failure is linked to genetic[48] and socio-economic factors[49] of the transplant population. Thus, generalisability of the proposed models to other settings outside Australia, needs to be assessed further after they have been developed.

Data availability

Underlying data

There were no underlying data associated with this article.

Extended data

Supplementary material : Independent variables that will be used in the models (donor and recipient characteristics)[22] https://doi.org/10.6084/m9.figshare.11923446.v1

Acknowledgements

We acknowledge the Australia and New Zealand Dialysis and Transplant Registry and for their support in providing data for this study. We also acknowledge all the renal units and patients across Australia who contribute data to the registry.

References

1. Wang T, Xi Y, Lubwawa RN, *et al.*: **Chronic Kidney Disease (CKD) in U.S. Adults with Self-Reported Cardiovascular Disease (CVD)—A National Estimate of Prevalence by KDIGO 2012 Classification.** *Am Diabetes Assoc.* 2018; **67**(Supplement 1).
Publisher Full Text
2. Valley TS, Nallamothu BK, Heung M, *et al.*: **Hospital Variation in Renal Replacement Therapy for Sepsis in the United States.** *Crit Care Med.* 2018; **46**(2): e158–e65.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Barsoum RS: **Chronic kidney disease in the developing world.** *N Engl J Med.* 2006; **354**(10): 997–9.
PubMed Abstract | Publisher Full Text
4. Brown TS, Elster EA, Stevens K, *et al.*: **Bayesian modeling of pretransplant variables accurately predicts kidney graft survival.** *Am J Nephrol.* 2012; **36**(6): 561–9.
PubMed Abstract | Publisher Full Text
5. Decruyenaere A, Decruyenaere P, Peeters P, *et al.*: **Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods.** *BMC Med Inform Decis mak.* 2015; **15**: 83.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Matas AJ, Gillingham KJ, Elick BA, *et al.*: **Risk factors for prolonged hospitalization after kidney transplants.** *Clin Transplant.* 1997; **11**(4): 259–64.
PubMed Abstract
7. Rosenthal J, Danovitch GM, Wilkinson A, *et al.*: **The high cost of delayed graft function in cadaveric renal transplantation.** *Transplantation.* 1991; **51**(5): 1115–8.
PubMed Abstract
8. Akl A, Ismail AM, Ghoneim M: **Prediction of graft survival of living-donor kidney transplantation: nomograms or artificial neural networks?** *Transplantation.* 2008; **86**(10): 1401–6.
PubMed Abstract | Publisher Full Text
9. Greco R, Papalia T, Lofaro D, *et al.*: **Decisional trees in renal transplant follow-up.** *Transplant Proc.* 2010; **42**(4): 1134–6.
PubMed Abstract | Publisher Full Text
10. Lin RS, Horn SD, Hurdle JF, *et al.*: **Single and multiple time-point prediction models in kidney transplant outcomes.** *J Biomed Inform.* 2008; **41**(6): 944–52.
PubMed Abstract | Publisher Full Text
11. Moore J, He X, Shabir S, *et al.*: **Development and evaluation of a composite risk score to predict kidney transplant failure.** *Am J Kidney Dis.* 2011; **57**(5): 744–51.
PubMed Abstract | Publisher Full Text
12. Foucher Y, Daguin P, Akl A, *et al.*: **A clinical scoring system highly predictive of long-term kidney graft survival.** *Kidney Int.* 2010; **78**(12): 1288–94.
PubMed Abstract | Publisher Full Text
13. Tiong HY, Goldfarb DA, Kattan MW, *et al.*: **Nomograms for predicting graft function and survival in living donor kidney transplantation based on the UNOS Registry.** *J Urol.* 2009; **181**(3): 1248–55.
PubMed Abstract | Publisher Full Text
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**(2): 231–6.
PubMed Abstract | Publisher Full Text
15. Kaplan B, Schold J: **Transplantation: neural networks for predicting graft survival.** *Nat Rev Nephrol.* 2009; **5**(4): 190–2.
PubMed Abstract | Publisher Full Text
16. Senanayake S, White N, Graves N, *et al.*: **Machine learning in predicting graft failure following kidney transplantation: A systematic review of published predictive models.** *Int J Med Inform.* 2019; **130**: 103957.
PubMed Abstract | Publisher Full Text
17. Topuz K, Zengul FD, Dag A, *et al.*: **Predicting graft survival among kidney transplant recipients: A Bayesian decision support model.** *Decision Support Systems.* 2018; **106**: 97–109.
Publisher Full Text
18. Krikov S, Khan A, Baird BC, *et al.*: **Predicting kidney transplant survival using tree-based modeling.** *ASAIO J.* 2007; **53**(5): 592–600.
PubMed Abstract | Publisher Full Text
19. Goldfarb-Rumyantzev AS, Scandling JD, Pappas L, *et al.*: **Prediction of 3-yr cadaveric graft survival based on pre-transplant variables in a large national dataset.** *Clin Transplant.* 2003; **17**(6): 485–97.
PubMed Abstract | Publisher Full Text
20. van der Ploeg T, Austin PC, Steyerberg EW: **Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints.** *BMC Med Res Methodol.* 2014; **14**: 137.
PubMed Abstract | Publisher Full Text | Free Full Text
21. 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**(1): 8904.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Senanayake S: **Data fields for the study.docx.** *figshare.* 2020 [cited 2020 March 3].
http://www.doi.org/10.6084/m9.figshare.11923446.v1
23. Gordon L, Olshen RA: **Tree-structured survival analysis.** *Cancer Treat Rep.* 1985; **69**(10): 1065–9.
PubMed Abstract
24. Ishwaran H, Kogalur UB, Blackstone EH, *et al.*: **Random survival forests.** *Ann Appl Stat.* 2008; **2**(3): 841–60.
Publisher Full Text
25. Fouodo CJ, König IR, Weihs C, *et al.*: **Support Vector Machines for Survival Analysis with R.** *R J.* 2018; **10**(1): 412–423.
Publisher Full Text
26. Cheadle C, Cho-Chung YS, Becker KG, *et al.*: **Application of z-score transformation to Affymetrix data.** *Appl Bioinformatics.* 2003; **2**(4): 209–17.
PubMed Abstract
27. Kim JH: **Multicollinearity and misleading statistical results.** *Korean J Anesthesiol.* 2019; **72**(6): 558–69.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Wold S, Esbensen K, Geladi P: **Principal component analysis.** *Chemometr Intell Lab Syst.* 1987; **2**(1–3): 37–52.
Publisher Full Text
29. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Stat Soc Series B Stat Methodol.* 2005; **67**(2): 301–20.
Publisher Full Text
30. Tibshirani R: **Regression shrinkage and selection via the lasso.** *J R Statist Soc B.* 1996; **58**(1): 267–88.
Reference Source
31. Wong NC, Lam C, Patterson L, *et al.*: **Use of machine learning to predict early biochemical recurrence after robot-assisted prostatectomy.** *BJU Int.* 2019; **123**(1): 51–57.
PubMed Abstract | Publisher Full Text
32. Core Team R: **R: A language and environment for statistical computing.** R Foundation for statistical computing. Vienna. 2013.
Reference Source
33. Fox J: **Cox proportional-hazards regression for survival data.** 2002; 2002.
Reference Source
34. Efron B, Tibshirani RJ: **An introduction to the bootstrap.** CRC press; 1994.
Reference Source
35. Breiman L: **Bagging predictors.** *Mach Learn.* 1996; **24**(2): 123–40.
Publisher Full Text

36. Podgorelec V, Kokol P, Stiglic B, *et al.*: **Decision trees: an overview and their use in medicine.** *J Med Syst.* 2002; **26**(5): 445–63.
PubMed Abstract | **Publisher Full Text**

37. Marshall RJ: **The use of classification and regression trees in clinical epidemiology.** *J Clin Epidemol.* 2001; **54**(6): 603–9.
PubMed Abstract | **Publisher Full Text**

38. Shaikhina T, Lowe D, Daga S: **Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation.** *Biomedical Signal Processing and Control.* 2019; **52**: 456–462.
Publisher Full Text

39. Hu X, Wong KK, Young GS, *et al.*: **Support vector machine multiparametric MRI identification of pseudoprogression from tumor recurrence in patients with resected glioblastoma.** *J Magn Reson Imaging.* 2011; **33**(2): 296–305.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

40. Zhao D, Liu H, Zheng YH, *et al.*: **A reliable method for colorectal cancer prediction based on feature selection and support vector machine.** *Med Biol Eng Comput.* 2019; **57**(4): 901–912.
PubMed Abstract | **Publisher Full Text**

41. Steck H, Krishnapuram B, Dehing-Oberije C, *et al.*: **On ranking in survival analysis: Bounds on the concordance index**. *Advances in neural information processing systems.* 2008.
Reference Source

42. McDonald SP, Russ GR: **Australian registries-ANZDATA and ANZOD.** *Transplant Rev (Orlando).* 2013; **27**(2): 46–9.
PubMed Abstract | **Publisher Full Text**

43. McDonald SP, Craig JC, Australian and New Zealand Paediatric Nephrology Association: **Long-term survival of children with end-stage renal disease.** *N Engl J Med.* 2004; **350**(26): 2654–62.
PubMed Abstract | **Publisher Full Text**

44. Brook NR, Gibbons N, Nicol DL, *et al.*: **Open and laparoscopic donor nephrectomy: activity and outcomes from all Australasian transplant centers.** *Transplantation.* 2010; **89**(12): 1482–8.
PubMed Abstract | **Publisher Full Text**

45. Vacher-Coponat H, McDonald S, Clayton P, *et al.*: **Inferior early posttransplant outcomes for recipients of right versus left deceased donor kidneys: an ANZDATA registry analysis.** *Am J Transplant.* 2013; **13**(2): 399–405.
PubMed Abstract | **Publisher Full Text**

46. Yousef AH: **Extracting software static defect models using data mining.** *Ain Shams Engineering Journal.* 2015; **6**(1): 133–44.
Publisher Full Text

47. Lorena AC, Garcia LP, Lehmann J, *et al.*: **How Complex is your classification problem? A survey on measuring classification complexity.** *arXiv preprint.* arXiv: 180803591. 2018.
Reference Source

48. Yanagawa B, Algarni KD, Singh SK, *et al.*: **Clinical, biochemical, and genetic predictors of coronary artery bypass graft failure.** *J Thorac Cardiovasc Surg.* 2014; **148**(2): 515–20.e2.
PubMed Abstract | **Publisher Full Text**

49. Molmenti EP, Alex A, Rosen L, *et al.*: **Recipient Criteria Predictive of Graft Failure in Kidney Transplantation.** *Int J Angiol.* 2016; **25**(1): 29–38.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✓ ? ✓

Version 2

Reviewer Report 04 May 2020

https://doi.org/10.5256/f1000research.25046.r62016

✓ **Slade Matthews**
Pharmacoinformatics Laboratory, Discipline of Pharmacology, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia

The authors present a clear account of their approach to creating a predictive model for failure of renal transplants. The statistical and machine learning descriptions are very clear and nicely laid out. I have a few comments for the authors below:

1. In the introduction you mention that previous studies lumped together data considering live and post-mortem renal donor transplants. And that you anticipate there to be such a large difference between these two conditions that you will build two separate models for these two situations. The point is made by reviewer Chenxi Huang that you could use all the data and have an indicator variable which shows whether living or deceased donors were the organ source. Is it that the variable will be so predictive that you fear it will swamp otherwise important sources of variability in the model? Either way it is actually a testable hypothesis since you could build three models, the last one being a combined model and then answer the questions as to the deleterious contribution of this data.

2. The list of independent variables of which there are 83 includes many that would make prediction easy such as “graft failure date” and some that could be not directly related to the outcome but might give a little too much away like “age at death”. Clearly a careful consideration of what these variables mean must be had when selecting the variables allowed to comprise the model. The authors mention that “only variables available before transplantation will be used in developing prediction models” which is great and they then go on to describe feature selection methods ML & EO, PCA and EN. Then seven possible combinations are outlined in Table 1. Once these sets of descriptor variables are included in the 4 ML models will further variable selection take place such as selecting variables with significant beta coefficients in Cox Proportional Hazards Regression?

3. The authors say that the model performance evaluation will use 2 methods, the concordance index and the C-statistic. The concordance index looks at pairs of patients at a time. It compares their risk of death given by the model with the survival time recorded in the clinical data. Is this correct? So the model output is a risk of death value? I found this section a tiny bit unclear and it could benefit from an explicit description of the anticipated numerical output from the model. The 2nd

method uses the C-statistic or concordance with the ROC curve. Here the risk of graft failure predicted by the models will be subjected to a series of thresholds and sensitivity is calculated as the probability that the predicted risk value will be above the threshold for individuals for whom the graft failed. If I am correct here then I think that a sentence like this should be included because as it stands the description is somewhat terse like a man page in unix…

4. Validation of predictive machine learning models in medicine has long been a problem with the lack of clinical data availability in this setting. Have you have considered contacting the researchers in USA you mention to see if they would be willing to run their clinical data through your completed model as an external validation set. This would really add a lot of impact to the final reporting of the model performance.

5. Finally, on the whole I think this is a terrific project and I anticipate that it will achieve some really useful results. Your approach is very well thought out and with the inclusion of a couple of additional explanations more people will be able to benefit from reading this work and be inspired to adopt similar approaches in their own research.

Is the rationale for, and objectives of, the study clearly described?
Yes

Is the study design appropriate for the research question?
Yes

Are sufficient details of the methods provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am a computational toxicologist and I have experience in machine learning applied to biomedical data and survival analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 21 April 2020

https://doi.org/10.5256/f1000research.25046.r62013

? **Chenxi Huang**
Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA

The authors aimed to develop risk prediction models for graft failure following kidney transplantation using machine learning techniques. They have provided a lot of details on how the data will be prepared, what modeling techniques will be implemented and how the derived models will be compared. I have the following suggestions:

1. The major question I have is that the authors proposed to develop two separate models, one for live donor and one for deceased. Why couldn't they consider adding an indicator variable for the source of donor and build a more general model for use? Any interactions that the authors argued any single model would miss can be captured in this way. Such a model can benefit from a larger sample size and preventing overfitting. And the authors can test their hypothesis that such interaction exists.

2. The authors proposed to use a single random split with a backup plan for cross-validation. The authors should first specify what kind of cross-validation they are considering: 5-fold, 10-fold? Second, is it possible to do non-random splitting, such as splitting based on region or time of transplantation? Random data splitting often lead to failure of identifying overfitting and unrealistically optimal results.

3. The authors proposed to implement several feature selection methods. It is not clear whether they will be implemented with all modeling techniques and how many final models will be derived and compared. Also the last two feature selection methods are based on linear modeling, which does not take into account of potential interactions and linearity that the tree-based methods may be able to capture. Thus variables that may be important for subgroups may not be identified with these methods. I would suggest the authors to consider permutation methods paired with the specific machine learning modeling technique.

4. For comparing models, the summary statistics the authors proposed to use can sometimes be uninformative. To get a more complete picture of the model performance in discrimination, suggest the authors to use decision-curve analysis or simply visualizing AUC curve at different decision points.

5. As an optional point, the authors did not consider any modeling techniques that involves neural networks. The small sample size may be why. But to be more complete in machine learning modeling techniques, could consider small networks to see if they provide any improvement.

Is the rationale for, and objectives of, the study clearly described?

Yes

Is the study design appropriate for the research question?

Partly

Are sufficient details of the methods provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: biostatistics, machine learning, image analysis and processing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 14 April 2020

https://doi.org/10.5256/f1000research.25046.r61037

✓ **Ramkiran Gouripeddi**

Department of Biomedical Informatics, Center for Clinical and Translational Sciences (CCTS) Biomedical Informatics Core, University of Utah, Salt Lake City, UT, USA

I have no further comments to make.

Is the rationale for, and objectives of, the study clearly described?
Not applicable

Is the study design appropriate for the research question?
Not applicable

Are sufficient details of the methods provided to allow replication by others?
Not applicable

Are the datasets clearly presented in a useable and accessible format?
Not applicable

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomedical Informatics, Translational Research Informatics, Machine Learning, Data Integration, Exposome Informatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 12 December 2019

https://doi.org/10.5256/f1000research.22723.r55930

? **Ramkiran Gouripeddi**
Department of Biomedical Informatics, Center for Clinical and Translational Sciences (CCTS) Biomedical Informatics Core, University of Utah, Salt Lake City, UT, USA

Congratulations to the authors for submitting this manuscript. Here are some comments for your consideration:

1. The authors suggest using model time-to-event modeling instead of binary outcomes. Providing references to show that this is beneficial over binary outputs would be useful.

2. Modeling time-to-event would provide continuous output variables, and therefore require regression methods and not classification methods. It is not clear if the time-to-event is being model as binary variable, in which case the novelty is reduced, or the authors plan to use only regression methods. Please clarify.

3. Regarding “Three machine learning methods (survival tree, random survival forest and survival support vector machine) and one traditional regression method, Cox proportional regression, will be used to develop the two predictive models.”, is this one model live and one model for deceased donors? Changing this here and elsewhere to read as “one model live and one model for deceased donors, respectively” would make this clear.

4. “However, evidence indicate that the health systems around the world, have failed to meet the increasing demand for kidney grafts. This is evident from the growing prevalence of ESRD in the world3.” These two sentences do not seem related as suggested by the authors.

5. Listing or providing the 83 variables as a supplementary material will be helpful in reproducing methods.

6. Providing references for many of the methods mentioned (e.g. VIF) would also help.

7. “Even though having a large number of input variables may potentially produce superior results with high accuracy, in a practical sense it is important that a manageable number of input variables are used in the model.” This might not always be true. The goal is to avoid the curse of dimensionality and have appropriate number of independent variables.

8. Consider using cross-validation for model evaluation.

9. How is survival SVM different from any SVM?

10. The authors could consider evaluating the machine learning models with currently used methods in clinical settings as mentioned in the introductory remarks.

Is the rationale for, and objectives of, the study clearly described?

Partly

Is the study design appropriate for the research question?
Partly

Are sufficient details of the methods provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomedical Informatics, Translational Research Informatics, Machine Learning, Data Integration, Exposome Informatics,

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 02 Mar 2020

Sameera Senanayake, Queensland University of Technology, Kelvin Grove, Australia

We are grateful for the comments made by the reviewer. They have been extremely useful for us in reviewing our protocol. We have made changes in response to the suggestions of the reviewer wherever possible and have explained our responses in detail below.

Comment 1 : The authors suggest using model time-to-event modeling instead of binary outcomes. Providing references to show that this is beneficial over binary outputs would be useful.
Response 1 : The benefit of modelling time-to-event information instead of binary information, has been included to the introduction.
"Time-to-event (survival) information had not been modelled in any of the machine learning-based prediction models available in the medical literature, to predict graft failure after kidney transplant. Instead, most have used the binary outcome of failure or not as the outcome variable. However, presence of censored observations makes predictions done using this type of prediction models less accurate. Therefore, incorporating the timing of the event to the prediction model could lead to better prediction models[21]."

Comment 2 : Modeling time-to-event would provide continuous output variables, and therefore require regression methods and not classification methods. It is not clear if the time-to-event is being model as binary variable, in which case the novelty is reduced, or the authors plan to use only regression methods. Please clarify.
Response 2 : The survival models proposed in this paper will model time-to-event data and so consider two outcomes, i.e. time and graft failure (yes/no) [1,2]. This uses more information than a simple binary approach, which is statistically better than a simple binary outcome, and it uses more clinically relevant information, as patients and clinicians would like to know how long the graft will last.
[1] Bou-Hamad I, Larocque D, Ben-Ameur H. A review of survival trees. Statistics Surveys. 2011;5:44-71.
[2] Banerjee M, Reyes-Gastelum D, Haymart MR. Treatment-Free Survival in Patients With

Differentiated Thyroid Cancer. The Journal of Clinical Endocrinology & Metabolism. 2018 Jul;103(7):2720-7.

Comment 3 : Regarding “Three machine learning methods (survival tree, random survival forest and survival support vector machine) and one traditional regression method, Cox proportional regression, will be used to develop the two predictive models.”, is this one model live and one model for deceased donors? Changing this here and elsewhere to read as “one model live and one model for deceased donors, respectively” would make this clear.
Response 3 : This point is well noted. Changes have been made to indicate that two separate models, one for live donors and one for deceased donors, will be developed using different data sets

Comment 4 : “However, evidence indicates that the health systems around the world, have failed to meet the increasing demand for kidney grafts. This is evident from the growing prevalence of ESRD in the world3.” These two sentences do not seem related as suggested by the authors.

Response 4 : Please note that the final stage of chronic kidney disease is called the End Stage Renal Disease (ESRD) and patients in this stage need to either transplant a kidney or undergo dialysis to sustain life. Evidence indicates that the incidence (new patients) of ESRD is increasing around the world and if the supply of donor grafts increases accordingly, the prevalence (all patients) of ESRD is expected to reduce (as they will not be in ESRD anymore). However, the prevalence of ESRD has been increasing over recent years, indicating the demand for kidney grafts has not been met by the health systems.
Therefore, the authors believe that these 2 sentences complement each other.

Comment 5 : Listing or providing the 83 variables as a supplementary material will be helpful in reproducing methods.
Response 5 : Supplementary material indicating all possible variables has been added.

Comment 6 : Providing references for many of the methods mentioned (e.g. VIF) would also help.
Response 6 : Reference to the methods mentioned in the text has been added.

Comment 7 : Even though having a large number of input variables may potentially produce superior results with high accuracy, in a practical sense it is important that a manageable number of input variables are used in the model.” This might not always be true. The goal is to avoid the curse of dimensionality and have appropriate number of independent variables.
Response 7 : Thank you for highlighting this. The following changes have been made to the methods section.
“Feature (variable) selection is the process of selecting the most relevant variables that should be included in the model. We will carefully select a potentially large set of variables to be used by the feature selection methods in discussion with clinical colleagues. We will use variable selection methods that aim to create a parsimonious set of predictor variables from the larger set using cross-validation. We will reflect on the variables selected with our clinical colleagues to verify that the model makes clinical sense. Since the predictive models might potentially be used in pre-transplant decision making, only variables available before transplantation will be used in developing prediction models.”

Comment 8 : Consider using cross-validation for model evaluation.
Response 8 : Thank you for highlighting this. The following changes have been made to the

methods section.
"During model training, the dataset is randomly divided into two parts: a training dataset and validation dataset. This prevents over-fitting and provides models that are more robust and give more realistic predictions of their prediction accuracy. Several spilt proportions have been used in models in relevant literature, such as 90:10% and 80:20%, with 70:30% being the most common [30]. Thus, in the present study dataset will be split into two parts, with 70% of the data to train the model and 30% to validate the developed models. Given our large sample size we expect that this approach would produce similar results to multiple cross-validations. However, in live donor transplant sample of around 3,758 we will use cross-validation to estimate the variability in our model evaluation statistics, and if the variability is large (more than 10% of the mean accuracy) then we will use cross-validation for this sample."

Comment 9 : How is survival SVM different from any SVM?
Response 9 : Usual SVM can be used only for dichotomous outcomes and does not account to time-to-event information. However, survival SVM accounts for time-to-event data as it considers outcomes, i.e. time and event [1].
[1] *Fouodo CJ, König IR, Weihs C, Ziegler A, Wright MN. Support Vector Machines for Survival Analysis with R. R Journal. 2018 Jul 1;10(1).*

Comment 10 : The authors could consider evaluating the machine learning models with currently used methods in clinical settings as mentioned in the introductory remarks.
Response 10 : Thank you for highlighting this important point. The outputs of different machine learning predictive models will be compared with Kidney Donor Risk Index (KDPI), commonly used to quantify graft failure risk before transplantation.

Competing Interests: None