Machine-Learning Algorithms Predict Graft Failure Following Liver Transplantation

Authors

Lawrence Lau MBBS (Hons), FRACS*[1], Yamuna Kankanige BSc Eng*[2], Benjamin Rubinstein PhD[2], Robert Jones MBChB, FRACS[1], Christopher Christophi MD, FRACS[1], Vijayaragavan Muralidharan PhD, FRACS**[1], James Bailey PhD **[2]

1. Department of Surgery, Austin Hospital, Heidelberg, Melbourne, Australia

2. Department of Computing and Information Systems, University of Melbourne, Australia

*Joint First Authors

**Joint Last Authors

Address of correspondence

Dr Lawrence Lau

Department of Surgery, Austin Hospital

Heidelberg, Melbourne, Australia

thelau@gmail.com

Funding

The authors were supported in part by the Royal Australasian College of Surgeons Surgeon Scientist Research Scholarship, the Avant Doctor in Training Research Scholarship and the Australian Postgraduate Award

Disclosure

The authors declare no conflicts of interest

Author Contributions

Lawrence Lau – Research Design, Data Collection, Manuscript Writing (thelau@gmail.com)

Yamuna Kankanige – Research Design, Data Analysis, Manuscript Writing (ykankanige@student.unimelb.edu.au)

Benjamin Rubinstein – Research Design, Data Analysis, Manuscript Revision (benjamin.rubinstein@unimelb.edu.au)

Robert Jones – Research Design, Manuscript Revision (robert.jones@austin.org.au)

Christopher Christophi – Research Design, Manuscript Revision (cchri@unimelb.edu.au)

Vijayaragavan Muralidharan – Research Design, Manuscript Revision (v.muralidharan@unimelb.edu.au)

James Bailey – Research Design, Data Analysis, Manuscript Revision (baileyj@unimelb.edu.au)

Abbreviations

ALT, alanine aminotransferase

AUC-ROC, area under the receiver operating characteristic curve

BMI, body mass index

CI, confidence interval

CMV, cytomegalovirus

DRI, donor risk index

Hb, haemoglobin

HCC, hepatocellular carcinoma

HSV, herpes simplex virus

ICU, intensive care unit

INR, international normalised ratio

MELD, model for end-stage liver disease

SOFT, survival outcomes following liver transplantation

Abstract

Background

Ability to predict graft failure or primary non-function at liver transplant decision time assists utilization of scarce resource of donor livers, while ensuring that patients who are urgently requiring a liver transplant are prioritized. An index that is derived to predict graft failure using donor and recipient factors, based on local datasets, will be more beneficial in the Australian context.

Methods

Liver transplant data from the Austin Hospital, Melbourne, Australia, from 2010-2013 has been included in the study. The top 15 donor, recipient and transplant factors influencing the outcome of graft failure within 30 days, were selected using a machine learning methodology. An algorithm predicting the outcome of interest was developed using those factors.

Results

Donor Risk Index (DRI) predicts the outcome with an area under the receiver operating characteristic curve (AUC-ROC) value of 0.680 (95% CI 0.669-0.690). The combination of the factors used in DRI with the model for end-stage liver disease (MELD) score yields an AUC-ROC of 0.764 (95% CI 0.756–0.771), whereas Survival outcomes following liver transplantation (SOFT) score obtains an AUC-ROC of 0.638 (95% CI 0.632– 0.645). The top 15 donor and recipient characteristics within random forests results in an AUC-ROC of 0.818 (95% CI 0.812-0.824).

Conclusions

Using donor, transplant and recipient characteristics known at the decision time of a transplant, high accuracy in matching donors and recipients can be achieved, potentially providing assistance with clinical decision making.

Introduction

Outcome following liver transplantation depends upon a complex interaction between donor, recipient and process factors. Driven by the disparity between the increasing number of potential transplant recipients and the limited number of suitable organ donors, there is increasing use of organs of marginal quality[1,2]. This shift brings into focus, the delicate balance with organ allocation, between organ utility and the potential to cause harm to the recipient. Add to this the significant financial costs and regulatory pressures with each transplant, a quantitative tool which can help the transplant surgeon optimize this decision-making process is urgently required.

Surgeon intuition in the evaluation of donor risk is inconsistent and often inaccurate[3]. Scoring indices such as the DRI[4] attempts to quantify the quality of the donor liver based on donor characteristics but include factors which may not be applicable internationally (e.g. ethnicity and regional location of donor), and does not include factors which are known to be strong predictors of outcome but may not be consistently appraised (e.g. Hepatic steatosis). DRI has not found wide adoption into routine practice[5].

Beyond the assessment of donor organ quality, is the concept of donor-recipient matching[6], in order to maximize organ utilization while protecting patients from post-transplant complications. Risk scores that use both donor and recipient characteristics such as SOFT[7] score have been proposed for this purpose. Theoretically, the success of a transplant may be altered if a given donor organ were transplanted into different recipients. Unfortunately, aside from blood group matching and recipient urgency, currently there is little that guides this decision and the ideal donor-recipient matching algorithm[6] remains a long-term vision. Attempts to match donors to recipients based on recipient MELD score have had conflicting results[8,9].

Machine-learning algorithms can be used to predict the outcome of a new observation, based on a training dataset containing previous observations where the outcome is known. They can detect complex non-linear relationships between numerous variables and are used for predictive applications in a wide range of fields including agriculture, financial markets, search engines and match-making[10-13]. They are also finding increasing application in medicine[14]. A machine-learning algorithm, developed from the experience of a particular liver transplant unit, may be able to predict the likelihood of transplant success which is unit-specific and potentially allow for evolving practice.

The objective of this study is to evaluate the utility of machine-learning algorithms such as random forests and artificial neural networks, in order to predict outcome based on donor and recipient variables which are known prior to organ allocation. The performance of these algorithms will be compared against current standards of donor and recipient risk assessment such as DRI, MELD and SOFT score in predicting transplant outcome. This risk quantification tool may potentially assist donor-recipient matching, with improved balancing of the considerable risks associated with liver transplantation.

Materials and Methods

Study cohort

This study included the Liver Transplant Database from Austin Health, Melbourne, Australia, from January 1988 to October 2013. Austin Health is one of five state-based liver transplant units within Australia and serves the population in the States of Victoria and Tasmania. Brain-dead and cardiac death organ donors of whole liver and split liver transplants were included. Transplants involving paediatric recipients (under 18 years of age) and transplants from living-related donors were excluded from the study. Although transplant records are available from 1988, due to the significant number of values not available in the records prior to 2010 (particularly with the factors used to calculate DRI), only transplants which occurred after January 1st 2010, were included for analysis. Transplants from November 2013 to May 2015 were used for validating the results. This research was approved by the Austin Health Human Research Ethics Committee (Project Number: LNR/14/Austin/368).

Dataset Collation

The prospectively maintained database contains comprehensive information about each transplant including donor factors, transplant factors, recipient factors as well as recipient outcomes. The database was collated into the working dataset, with all fields arranged into categorical, ordinal or continuous variables.

Model Development

Well-known machine learning techniques such as random forests[15,16], artificial neural networks and logistic regression were employed for model development[17]. However, logistic regression was not used for models with many factors due to its comparatively poor performance during initial testing.

Training and test datasets were created by bootstrap sampling with replacement. In brief, an equivalent number of cases from the original dataset were randomly selected with duplicates to create a sample training set. It has been shown in literature that such a bootstrap sample will contain about 63% unique cases from the original dataset[18]. The remaining transplants, not included in the training set were allocated as the corresponding test set. This methodology known as out-of-bag error estimation, ensures that there will be no overlaps between the training and test sets[18], and is similar to the leave-one-out bootstrap technique for estimating prediction error[19]. This process was then repeated 1000 times to yield a set of 1000 training and corresponding testing datasets. Performances of all the algorithms were evaluated by the average of AUC-ROC values for the corresponding 1000 testing samples. Random forest and artificial neural network implementations in Weka data mining software were used for the experiments (Refer Appendix 2 for further information).

First, random forest algorithms and artificial neural networks were trained using all available characteristics for the 1000 bootstrapped samples.

Next all the characteristics were ranked per training sample using AUC-ROC based characteristic ranking method, which is suitable for datasets with high number of factors, missing values and imbalanced class sizes[20,21]. The implementation on "party package" for R statistical software[22] was used for this task. By scoring the characteristics according to their importance per each sample, over the 1000 samples, we determined the overall ranks of the characteristics for our training data.

As the next step, the top 15 factors for each sample were trained and evaluated using the random forests and artificial neural networks. Fifteen was chosen as the number of factors to be considered based on clinical utility. When training random forests, the following standard parameters were used[23]: 5000 as the number of trees, the square root of the number of available factors as the number of randomly selected factors considered at each decision point. Two hidden layers were used when training artificial neural networks.

Random forests and artificial neural networks with the overall top 15 ranked characteristics were employed to determine the performance with the validation data.

Outcome Parameters

The primary outcome parameter used, to develop and evaluate the prediction model was graft failure or primary non-function, as defined by death or re-transplantation, within 30 days of the transplant. As a secondary outcome parameter, the performance of the developed model to predict graft failure at 3 months was evaluated, using a separate validation dataset.

Donor Risk Index

As a comparative predictor of outcome, the DRI was calculated using the definition provided by Feng S et al.[4]. In the dataset, some factors required to calculate DRI for a particular donor may not have been recorded. DRI was considered as missing for that record, if any of the factors that are used in DRI were missing; age, cause of death (stroke, anoxia, trauma, other), whether the organ offer is after brain death or cardiac death, height, race (white, African American, other), donor hospital location (local, regional, national), cold ischemia time, partial/split liver. Actual cold ischemia time recorded was used in the calculations. Donor hospital location was assigned as follows: offers from hospitals in Melbourne metropolitan area as local, within Victoria state as regional, and others as national. Logistic regression was used to evaluate the performance of the samples with DRI.

DRI +/- MELD by Random Forest

The coefficients of the factors in DRI were derived in accordance to a Cox linear regression analysis of a large dataset from the United States[4]. It is possible that if the coefficients were recalculated or used to develop a non-linear model, the factors considered in DRI may be more specific to the local Australian context. Therefore, a random forest algorithm was developed using the DRI factors to assess their predictive capability.

A further random forest algorithm was developed using the factors required to calculate the DRI and the MELD score. This was an attempt to consider both donor and recipient factors in their contribution to transplant outcome.

SOFT Score

We calculated SOFT score as another comparative predictor of the outcome concerned, using the definition provided by Rana A et al.[7]. Portal bleed 48 hours pre-transplant was removed from the formula due to its unavailability in the dataset. SOFT score was considered as missing for a record, if any of the 18 factors used for SOFT score calculations were missing. Due to the high number of missing values in SOFT score (56%), performance with SOFT score was evaluated using random forests.

Statistical Analysis

The predictive performance of all the models was assessed using AUC-ROC analysis, a measurement of the discriminative ability of the model which is especially suited for imbalanced class classification[24-26]. AUC-ROC values vary from 0 to 1, where > 0.9 is considered excellent discrimination, > 0.75 is considered good discrimination and 0.5 is equivalent to random guessing[24]. AUC-ROC values were computed for each of the 1000 sample training/testing datasets and 95% confidence intervals were determined.

Results

Dataset Characteristics

The final dataset had 180 transplants, including 16 retransplants, with 11 graft failures (6.1%) within 30 days. 276 available donor and recipient characteristics (95 dichotomous, 25 non-dichotomous, 156 numerical) were included for characteristic selection, where 32% of the values in the dataset were missing values. One hundred seventy-three (173) donor characteristics, including demographic, clinical and logistical information were included. The recipient characteristics used in the study included 103 demographic and pre-transplant clinical information. A summary of the donor and recipient demographic and clinical characteristics are shown in Table 1 and the full list of characteristics are given in the appendix.

Table 1: Summary of donor and recipient characteristics

Characteristics	*Average ± standard deviation (range) for numerical factors, % for nominal factors*	
Donor Factors	**Study dataset**	**Validation dataset**
Age	45.8 ± 16.8 (14-78)	45.4 ± 16.2 (14-78)
Gender Male	 52.8%	 53.3%

Female Not recorded	46.7% 0.5%	46.7% 0%
BMI	26.3 ± 4.5 (17.6-40.4)	26.9 ± 5.6 (16.8-54.5)
Number of organs from donor	2.5 ± 0.8 (1-4)	2.6 ± 0.9 (1-4)
Donor offer Donation after brain death Donation after cardiac death Not recorded	 91.1% 8.9% 0%	 91.1% 5.6% 3.3%
Ethnicity Caucasian Other Not recorded	 87.2% 8.3% 4.5%	 76.7% 7.8% 15.5%
Cause of death Stroke Anoxia Trauma Other Not recorded	 65% 16.1% 10.6% 7.8% 0.5%	 56.7% 22.2% 10% 8.9% 2.2%
Donor pancreas retrieved Yes No Not recorded	 36.7% 53.9% 9.4%	 27.8% 72.2% 0%
Smoking history Yes No Ex-smoker Not recorded	 56.1% 37.2% 5% 1.7%	 55.6% 27.8% 14.4% 2.2%
Insulin use Yes No Not recorded	 41.1% 40.6% 18.3%	 6.7% 21.1% 72.2%

Alcohol consumption No Yes (quantity unknown) Mild (<1/d) Mod (2-4/d, up to 14/w) Heavy (>4/d, >14/w) Not recorded	 19.4% 27.8% 33.3% 11.1% 6.7% 1.7%	 15.6% 25.5% 38.9% 1.1% 8.9% 10%
Bilirubin	13.4 ± 17.1 (2-166)	9.5 ± 6.2 (2-37)
Plasma sodium	144.3 ± 6.5 (128-164)	140.4 ± 4.2 (133-156)
Creatinine	86.8 ± 48.4 (26-392)	94.1 ± 47.4 (39-305)
ALT	77.7 ± 107.5 (5-733)	110.8 ± 166.9 (10-668)
Hb	116.7 ± 26.4 (60-183)	128.0 ± 23.5 (74-175)
Cold ischemia time	6.4 ± 2.0 (3-18.8)	6.5 ± 2.6 (0.9-20.3)
Cut down Whole Split	 95.6% 4.4%	 95.6% 4.4%
Recipient Factors		
Age at transplant	50.6 ± 11.6 (19.3-70.9)	53.5 ± 11.3 (20.8-66.8)
Gender Male Female	 66.1% 33.9%	 72.2% 27.8%
MELD score	18.2 ± 7.5 (6-43)	19.6 ± 8.6 (6-46)
Re-transplant No Yes	 91.1% 8.9%	 98.9% 1.1%
If HCC, number of tumours	1.4 ± 0.5 (1-3)	2.1 ± 1.1 (1-6)
Oesophageal varices < ¼ of lumen, not bandable	 31.1%	 25.5%

Large	25.6%	16.7%
Not present	17.2%	6.7%
Not recorded	26.1%	51.1%
Bilirubin	134.6 ± 172.0 (5-902)	94.9 ± 131.0 (4-682)
INR	1.6 ± 0.5 (1-3.8)	1.5 ± 0.4 (1-3.2)
Albumin	29.3 ± 6.4 (13-47)	30.1 ± 7.8 (16-44)
Portal vein patency		
Patent	78.9%	82.2%
Thrombosed	3.3%	4.5%
Partial Thrombosis	2.2%	6.7%
Patent transjugular transhepatic portosystemic shunt	1.7%	2.2%
Not recorded	13.9%	4.4%
Ethnicity		
Caucasian	55%	37.8%
Asian	7.8%	8.9%
Other	3.3%	3.3%
Not recorded	33.9%	50%
Primary diagnosis / Disease category		
Hepatitis C	22.8%	11.2%
Malignancy / Hepatoma	14.4%	37.8%
Primary Sclerosing Cholangitis	10.6%	14.4%
Alcoholic Cirrhosis	8.9%	6.7%
Other	8.9%	13.3%
Chronic Active Hepatitis	5.6%	1.1%
Metabolic Disease	4.4%	3.3%
Primary Biliary cirrhosis	4.4%	5.6%
Acute Hepatic Necrosis	3.9%	1.1%
Cirrhosis-Cryptogenic	3.9%	3.3%
Chronic Active Hepatitis B	2.8%	1.1%
Biliary Atresia		0%

Not recorded	0.5% 8.9%	1.1%

Algorithm Performances

The ranks of the factors were determined from the sample training datasets using random forest characteristic importance method and the overall top 15 predictive donor and recipient factors were selected.

These donor factors were: cause of death (stroke, anoxia, trauma, other), serum albumin level, donation after brain or cardiac death, the state in which the donor hospital is located, alcohol consumption (no, unknown quantity, <1, 2-4, >4 drinks per day), haemoglobin level, total protein level, insulin usage, age, previous surgery, whether pancreas was retrieved concurrently, and donor cytomegalovirus status.

The recipient factors were: disease category, medical status at activation (home, frequent hospital care, hospital bound, ICU, ventilated) and serum herpes simplex antibodies. Table 2 provides the ranking of overall top 15 factors with their percentages of missingness in the study and validation datasets. It is noteworthy that most of these top predictors have less missing percentages when compared with the average of 32%.

Table 2: Overall top 15 predictors with the percentage of missing values in the study data and validation data

Characteristic	*Average rank sum*	*Missing % in study data*	*Missing % in validation data*
Recipient disease category	1.619	8.89	1.11
Donor serum albumin level	18.836	8.89	36.67
Donor cause of death	20.420	0.56	2.22

Donation after brain or cardiac death	24.931	0	3.33
Donor haemoglobin level	30.375	16.67	45.56
Donor alcohol consumption	30.805	1.67	10
The state in which the donor hospital is located	31.373	0	0
Donor total protein level	32.441	18.89	41.11
Donor insulin usage (dichotomous)	35.011	18.33	72.22
Recipient medical status at activation	36.285	33.89	27.78
Donor pancreas retrieved (dichotomous)	38.166	9.44	0
Donor age	38.412	0	0
Serum herpes simplex antibodies	38.654	12.78	8.89
Donor previous surgery (dichotomous)	41.505	2.78	0
Donor cytomegalovirus (CMV) Status (dichotomous)	42.083	0	0

Without characteristic selection, neural networks had an average AUC-ROC of 0.734 (95% CI 0.729-0.739) while random forests achieved 0.787 (95% CI 0.782-0.793). By comparison, when using the top 15 factors of each sample for 30 day graft failure, the predictive ability had an average AUC-ROC value of 0.818 (95% CI 0.812-0.824) with random forests and 0.835 (95% CI 0.831-0.840) with neural networks.

The validation dataset contained 90 transplants with 3 graft failures within 3 months, which was selected as the outcome for validation due to the lack of graft failures within 30 days. When the performance of the final model with the overall top 15 factors, trained for graft failure at 30 days, was assessed in its prediction ability for graft failure at 3 months, random forests achieved an average AUC-ROC value of 0.715 (95% CI 0.705-0.724), whereas neural networks yielded 0.559 (95% CI 0.548-0.569).

DRI, SOFT score and DRI +/- MELD by Random Forest Performance

To compare, the DRI for each donor in our dataset was calculated with a mean value of 1.56 (± 0.37). DRI predicted graft failure within 30 days with an average AUC-ROC value of

0.680 (95% CI 0.669-0.690). Using DRI trained for graft failure at 30 days, to predict graft failure at 3 months for the validation dataset, the average AUC-ROC value was 0.595 (95% CI 0.587-0.602).

Using the same factors that are used in DRI, we developed a model using Random Forests. This model achieved an average AUC-ROC of 0.697(95% CI 0.688- 0.705). When MELD score were added to the DRI factors for Random Forest modelling, a predictive average AUC-ROC of 0.764 (95% CI 0.756 – 0.771) was observed.

The SOFT score was also assessed and had a mean value of 5.5 (± 4.3). As a predictor for 30 day graft failure, it had average AUC-ROC of 0.638 (95% CI 0.632 – 0.645).

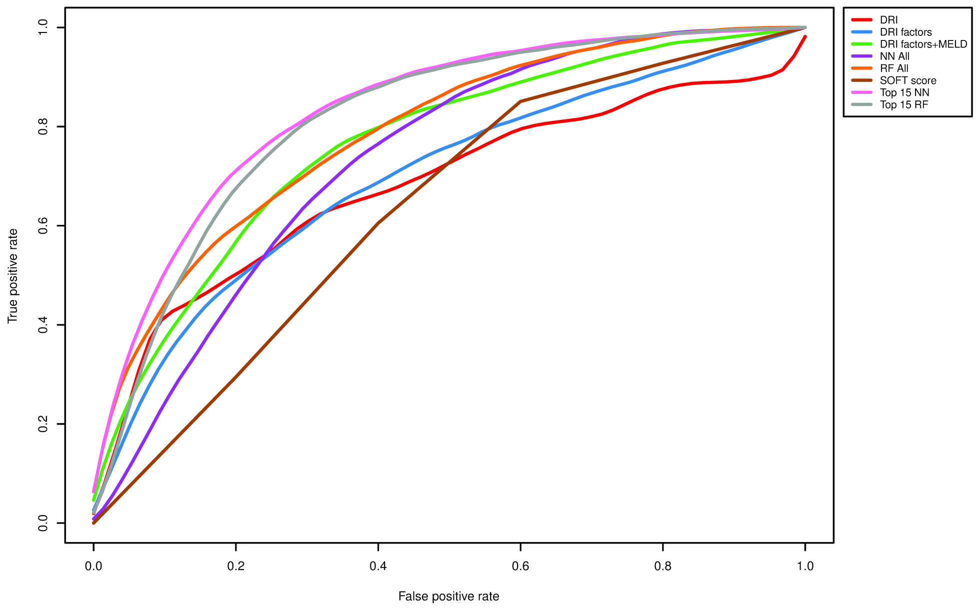
A comparison of all the results with the study dataset is given in Table 3 and Figure 1.

Table 3: Comparison of AUC-ROC values of different models created during the study

Characteristics used	*AUC-ROC (95% CI)*
Donor risk index	0.680 (0.669-0.690)
SOFT score	0.638 (0.632-0.645)
Neural network with all the factors	0.734 (0.729-0.739)
Random forest with all the factors	0.787 (0.782-0.793)
DRI characteristics in random forest	0.697 (0.688-0.705)
DRI characteristics and MELD score in random forest	0.764 (0.756-0.771)
Random forest with characteristic selection (Top 15)	0.818 (0.812-0.824)
Neural network with random forest characteristic selection (Top 15)	0.835 (0.831-0.840)

DRI – donor risk index factors: age, cause of death, race, partial/split, height, regionality, cold ischaemia time; MELD – model for end stage liver disease factors: recipient creatinine, bilirubin and INR; SOFT- Survival outcomes following liver transplantation score factors-age, BMI, number of previous transplants, previous abdominal surgery, albumin, dialysis prior to transplantation, intensive care unit pre-transplant, admitted to hospital pre-transplant, MELD, life support pre-transplant, encephalopathy, portal vein thrombosis, ascites pre-transplant, portal bleed 48 h pre-transplant, donor age, donor cause of death from cerebral vascular accident, donor creatinine, national allocation, cold ischemia time.

Figure 1: ROC curve comparison of different models created during the study



Discussion

This study is a proof-of-concept that machine-learning algorithms can be an invaluable tool, supporting the decision-making process for liver transplant organ allocation. This is particularly relevant in the current high-stakes environment where suboptimal organ utility leads to either increased waiting list mortality or patient mortality following transplantation.

The results of this study revealed that using 15 of the top-ranking donor and recipient variables available prior to transplantation were the best predictors of outcome with an average AUC-ROC of 0.818 with the random forest algorithm and 0.835 with artificial neural networks. Both machine learning techniques showed significant improvements in AUC-ROC with characteristic selection. This was followed by training the random forest classifier with the variables used to calculate DRI plus MELD score (AUC-ROC=0.764). Using the random forest classifier with the factors used to calculate DRI improved the discrimination of DRI from 0.680 to 0.697. SOFT score achieved an average AUC-ROC of 0.638. Assessing the predictive accuracy of the final models with top 15 factors, as trained for 30 day outcome, for graft failure at 3 months, the AUC-ROC value decreased from 0.818 to 0.715 with random forests and 0.835 to 0.559 with neural networks. By comparison, DRI prediction of 3 month graft failure was 0.595.

There are many machine-learning paradigms, of which two of the most widely used are artificial neural networks and random forest classifiers. In a recent landmark paper where the performance of 179 different machine-learning classifiers were used to classify all 121 datasets, representing the entire University of California Irvine Machine Learning Repository, random forest classifiers were found to be the most accurate[27]. There are four reports using artificial neural networks to predict transplant outcome in literature[28-31]. The present study is the first report using a random forest machine-learning algorithm for predicting outcome following liver transplantation.

There are multiple theoretical advantages with the use of random forest algorithms in this application. It is well known in machine learning literature that artificial neural networks are prone to overfitting and learning noise in data, resulting in unstable models with poor generalization ability[32-35]. However, by design, random forest classifiers are less prone to overfitting producing more stable models[36-38]. In medical datasets, there is frequently a large degree of missing data since the data is often not collected for research purposes, and some tests are not routinely performed even though they may be highly prognostic (e.g. donor liver biopsy for assessment of steatosis). Simply excluding these cases may bias the results due to the fact that the "missing-ness" of the data is not completely at random[39,40]. Random forest algorithms are superior in handling datasets missing a significant proportion of input data such as with this study[41]. Furthermore, while artificial neural networks are essentially, a "black-box" into which data is inputted and a prediction is outputted, the characteristic importance measure with random forest can indicate the importance of each variable in the dataset thereby improving the transparency of the algorithm[38,41,42].

Myriad factors interact to influence liver transplant including donor, recipient and locally specific transplant factors. There have been many attempts to predict graft failure, following liver transplant in literature[7,8,43-48]. Some studies looked at predicting graft failure using either donor factors, recipient factors[43], or a combination of both[7,8,45-48]. However, these approaches have all failed to gain greater adaptability because they are developed from patient populations which may not be generalizable to other centres due to regional differences in patient, donor or process factors, or changes in practice since their development[5,6]. Furthermore, they are calculated from simple multiple regression statistical models which assumes the linear influence of different variables. A predictive model required to enable effective organ allocation needs to be locally and temporally applicable, and account for the complex interactions within the data available prior to transplantation.

Currently, decisions for organ allocation are largely subjective or based on a recipient "sickest-first" or "waiting-time" approach rather than an outcome-based approach. Machine-learning algorithms are increasingly used for modern clinical decision-making. Compared to current methods, they are data driven, able to accommodate numerous interdependent variables and specific to the population from which they were trained on. In addition, compared with static indices, they are dynamic, able to "learn" case-by-case with the expansion of the training set.

Using characteristic importance measure, the most influential donor and recipient variables were determined. Most of these factors such as donor age, whether the offer is after brain death or cardiac death, donor cause of death, donor hospital State (geographical distance), donor alcohol consumption, recipient disease category and medical status at activation are already known as important factors[4,45,49,50]. Donor haemoglobin, protein level and insulin usage were also top-ranking predictive characteristics which make sense clinically. Donor CMV and recipient HSV status were also predictive and although less intuitive, has been shown to be associated with acute viral infection and rejection[51,52]. Interestingly, the decision to retrieve the pancreas for islet cell or whole organ transplant was also a top-ranking factor, although the decisions to retrieve kidneys, lungs or heart were not significant factors. This is likely because the decision for pancreas retrieval is usually more stringent, requiring more ideal donor conditions.

This study highlights the importance of characteristic selection and tailoring in predictive modelling. The predictive accuracy of the well-known DRI was improved when tailored to the specific influences at the Austin Health Liver Transplant Unit. Accuracy was further improved with the addition of recipient MELD characteristic with the best accuracy found with the application of a unit-specific Random Forest algorithm using the top-ranking predictive factors.

The main limitations of machine-learning algorithms are that they are best suited to predicting outcome in the environment from which they are derived. Conversely, this limitation is also its strength, in that it is highly specific to the peculiarities of a particular transplant centre, enabling the best decision for each individual transplant. Therefore, while it is not ideal to export a trained algorithm from one transplant centre to the next, certainly, the approach, with an algorithm tailored to each transplant centre is possible. A further limitation of this algorithm is that while it is trained to predict 30 day graft failure, its predictive accuracy may not extend to other important liver transplant outcomes such as 3, 6 or 12 month graft failure, early graft dysfunction, acute/chronic rejection, infections, immunosuppression or late biliary strictures. Each of these outcomes might require a separately trained algorithm.

A limitation of this study is that the machine-learning algorithm was derived from an observational database. While the bootstrapping with replacement methodology is well validated for the development of robust predictive machine-learning models[53,54], and our attempts to predict 3 month graft failure for a separate validation dataset looks promising, prospective validation for 30 day graft failure would be valuable to confirm the predictive ability .

This study confirms that machine-learning algorithms based on donor and recipient variables which are known prior to organ allocation can be utilized to predict transplant outcomes. This approach may be used as a tool for transplant surgeons to improve organ allocation decisions. The ability to quantify risk may allow for improved confidence with the use of marginal organs and better outcome following transplantation.

Acknowledgements

The authors gratefully acknowledge Angela Li, and the staff of the Liver Transplant Unit at Austin Hospital for their invaluable support for this study. This study was supported in part by the Royal Australasian College of Surgeons Surgeon Scientist Research Scholarship the Avant Doctor in Training Research Scholarship and the Australian Postgraduate Award.

References

1. Busuttil RW, Tanaka K. The utility of marginal donors in liver transplantation. *Liver transplantation.* 2003;9(7):651-663.
2. Tector AJ, Mangus RS, Chestovich P, et al. Use of extended criteria livers decreases wait time for liver transplantation without adversely impacting posttransplant survival. *Annals of surgery.* 2006;244(3):439-450.
3. Volk ML, Roney M, Merion RM. Systematic bias in surgeons' predictions of the donor ‐ specific risk of liver transplant graft failure. *Liver Transplantation.* 2013;19(9):987-990.
4. Feng S, Goodrich N, Bragg ‐ Gresham J, et al. Characteristics associated with liver graft failure: the concept of a donor risk index. *American Journal of Transplantation.* 2006;6(4):783-790.
5. Mataya L, Aronsohn A, Thistlethwaite JR, Friedman Ross L. Decision making in liver transplantation—limited application of the liver donor risk index. *Liver Transplantation.* 2014;20(7):831-837.
6. Briceño J, Ciria R, de la Mata M. Donor-recipient matching: myths and realities. *Journal of hepatology.* 2013;58(4):811-820.
7. Rana A, Hardy M, Halazun K, et al. Survival outcomes following liver transplantation (SOFT) score: a novel method to predict patient survival following liver transplantation. *American Journal of Transplantation.* 2008;8(12):2537-2546.
8. Halldorson J, Bakthavatsalam R, Fix O, Reyes J, Perkins J. D ‐ MELD, a Simple Predictor of Post Liver Transplant Mortality for Optimization of Donor/Recipient Matching. *American Journal of Transplantation.* 2009;9(2):318-326.

9. Croome K, Marotta P, Wall W, et al. Should a lower quality organ go to the least sick patient? Model for End-Stage Liver Disease score and donor risk index as predictors of early allograft dysfunction. Paper presented at: Transplantation proceedings2012.
10. Feyyad U. Data mining and knowledge discovery: making sense out of data. *IEEE expert.* 1996;11(5):20-25.
11. Kaur M, Gulati H, Kundra H. Data Mining in Agriculture on Crop Price Prediction: Techniques and Applications. *International Journal of Computer Applications.* 2014;99(12):1-3.
12. Joachims T. Optimizing search engines using clickthrough data. Paper presented at: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining2002.
13. Langley P. Machine learning for adaptive user interfaces. Paper presented at: KI-97: Advances in artificial intelligence1997.
14. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine.* 2001;23(1):89-109.
15. Breiman L. Random forests. *Machine learning.* 2001;45(1):5-32.
16. Liaw A, Wiener M. Classification and regression by randomForest. *R news.* 2002;2(3):18-22.
17. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics.* 2008;77(2):81-97.
18. Breiman L. *Out-of-bag estimation.* Statistics Department, University of California Berkeley, Berkeley CA 94708;1996.
19. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association.* 1983;78(382):316-331.
20. Janitza S, Strobl C, Boulesteix A-L. An AUC-based permutation variable importance measure for random forests. *BMC bioinformatics.* 2013;14(1):119.
21. Hapfelmeier A, Hothorn T, Ulm K, Strobl C. A new variable importance measure for random forests with missing data. *Statistics and Computing.* 2014;24(1):21-34.
22. Hothorn T, Hornik K, Strobl C, Zeileis A. Party: A laboratory for recursive partytioning. 2010.
23. Lunetta KL, Hayward L, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics.* 2004;5(1):1.
24. Ray P, Le Manach Y, Riou B, Houle TT. Statistical evaluation of a biomarker. *The Journal of the American Society of Anesthesiologists.* 2010;112(4):1023-1040.
25. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition.* 1997;30(7):1145-1159.
26. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.).* 2010;21(1):128.
27. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research.* 2014;15(1):3133-3181.
28. Dvorchik I, Subotin M, Marsh W, McMichael J, Fung J. Performance of multi-layer feedforward neural networks to predict liver transplantation outcome. *Methods of information in medicine.* 1996;35(1):12-18.
29. Matis S, Doyle H, Marino I, Mural R, Uberbacher E. Use of neural networks for prediction of graft failure following liver transplantation. Paper presented at:

Computer-Based Medical Systems, 1995., Proceedings of the Eighth IEEE Symposium on1995.

30. Briceño J, Cruz-Ramírez M, Prieto M, et al. Use of artificial intelligence as an innovative donor-recipient matching model for liver transplantation: results from a multicenter Spanish study. *Journal of hepatology.* 2014;61(5):1020-1028.
31. Cruz-Ramírez M, Hervás-Martínez C, Fernandez JC, Briceno J, De La Mata M. Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks. *Artificial intelligence in medicine.* 2013;58(1):37-49.
32. Cheng B, Titterington DM. Neural networks: A review from a statistical perspective. *Statistical science.* 1994:2-30.
33. Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment.* 1998;32(14):2627-2636.
34. Adya M, Collopy F. How effective are neural networks at forecasting and prediction? A review and evaluation. *J. Forecasting.* 1998;17:481-495.
35. Zhang GP. Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on.* 2000;30(4):451-462.
36. Anaissi A, Kennedy PJ, Goyal M, Catchpoole DR. A balanced iterative random forest for gene selection from microarray data. *BMC bioinformatics.* 2013;14(1):261.
37. Amaratunga D, Cabrera J, Lee Y-S. Enriched random forests. *Bioinformatics.* 2008;24(18):2010-2014.
38. Cutler DR, Edwards TC, Beard KH, et al. Random forests for classification in ecology. *Ecology.* 2007;88(11):2783-2792.
39. Acuna E, Rodriguez C. The treatment of missing values and its effect on classifier accuracy. *Classification, clustering, and data mining applications*: Springer; 2004:639-647.
40. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological methods.* 2002;7(2):147.
41. Pantanowitz A, Marwala T. Missing data imputation through the use of the Random Forest Algorithm. *Advances in Computational Intelligence*: Springer; 2009:53-62.
42. Ball RL, Tissot P, Zimmer B, Sterba-Boatwright B. Comparison of random forest, artificial neural network, and multi-linear regression: a water temperature prediction case. Paper presented at: Seventh Conference on Artificial Intelligence and its Applications to the Environmental Sciences. New Orleans, LA2009.
43. Desai NM, Mange KC, Crawford MD, et al. Predicting outcome after liver transplantation: utility of the model for end-stage liver disease and a newly derived discrimination function1. *Transplantation.* 2004;77(1):99-106.
44. Avolio A, Siciliano M, Barbarino R, et al. Donor risk index and organ patient index as predictors of graft survival after liver transplantation. Paper presented at: Transplantation proceedings2008.
45. Ioannou GN. Development and validation of a model predicting graft survival after liver transplantation. *Liver transplantation.* 2006;12(11):1594-1606.
46. Amin MG, Wolf MP, TenBrook JA, et al. Expanded criteria donor grafts for deceased donor liver transplantation under the MELD system: a decision analysis. *Liver transplantation.* 2004;10(12):1468-1475.

47. Avolio AW, Cillo U, Salizzoni M, et al. Balancing Donor and Recipient Risk Factors in Liver Transplantation: The Value of D‐MELD With Particular Reference to HCV Recipients. *American Journal of Transplantation.* 2011;11(12):2724-2736.
48. Dutkowski P, Oberkofler CE, Slankamenac K, et al. Are there better guidelines for allocation in liver transplantation?: A novel score targeting justice and utility in the model for end-stage liver disease era. *Annals of surgery.* 2011;254(5):745-754.
49. Mateo R, Cho Y, Singh G, et al. Risk factors for graft survival after liver transplantation from donation after cardiac death donors: an analysis of OPTN/UNOS data. *American journal of transplantation.* 2006;6(4):791-796.
50. Moore DE, Feurer ID, Speroff T, et al. Impact of donor, technical, and recipient risk factors on survival and quality of life after liver transplantation. *Archives of Surgery.* 2005;140(3):273-277.
51. Linares L, Sanclemente G, Cervera C, et al. Influence of cytomegalovirus disease in outcome of solid organ transplant patients. Paper presented at: Transplantation proceedings2011.
52. Pedersen M, Seetharam A. Infections after orthotopic liver transplantation. *Journal of clinical and experimental hepatology.* 2014;4(4):347-360.
53. Breiman L. *Out-of-bag estimation.* Citeseer;1996.
54. Austin PC, Tu JV. Bootstrap methods for developing predictive models. *The American Statistician.* 2004;58(2):131-137.