Interpretable Machine Learning in Kidney Offering: Multiple Outcome Prediction for Accepted Offers

Achille Salaün (achille.salaun@eng.ox.ac.uk)[1,*], **Simon Knight** (simon.knight@nds.ox.ac.uk)[2], **Laura Wingfield** (lrwingfield@gmail.com)[2], **and Tingting Zhu** (tingting.zhu@eng.ox.ac.uk)[1]

[1]Institute of Biomedical Engineering, Department of Engineering, University of Oxford, Oxford, OX3 7DQ, United-Kingdom
[2]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, OX3 9DU, United-Kingdom
[*]Corresponding author: achille.salaun@eng.ox.ac.uk

ABSTRACT

The decision to accept an organ offer for transplant, or wait for something potentially better in the future, can be challenging. Especially, clinical decision support tools predicting transplant outcomes are lacking. This project uses interpretable methods to predict both graft failure and patient death using data from previously accepted kidney transplant offers. Precisely, using more than twenty years of transplant outcome data, we train and compare several survival analysis and classification models in both single and multiple risk settings. In addition, we use *post hoc* interpretability techniques to clinically validate these models. In a single risk setting, neural networks provide comparable results to the Cox proportional hazard model, with 0.71 and 0.81 AUROC for predicting graft failure and patient death at year 10, respectively. Recipient and donor ages, primary renal disease, donor eGFR, donor type, and the number of mismatches at DR locus appear to be important features for transplant outcome prediction. We also extended the neural network approach to multiple outcome prediction, maintaining consistent performances and clinical interpretation. Thus, owing to their good predictive performance and the clinical relevance of their *post hoc* interpretation, neural networks represent a promising core component in the construction of future decision support systems for transplant offering.

Introduction

Around 2,500 deceased donor kidney transplants are performed in the UK each year. At any time, there are around 5,000 patients on the kidney transplant waiting list with an average wait of 2-3 years. The shortage of organs available for transplant means that some patients become unfit for surgery or die whilst waiting. Because of this, clinicians often consider organ offers from less-than-optimal donors with existing comorbidities or older age. Decisions around organ offers are made by clinicians based upon the information available at the time of offer, including donor and recipient demographic and medical details. Clinicians use their clinical experience, but do not have reliable tools available to help them predict what would happen if they choose to accept or decline an offer and wait for the next available one. This uncertainty leads to considerable variability in organ decline rates and waiting times between clinicians and centres. A computerised decision support (CDS) system that accurately predicts transplant outcomes, both in terms of graft failure and patient death, as well as indicating what would happen if the organ offer was declined (in terms of future offers and likely waiting time), may help to support clinicians in making these difficult decisions. As decisions must remain under the responsibility and control of the clinician, any CDS tool must be easy to use, and predictions must be interpretable from a clinician's perspective. Interpretability and usability are also important to patients, allowing better explanations of likely outcomes during the informed consent process.

The aim of this study is to predict transplant outcomes in the scenario of an accepted kidney offer. We rely on more than twenty years of registry data, containing over 36,000 accepted kidney transplant offers, with graft and patient survival information. These data have been provided by National Health Service Blood and Transplant (NHSBT) with ethical approval. Using these data, we have trained and compared several survival analysis and machine learning classification models, in both single and multiple-risk settings. In addition, we use *post hoc* interpretability techniques to clinically validate these models.

Predicting the time of occurrence of an event (such as patient death or graft failure) from censored data has been extensively

studied under the name of survival analysis. This has many applications in health informatics such as predicting strokes [1], oral cancer [2], or graft outcome prediction. Censored data are common in such contexts, resulting from loss of follow-up, competing events, or the end of the study. In the context of graft outcome prediction, [3–5] use the Cox proportional hazard (PH) model to predict kidney graft or recipient survival. The Cox PH model is a classic time-to-event approach that models the hazard function, as in the failure rate of a system according to time [6]. This approach is not only robust and reliable, but also simple to use and well understood by clinicians. Several generalisations of this model have been proposed. For instance, DeepSurv [7] aims at increasing the modelling power of the Cox model by replacing the linear contribution of the covariates with a neural network. Since the Cox model was originally designed to handle a single type of event, generalisations to multiple risks (e.g. predicting both graft and patient survival) have also been proposed. However, the effects of the regression coefficients on cause-specific survivability are not interpretable [8].

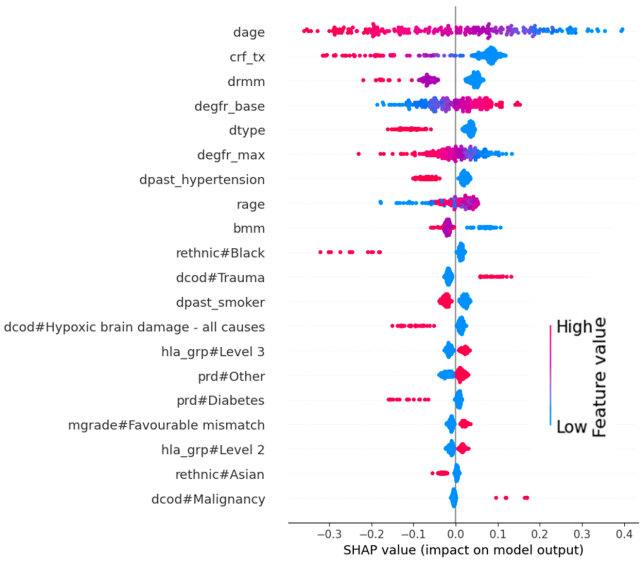
In general, one can distinguish two approaches for survival analysis: either providing a description of survivability over time, or predicting the state of the subject (e.g., graft, recipient) at arbitrary time points. While the Cox PH model follows the first approach, machine learning models can be used after converting the survival analysis problem into a classification problem at a given point in time. In the case of predicting transplant outcomes, predictions at specific milestones (e.g., graft and patient survival at years 1, 5, or 10) are generally sufficient: for example, existing risk communication tools such as [9] identify survival functions obtained from the Cox PH model at these time points. Many previous publications directly address this approach. For instance, [10] predicts kidney graft survival using tree-based models, [11] investigates several techniques, including random forests and neural networks. In [3], the neural network's ability to predict both graft and patient survival in kidney transplant is compared to one of the regression techniques (such as Cox). Predictions at different time points were either modelled independently or through multiple-output neural networks. [12] compares multilayer perceptrons and Bayesian networks, [13] uses Bayesian belief networks. Whilst many of these previous studies demonstrate acceptable predictive performance, none challenged their models' validity through the lens of clinical interpretability.

Interpretability is another important criterion in the construction of a CDS tool for predicting graft outcomes. Informally, interpretability is the extent to which the prediction of a model can be understood by a human [14]. This way, users can build trust regarding the model's results and remain in control of the associated outcomes. Moreover, a good model should always be *intrinsically* interpretable to a certain degree. Indeed, interpretable models have been shown to be more robust to adversarial attacks [15]. Unfortunately, this is not the case with the approaches mentioned above. Although the Cox PH model is interpretable in the single risk case, its generalisation to competing risks is not [8]. It is possible to interpret *a posteriori* a black box model with the help of *post hoc* interpretability methods. One can provide a local explanation of a given prediction. For instance, LIME [16] locally samples data points around the input and returns a linear explanation of the predictions made by the black-box model from these data points. Unfortunately, this solution is unstable; explanations depend highly on the sampled data points, harming the trustworthiness of the explanations. Similarly to LIME, SHAP [17] is a *post hoc* interpretability method relying on additive feature attribution models, i.e. linear functions as local explanation models. It provides explanations *via* game theory: each prediction is seen as a game where the features are players contributing to that game. Feature contributions are computed by considering all possible coalitions of features and the marginal contribution of each feature within these coalitions. Hence, SHAP can be considered as a gold standard in terms of *post hoc* interpretability methods.

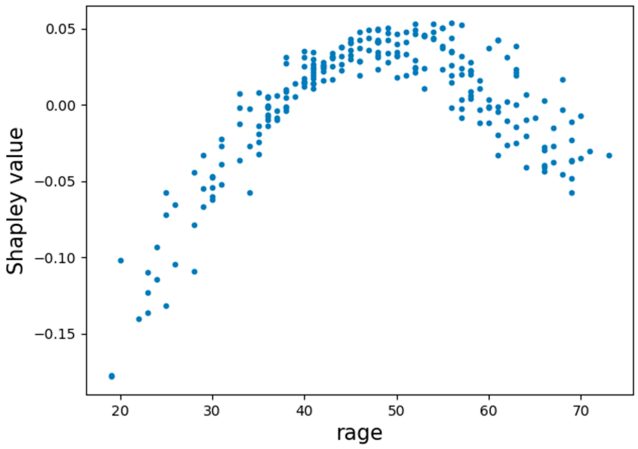
Results

Single Outcome Prediction

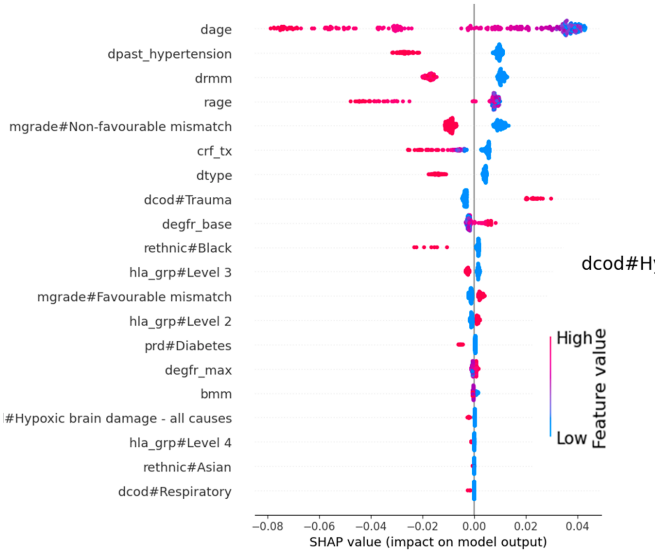
During the feature selection stage, we identified inconsistencies between SHAP interpretations of the obtained models and clinical expectations. For example, survivability is expected to decrease with the number of times a patient has been transplanted, which is not what we observed in the models produced. After further investigations, it appears that this feature is biased with all non *primo* recipients having a successful graft. The data set has indeed been built from various heterogeneous sources, with some outcomes not available for subsets of the data. From now on, we discard this feature. Finally, 15 and 10 features are



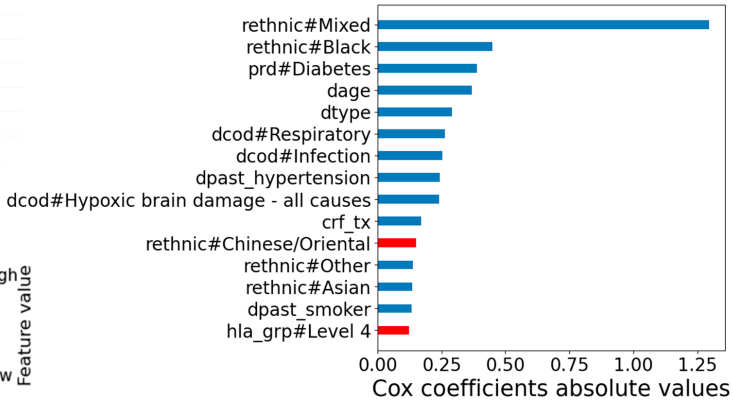
(a) Neural network's SHAP-based feature importance.
A negative SHAP value indicates a negative impact on graft survival.



(b) Neural network's dependence on recipient age (`rage`)



(c) Random forest SHAP-based feature importance.
A negative SHAP value indicates a negative impact on graft survival.



(d) Largest Cox PH model's coefficients.
Blue and red bars represent positive and negative coefficients, respectively.

Figure 1. Interpreting single outcome prediction models. Both models have been trained to predict graft failure at 10 years.

		Random forest		Cox PH model		Neural network	
		AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
Year 1	Without feature selection	**.62** ($\pm 2e^{-4}$)	.14 ($\pm 2e^{-6}$)	.61 ($\pm 1e^{-2}$)	.15 ($\pm 1e^{-5}$)	**.62** ($\pm 2e^{-4}$)	**.17** ($\pm 9e^{-5}$)
	With feature selection	.61 ($\pm 3e^{-4}$)	.14 ($\pm 4e^{-6}$)	.61 ($\pm 3e^{-4}$)	.15 ($\pm 1e^{-5}$)	**.62** ($\pm 3e^{-4}$)	**.18** ($\pm 7e^{-5}$)
Year 5	Without feature selection	.62 ($\pm 2e^{-4}$)	.34 ($\pm 7e^{-7}$)	**.64** ($\pm 2e^{-4}$)	**.37** ($\pm 5e^{-5}$)	**.64** ($\pm 2e^{-4}$)	**.37** ($\pm 2e^{-4}$)
	With feature selection	.60 ($\pm 1e^{-4}$)	.35 ($\pm 8e^{-6}$)	**.63** ($\pm 1e^{-4}$)	**.36** ($\pm 1e^{-4}$)	**.63** ($\pm 1e^{-4}$)	**.36** ($\pm 1e^{-4}$)
Year 10	Without feature selection	.68 ($\pm 2e^{-4}$)	.64 ($\pm 6e^{-5}$)	**.71** ($\pm 1e^{-4}$)	**.65** ($\pm 7e^{-5}$)	**.71** ($\pm 1e^{-4}$)	.61 ($\pm 2e^{-4}$)
	With feature selection	.68 ($\pm 1e^{-4}$)	.62 ($\pm 5e^{-5}$)	.70 ($\pm 1e^{-4}$)	**.63** ($\pm 4e^{-5}$)	**.71** ($\pm 9e^{-5}$)	.61 ($\pm 1e^{-4}$)

Table 1. Performance of models for graft survival prediction. Graft failure ratios for years 1, 5, and 10 are 7%, 20%, 44%, respectively. Best scores are highlighted in bold.

		Random forest		Cox PH model		Neural network	
		AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
Year 1	Without feature selection	.73 ($\pm 4e^{-4}$)	**.15** ($\pm 1e^{-4}$)	**.75** ($\pm 3e^{-4}$)	**.15** ($\pm 1e^{-4}$)	.71 ($\pm 4e^{-4}$)	.12 ($\pm 6e^{-5}$)
	With feature selection	**.74** ($\pm 3e^{-4}$)	**.15** ($\pm 3e^{-4}$)	.73 ($\pm 3e^{-4}$)	.14 ($\pm 8e^{-5}$)	**.74** ($\pm 3e^{-4}$)	.12 ($\pm 4e^{-5}$)
Year 5	Without feature selection	.78 ($\pm 1e^{-4}$)	.42 ($\pm 1e^{-4}$)	**.79** ($\pm 1e^{-4}$)	**.43** ($\pm 1e^{-4}$)	**.79** ($\pm 1e^{-4}$)	.41 ($\pm 2e^{-4}$)
	With feature selection	.76 ($\pm 1e^{-4}$)	.39 ($\pm 1e^{-4}$)	**.77** ($\pm 1e^{-4}$)	**.41** ($\pm 1e^{-4}$)	.76 ($\pm 1e^{-4}$)	.39 ($\pm 1e^{-4}$)
Year 10	Without feature selection	.80 ($\pm 1e^{-4}$)	.67 ($\pm 1e^{-4}$)	**.82** ($\pm 9e^{-5}$)	**.69** ($\pm 1e^{-4}$)	**.82** ($\pm 9e^{-5}$)	.68 ($\pm 1e^{-4}$)
	With feature selection	.80 ($\pm 9e^{-5}$)	**.66** ($\pm 1e^{-4}$)	**.81** ($\pm 7e^{-5}$)	**.66** ($\pm 1e^{-4}$)	**.81** ($\pm 7e^{-5}$)	**.66** ($\pm 1e^{-4}$)

Table 2. Performance of models for patient survival prediction. Patient death ratios for years 1, 5, and 10 are 3%, 14%, and 37%, respectively. Best scores are highlighted in bold.

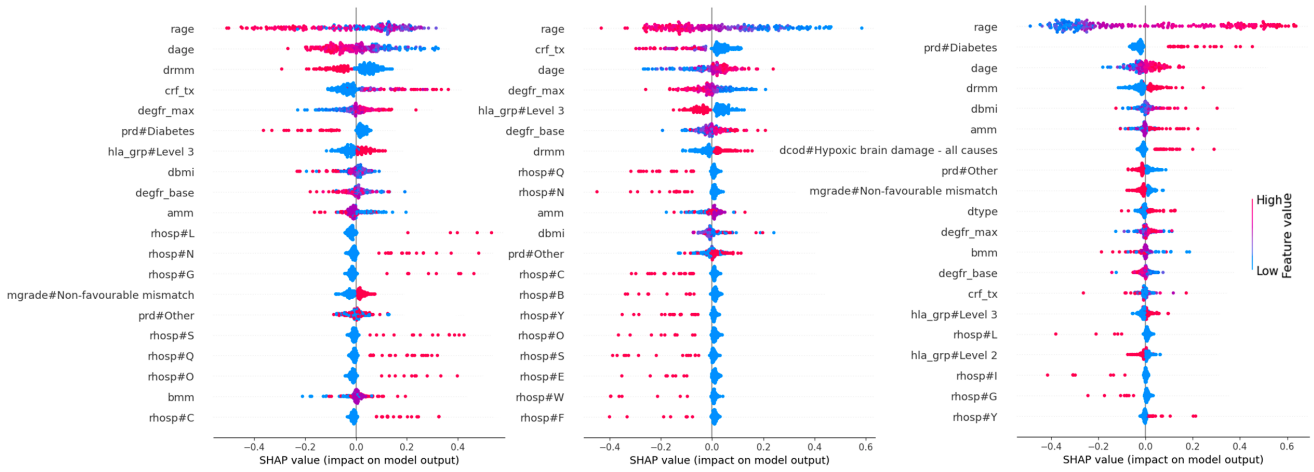
		Alive with functioning graft		Alive with failed graft		Dead	
		AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
Year 1	Without feature selection	.59 ($\pm 2e^{-4}$)	.93 ($\pm 8e^{-7}$)	.64 ($\pm 1e^{-3}$)	.09 ($\pm 2e^{-4}$)	.65 ($\pm 8e^{-4}$)	.10 ($\pm 1e^{-4}$)
	With feature selection	.59 ($\pm 1e^{-4}$)	.83 ($\pm 2e^{-5}$)	.59 ($\pm 3e^{-4}$)	.06 ($\pm 2e^{-4}$)	.69 ($\pm 6e^{-5}$)	.12 ($\pm 3e^{-5}$)
Year 5	Without feature selection	.64 ($\pm 6e^{-5}$)	.81 ($\pm 3e^{-5}$)	.58 ($\pm 3e^{-4}$)	.11 ($\pm 4e^{-4}$)	.74 ($\pm 4e^{-5}$)	.36 ($\pm 8e^{-5}$)
	With feature selection	.62 ($\pm 4e^{-5}$)	.74 ($\pm 2e^{-5}$)	.63 ($\pm 3e^{-4}$)	.09 ($\pm 3e^{-4}$)	.73 ($\pm 1e^{-4}$)	.38 ($\pm 2e^{-4}$)
Year 10	Without feature selection	.71 ($\pm 8e^{-5}$)	.70 ($\pm 8e^{-5}$)	.72 ($\pm 9e^{-4}$)	.18 ($\pm 1e^{-4}$)	.79 ($\pm 2e^{-5}$)	.63 ($\pm 6e^{-5}$)
	With feature selection	.71 ($\pm 6e^{-5}$)	.65 ($\pm 3e^{-5}$)	.71 ($\pm 5e^{-4}$)	.23 ($\pm 3e^{-4}$)	.79 ($\pm 3e^{-5}$)	.65 ($\pm 9e^{-5}$)

Table 3. Performance of multi outcome prediction models. Outcome ratios (alive with a functioning graft, alive with failed graft, and dead) for years 1, 5, and 10 are (.94, .03, .03), (.81, .15, .05), and (.56, .37, .07), respectively.

selected to predict graft failure and patient death, respectively. Notably, recipient and donor ages, primary renal disease, donor eGFR, donor type, and the number of mismatches at DR locus are important features to predict both outcomes.

Tables 1 and 2 provide both the AUROC and the F1-Score of each model on predicting graft failure and patient death, respectively, for observations years 1, 5, and 10. Performance before and after feature selection are presented. One can observe that overall performances increase with the observation time, being maximal at year 10. More specifically, the neural network has similar performances as the random forest and the Cox PH model, slightly outperforming them on the graft failure prediction task.

From an interpretability viewpoint, the neural network, when combined with SHAP, provides a richer clinical depiction of the data than Cox or the random forest. The features that are important to clinicians are also considered important to the neural network. For example, among predominant features for graft failure prediction (cf. Figure 1.a), recipient and donor age, donor type, donor past hypertension, or eGFRs are also features commonly used by regression models from the transplant literature [4, 18, 19]. The effect of feature values on predictions also matches clinical knowledge. For instance, patients with diabetes are likely to have inferior survival. This is reflected through the lower SHAP values regarding graft survival when `prd#Diabetes` is equal to one. The effect of covariates on survivability can be non-linear, as illustrated by the recipient age (`rage`; see Figure 1.b). Indeed, it is commonly recognised that younger patients can be less adherent to medication, hence increasing the risk of transplant failure. This phenomenon vanishes with older patients, and age then becomes a penalising



(a) Alive with functioning graft (b) Alive with failed graft (c) Deceased

Figure 2. SHAP values for multiple outcome predictions at year 10.

feature for survivability. In contrast, explanations obtained from the Cox PH model or the random forest do not highlight such behaviours (see Figures 1.d and 1.c), being limited to less expressive covariate effects. By design, it can be summarised as a linear function in the case of Cox, and the random forest sometimes fails to represent relevant dependencies between survival and numerical values (e.g. the recipient's age in Figure 1.c).

Multiple Outcome Prediction

In the multiple outcome case, 15 features are selected. Similarly to single outcome prediction, recipient and donor age, primary renal disease, eGFR, donor type, and number of mismatches at DR locus are present in this selection.

Table 3 stores the cause-specific AUROC and F1-Scores obtained by our neural network. Regarding AUROC, these results match the ones obtained in the single risk case. Notably, overall performances improve with $t^\star$. However, one can notice that the multiple outcome prediction problem is more subject to class imbalance. For instance, patients that are *alive with a failed graft* represent 3% of the total uncensored population at year 1, 5% at year 5, and 7% at year 10. Thus, according to F1-Scores, the model performs better in predicting the classes *alive with functioning graft* and *dead*. This is consistent with the results observed in the single risk case as better prediction was obtained regarding patient death over graft failure. From the interpretability viewpoint, we retrieve clinically coherent SHAP values (cf. Figure 2). Notably, we obtain interpretations that are consistent with the ones obtained in a single risk setting. Similarly, our neural network can reflect non-linear covariate effects.

Discussion

Neural networks have shown comparable performances to tools generally used by clinicians when predicting kidney transplant outcomes. In particular, they perform well when predicting long-term outcomes, which is a necessary property when considering the acceptance of an organ offer. The Cox PH model remains a robust solution in terms of performance, with little to no hyperparameter tuning. It is simple to use, leads to reliable predictions, and is easy to understand by clinicians. However, despite their black-box nature, neural networks stand out in terms of interpretability. Indeed, using SHAP allows us to have fine-grain interpretations of these models, which is not possible with the Cox PH model or random forests. This level of interpretability allows us to clinically validate these models, making them more trustworthy and explainable to patients. SHAP

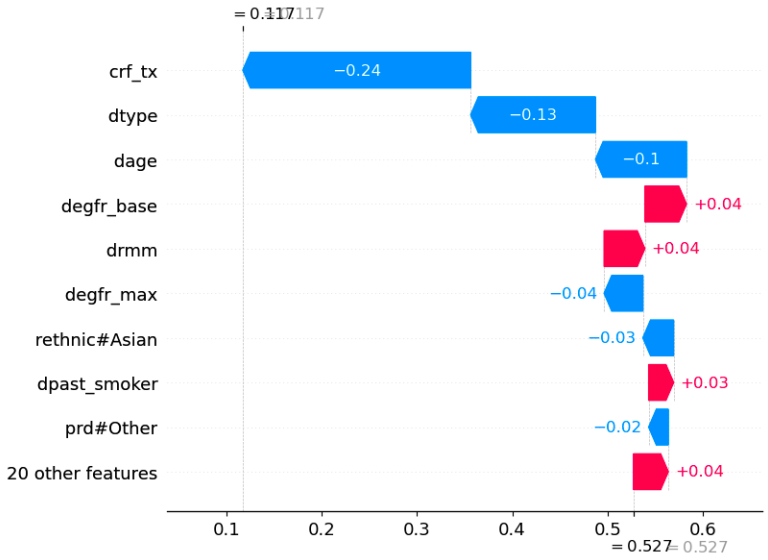


Figure 3. Explanation of a given prediction of graft survival at 10 years. The calculated reaction frequency at transplant is equal to 84%; donation occurred after circulatory death; the donor was 56 years old and non-smoker; eGFR remained equal to 113; no mismatches at the DR locus has been recorded; finally, the recipient is Asian with diabetes. Negative SHAP values indicate a negative impact on survivability, and *vice versa*.

can also highlight interesting relationships between covariates and transplant outcomes. Previous analyses of the UK registry data show that outcomes from kidney donations before and after circulatory death are equivalent regarding both patient and graft survival [20, 21]. Nonetheless, our models, trained from a larger dataset, suggest that donor type can have an impact on long-term transplant outcomes (see Figure 1.d, `dtype`). In practice, this level of interpretability is useful to explain individual prediction through the lens of SHAP. For example, Figure 3 shows the contribution of each features to a given prediction of graft survival at year 10. In this case, predicted survivability is mainly lowered by the calculated reaction frequency at transplant and the donor type. Neural networks can also deal with multiple outcomes, providing a more comprehensive prediction of the future state of the recipient, where patient death and graft failure are modelled jointly. There may be some clinical relevance to distinguishing the state of graft function at the time of death (*i.e.* death with or without graft failure). Unfortunately, the data set is too unbalanced and the population *dead with graft failure* is not large enough to provide such a distinction. These outcomes are not competing: a patient being alive with a failed graft at some point could die later, which remains an event of interest.

To conclude, we have trained several models to predict transplant outcomes from kidney offers, based on twenty years of registry data.Neural networks provide comparable results to classic survival analysis models, and can be easily extended to multiple outcome prediction. By using SHAP, we provide clinically validated interpretations of these models. This level of interpretability is especially relevant to enable validation from clinicians and to involve patients in the decision-making process. Therefore, neural networks represent a promising core component in the construction of future CDS for transplant offering. However, predicting transplant outcomes is only one aspect of the construction of a CDSS for kidney offering. Predicting what could be the consequences of refusing an organ offer in terms of future transplant opportunities, death, or removal from the waiting list is another key step. Having a good understanding of the outcomes in both scenarios is indeed necessary to predict individualised treatment effects. Uncertainty quantification is another critical research direction regarding the construction of a CDS tool for organ offering. Indeed, it can improve the trustworthiness of the tool by giving more insights about how difficult a given prediction is, and why. This can be achieved through post hoc error prediction using meta-modeling.

Methods

Data

Our work is based on the analysis of a data set from the UK Transplant Registry, provided with ethical approval by NHSBT. It describes 36,653 accepted kidney transplants, which have been performed between the years 2000 and 2020, across 24 UK transplant centres. The total follow-up duration is around 22 years. Each transplant is originally described with 3 identifiers, 12 immunosuppression follow-up indicators, 143 donor, recipient and transplant characteristics, and 7 entries describing targeted outcomes. Considering transplants as independent, we exclude the transplant, donor, and recipient identifiers. Additionally, information regarding post-transplant immunosuppression is discarded as this is not available at the time of the offer decision. The donor, recipient and transplant characteristics serve as input features for modelling. Among them, 24 describe the recipient, 109 represent the donor, and 10 refer to the overall transplant. Both recipient and donor characteristics contain generic information such as gender, ethnicity, age, blood group, height, weight, or body mass index (BMI). More specific information is also available, such as the transplant centre, number of previous transplants, waiting time, ease of matching, and the dialysis status. Donor data include the cause of death, past medical history and results of blood tests including kidney function (estimated glomerular filtration rate, eGFR). Transplant data include the donor-recipient immunological match.

Duplicate rows are removed, and we ensure that numerical values are within a plausible clinical range. Categorical values are checked by clinicians and simplified (or removed) if needed. BMI is recomputed based on weight and height. Both weights and heights are discarded to limit redundant information. Blood measurements are harmonised across the data set by selecting for each transplant the first measurement ever taken (generally at registration) and the maximum value ever recorded. Since the calculation of eGFR varies across hospitals, this metric is recomputed over the whole data set using a consistent definition (see appendix). Recipient dialysis status is also simplified into a dialysis duration and dialysis modality at time of transplant (predialysis, haemodialysis or peritoneal dialysis). Transplant offers not meeting the inclusion criteria, such as those leading to the transplantation of multiple organs, are discarded.

Outcomes present in the dataset include information about graft failure, patient death, and transplant failure. Transplant failure denotes either the graft failure or death. Each outcome is represented as a pair containing an event time and a right-censoring indicator. Right-censoring is a common type of censoring in survival analysis that describes the loss of follow-up on the event of interest. It can occur for various reasons, such as the end of the study, competing events, etc. Thus, right-censored information provides some partial information about the survival time, where it is only known to be greater than the censoring time. Minor missing censored indicators related to patient death are thus imputed based on graft information, and transplant outcomes are recomputed for the sake of consistency.

After removing the features presenting more than 5% missing values across the whole data set, and excluding any offer containing missing information, the resulting data set contains 25,370 transplants described through 45 input variables. A summary of the data-cleaning process is given in Figure 4. Additionally, an exhaustive list of the features and targets considered at the latest stage of this process is given in appendix.

Methodology

In this article, we first compare the Cox PH model to classification methods in a single risk setting. Subsequently, multiple outcome predictions are conducted by employing neural networks. The different models are interpreted *a posteriori*, and their performances are discussed. In both cases, the following methodology is applied. First, numerical values are standardised and categorical ones are one-hot-encoded. Standardisation appears to be more relevant than normalisation due to the presence of outliers in the data. Training is then performed through 5-cross validation on 80% of the data. Finally, the relevant performance metrics are computed and averaged from 100 bootstraps of the remaining 20%. The split into training and test data is done randomly, in a stratified manner with regards to censoring indicators. Due to matching policy changes and follow-up time differences between training and testing cohorts, we cannot split the data according to transplant dates. Classifiers are clinically interpreted using SHAP. When relevant, the coefficients of intrinsically interpretable models are also investigated (Code used

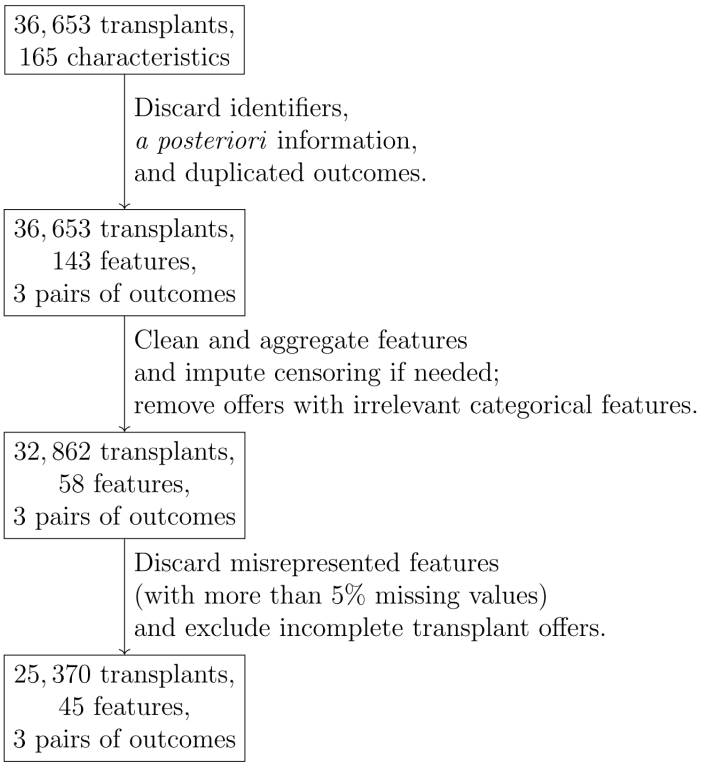


Figure 4. Data pre-processing steps.

for experiments can be found at https://github.com/AchilleSalaun/Xamelot).

Single Outcome Prediction

For a single type of event interest (e.g. graft failure), we want to predict the occurrence of that event before an arbitrary time point $t^\star$. The Cox PH model and classifiers consider different kinds of input data, tackling different formulations of that problem. Cox returns a survival function, describing the probability of survival (as in the absence of an event) with regard to time. Therefore, the probability of an event at time $t^\star$ can be obtained by evaluating the survival function at that time. Conversely, the censored data must be converted into labels to be fed into classifiers. Let us denote the pair (event time, censoring indicator) by (t, c). Then, the graft is *functioning* (or the patient is *alive*) if $t^\star < t$; the graft *failed* (the patient is *dead*) if $t^\star \geq t$ and the event has not been censored ($c = 1$); finally, the status of the graft (or patient) is *unknown* if $t^\star \geq t$ and the event has been censored ($c = 0$) (see Figure 5). Dropping unknown labels induces a binary classification problem since the graft is now either functioning or failed, or the patient is either alive or dead. This last operation generally assumes censoring and events to be independent, which often leads to biases in practice. However, the Cox PH model relies on the same assumption: likelihood's maximisation is achieved by managing a risk set over time, that is a set of subjects that are still under follow-up [6]. Hence, censored events with lost follow-up are implicitly dropped while training Cox PH models. The shorter the observation time $t^\star$, the more imbalanced the outcome distribution is, with failure or death being under-represented regarding survival. For instance, the class imbalance goes from 7% of graft failures at 1 year, to 44% at 10 years. As we drop censored events with a censoring time that is lower than the observation time, the number of data points used for training also varies with regard to time. Thus, the training data sets used for predicting transplant outcomes at years 1 and 10 comprise 23,422 and 10,017 data points, respectively.

We can now compare the respective abilities of the Cox PH model, random forests, and neural networks to predict transplant outcomes using the area under the receiver operating characteristic (AUROC). The choice of AUROC is motivated by two aspects. First, it is conceptually close to concordance which is the metric generally used for survival analysis. Second, it is a

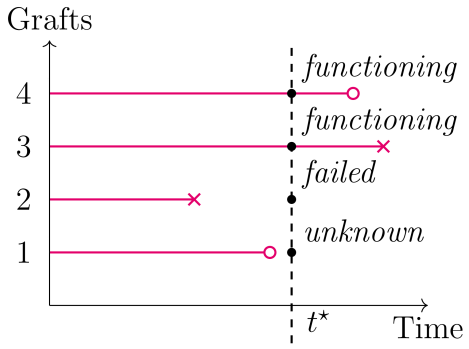


Figure 5. Translation of a single risk survival analysis problem into a classification task. Event times are given by the x-axis; $\times$ indicates the observation of an event, $\circ$ indicate censoring. For any time $t^\star$, graft (or patient) status can be derived from survival data.

good metric when dealing with balanced data, which is the case when predicting transplant outcomes after 10 years. Therefore, this metric is relevant since we have a particular interest in predicting long-term outcomes. However, as class imbalance remains a recurrent difficulty, we also compute the F1-Score. For a given model, we select the classification threshold that leads to the best possible F1-Score. From an instantiation viewpoint, Breslow's estimator is used to derive the Cox PH model's baseline [22]. In addition, a regularisation parameter is introduced and set to $1e^{-4}$ to deal with colinearities in the data. The random forest [23] contains 1,000 trees, relies on Gini's criteria, and adjusts class weights automatically. To predict graft survival (or patient death) at years 1, 5, and 10, a neural network is instantiated with one hidden layer of 400 (200, respectively) neurons, activated by a ReLU, and with 10% dropout. The loss is a weighted cross to handle class imbalance. Finally, the training is done through 20 epochs, with batches of size 8 (32, respectively), using RMSProp and a learning rate equal to $1e^{-4}$.

Feature selection is performed after preliminary training. We first inspect the effect of each feature on prediction to detect potential inconsistencies in the data. Then, we temporarily add random noise: features that are shown to be less important in the prediction than noise are discarded. Finally, we progressively remove the less relevant features from the set of selected features until a noticeable decrease in performance is observed. For a given type of outcome, the set of selected features corresponding to year 10 includes important features for prediction in earlier years; therefore we use the same set of features for years 1 and 5.

Multiple Outcome Prediction

For further analysis, we generalise our approach to multiple outcomes prediction. For an arbitrary time point $t^\star$, we want to know whether the patient is *alive with a functioning graft*, *alive with a failed graft*, or *dead*. The construction of these labels from censored data is similar to the one performed in the single risk case. Transplants with unknown status due to censoring are discarded. To address this new problem, we focus on neural networks as they appeared to be a promising approach in the single risk case (see Results and Discussion). We train a neural network with one hidden layer of 1000 neurons activated by ReLU. A dropout is set to 10%. The loss is a weighted cross entropy and the training is done through 100 epochs, with batches of size 64. Optimisation relies on RMSProp with a learning rate set to $1e^{-3}$. Similar to the single risk case, feature selection is conducted after initial training using all features.

Data availability statement

The dataset supporting this study is the property of NHSBT (National Health Service Blood and Transplant). Due to the presence of patient data that could potentially lead to identification, obtaining prior ethical approval from NHSBT is mandatory. Consequently, the supporting data is not accessible for external use.

References

1. M. Chun, R. Clarke, B. J. Cairns, D. Clifton, D. Bennett, Y. Chen, Y. Guo, P. Pei, J. Lv, C. Yu, *et al.*, "Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults," *Journal of the American Medical Informatics Association*, vol. 28, no. 8, pp. 1719–1727, 2021.

2. D. W. Kim, S. Lee, S. Kwon, W. Nam, I.-H. Cha, and H. J. Kim, "Deep learning-based survival prediction of oral cancer patients," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

3. R. S. Lin, S. D. Horn, J. F. Hurdle, and A. S. Goldfarb-Rumyantzev, "Single and multiple time-point prediction models in kidney transplant outcomes," *Journal of biomedical informatics*, vol. 41, no. 6, pp. 944–952, 2008.

4. P. S. Rao, D. E. Schaubel, M. K. Guidinger, K. A. Andreoni, R. A. Wolfe, R. M. Merion, F. K. Port, and R. S. Sung, "A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index," *Transplantation*, vol. 88, no. 2, pp. 231–236, 2009.

5. A. J. Vinson, B. A. Kiberd, R. B. Davis, and K. K. Tennankore, "Nonimmunologic donor-recipient pairing, HLA matching, and graft loss in deceased donor kidney transplantation," *Transplantation direct*, vol. 5, no. 1, 2019.

6. D. R. Cox, "Regression models and life-tables," *J R Stat Soc*, 1972.

7. J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Med. Res. Methodol.*, 2018.

8. H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in biostatistics: competing risks and multi-state models," *Stat. Med.*, 2007.

9. "NHS risk communication tools." https://www.odt.nhs.uk/transplantation/tools-policies-and-guidance/risk-communication-tools/.

10. S. Krikov, A. Khan, B. C. Baird, L. L. Barenbaum, A. Leviatov, J. K. Koford, and A. S. Goldfarb-Rumyantzev, "Predicting kidney transplant survival using tree-based modeling," *Asaio Journal*, vol. 53, no. 5, pp. 592–600, 2007.

11. S. A. A. Naqvi, K. Tennankore, A. Vinson, P. C. Roy, and S. S. R. Abidi, "Predicting kidney graft survival using machine learning methods: prediction model development and feature significance analysis study," *Journal of Medical Internet Research*, vol. 23, no. 8, p. e26843, 2021.

12. V. Rao, R. S. Behara, and A. Agarwal, "Predictive modeling for organ transplantation outcomes," in *2014 IEEE International Conference on Bioinformatics and Bioengineering*, pp. 405–408, IEEE, 2014.

13. K. Topuz, F. D. Zengul, A. Dag, A. Almehmi, and M. B. Yildirim, "Predicting graft survival among kidney transplant recipients: A bayesian decision support model," *Decision Support Systems*, vol. 106, pp. 97–109, 2018.

14. C. Molnar, *Interpretable machine learning*. self published, 2020.

15. A. Noack, I. Ahern, D. Dou, and B. Li, "An empirical study on the relation between network interpretability and adversarial robustness," *SN comput. sci.*, 2021.

16. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

17. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

18. C. J. Watson, R. J. Johnson, R. Birch, D. Collett, and J. A. Bradley, "A simplified donor risk index for predicting outcome after deceased donor kidney transplantation," *Transplantation*, vol. 93, no. 3, pp. 314–318, 2012.

19. M. Z. Molnar, D. V. Nguyen, Y. Chen, V. Ravel, E. Streja, M. Krishnan, C. P. Kovesdy, R. Mehrotra, and K. Kalantar-Zadeh, "Predictive score for posttransplantation outcomes," *Transplantation*, vol. 101, no. 6, p. 1353, 2017.

20. D. M. Summers, R. J. Johnson, A. Hudson, D. Collett, C. J. Watson, and J. A. Bradley, "Effect of donor age and cold storage time on outcome in recipients of kidneys donated after circulatory death in the UK: a cohort study," *The Lancet*, vol. 381, no. 9868, pp. 727–734, 2013.

21. D. M. Summers, C. J. Watson, G. J. Pettigrew, R. J. Johnson, D. Collett, J. M. Neuberger, and J. A. Bradley, "Kidney donation after circulatory death (DCD): state of the art," *Kidney International*, vol. 88, no. 2, pp. 241–249, 2015.

22. D. Lin, "On the Breslow estimator," *Lifetime data analysis*, vol. 13, pp. 471–480, 2007.

23. L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

Legends

Figure 1 Interpreting single outcome prediction models. Both models have been trained to predict graft failure at 10 years.

Table 1 Performance of models for graft survival prediction. Graft failure ratios for years 1, 5, and 10 are 7%, 20%, 44%, respectively. Best scores are highlighted in bold.

Table 2 Performance of models for patient survival prediction. Patient death ratios for years 1, 5, and 10 are 3%, 14%, and 37%, respectively. Best scores are highlighted in bold.

Table 3 Performance of multi outcome prediction models. Outcome ratios (alive with a functioning graft, alive with failed graft, and dead) for years 1, 5, and 10 are (.94, .03, .03), (.81, .15, .05), and (.56, .37, .07), respectively.

Figure 2 SHAP values for multiple outcome predictions at year 10.

Figure 3 Explanation of a given prediction of graft survival at 10 years. The calculated reaction frequency at transplant is equal to 84%; donation occurred after circulatory death; the donor was 56 years old and non-smoker; eGFR remained equal to 113; no mismatches at the DR locus has been recorded; finally, the recipient is Asian with diabetes. Negative SHAP values indicate a negative impact on survivability, and *vice versa*.

Figure 4 Data pre-processing steps.

Figure 5 Translation of a single risk survival analysis problem into a classification task. Event times are given by the x-axis; $\times$ indicates the observation of an event, $\circ$ indicate censoring. For any time $t^\star$, graft (or patient) status can be derived from survival data.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions. This work has been supported by funds from the NIHR (AI Award 2020 Phase 1: `AI_AWARD02316`). T.Z. was supported by the Royal Academy of Engineering under the Research Fellowship scheme.

Author contributions statement

A.S. undertook the data cleaning, model building, and redaction of this paper. S.K. and L.W. provided clinical input to data cleaning and model design. S.K. and T.Z. co-ordinate the overall project, providing respectively clinical and machine learning insights.

Additional information

No competing interest is declared.