Predicting transplant and patient survival following liver transplantation using machine learning

Simon Thorogood

Master of Science

School of Informatics
University of Edinburgh

2019

Abstract

The task of allocating donor livers to recipients for the purpose of transplantation is a critical and complex task that is exacerbated by a mismatch between supply and demand. Currently score-based approaches are used to rank recipients either in terms of their need or the likely benefit to them of a transplant in terms of increased survival. These scores are typically derived from linear statistical models such as Cox proportional hazards regression.

In this project we use a large-scale UK liver transplant dataset to compare the performance of liver transplant scores and traditional statistical models with a number of machine learning approaches when applied to the task of predicting survival following liver transplantation. The machine learning approaches considered include random forests and ANN classifiers as well a number of approaches that are specifically designed for use in a survival analysis context. We show that several of the machine learning models evaluated are able to predict transplant and patient survival at 1 and 5 years post-transplant significantly better than the traditional approaches.

Acknowledgements

I would like to thank Dr. Jacques Fleuriot for his support, guidance and enthusiasm throughout this project. The timely guidance and encouragement provided by the team at the Centre for Evidence in Transplantation (CET) in Oxford was also greatly beneficial to this work and was much appreciated.

I'm also very grateful for the support, patience and understanding of my family Clare and Anna, not just during this final project phase but throughout the past year.

Table of Contents

Chapter 1

Introduction

In this introductory chapter we provide an overview of the field of liver transplantation and the process by which donor organs are currently allocated. A number of the challenges faced by clinicians in this area are also highlighted. This is followed by an outline of the motivation for this work and the project objectives.

1.1 Liver transplant allocation

First becoming a widespread procedure in the 1980s, liver transplantation is currently the only effective treatment for end-stage liver disease, hepatocellular cancer (cancer of the liver) and a number of metabolic disorders. Globally, the number of patients awaiting a liver transplant (or graft) exceeds the number of donor livers available. This results in long waiting times and ultimately in the deaths of a significant number of patients before a suitable donor can be found [1]. This supply imbalance is further compounded by the fact that the overall quality of livers made available for transplant has decreased over time [2]. As a result, clinicians are increasingly forced to consider the merits of transplanting marginal or extended criteria donor (ECD) livers against the risk of a patient continuing to remain on the waiting list. Examples of marginal or ECD livers include those donated after cardiac death (DCD), livers from older donors (e.g. those over 65) or those from donors with steatosis (an accumulation of fat in the liver) [3]. The same organ transplanted into two different recipients might have quite different outcomes and making the wrong decision can result in graft failure leading to the death of the recipient or the need for a costly second transplant. The process of allocating livers is therefore a complex and finely-balanced one, with many donor and recipient factors to be considered and often multiple potential recipients for a single

donor organ.

To address these challenges, a number of scoring approaches have been developed in the past 20 years which form the basis of most liver transplant allocation processes today (see section 2.3). At the point of allocation such scores are used to quantify the predicted benefit for a recipient of receiving a donated organ and/or the risk associated with them remaining on the waiting list. Most of these scoring approaches are derived from traditional statistical survival analysis models that estimate the importance of different factors on survival. Whilst the use of such scores have been shown to improve overall patient outcomes and survival rates, they are limited in the number of factors that they can consider and their basis in linear statistical models mean they are unable to capture more complex interactions between factors. Recently, there has been growing evidence that the application of supervised machine learning (ML) techniques to this task can help improve predictive accuracy further (see section 2.4).

1.2 Project scope and objectives

Clearly toolsets that help clinicians predict the outcomes of transplantation can form a useful adjunct to clinical judgement during the allocation process. They have the potential to help maximise both the utilisation of organs and the survival benefits for patients.

In this project we explore the benefits of applying ML techniques to the prediction of survival following liver transplantation. ML models are known to have a number of potential advantages over traditional statistical models. In particular, their ability to learn complex non-linear relationships can lead to better predictive accuracy. They are also able to handle large numbers of input variables with ease and tend to be more robust to missing data. Finally, since they are driven primarily by the data, they can be easily trained and validated for a specific population and kept up to date as new survival data becomes available. Conversely, the predictions made by ML models can be less straight-forward to interpret than those of regression-based models, an important consideration in any medical field and particularly for a sensitive task such as organ allocation.

Prediction of patient and transplant survival over the short- (i.e. 1 year) and longer-term (i.e. 5 years) are addressed since both are considered during the organ allocation process. Separate models based on donor factors only and on combined donor and recipient factors are considered in order to provide tools suitable both for the evaluation

of a donor liver at the point of retrieval and for the evaluation of a specific donor-recipient (D-R) pair[1].

The project was carried out in collaboration with the Centre for Evidence in Transplantation (CET) at the University of Oxford, who provided subject matter expertise and guidance throughout the project. A dataset provided by National Health Service Blood and Transplant (NHSBT) was used to carry out the work.

1.3 Thesis organisation

The remainder of this thesis is structured as follows. Survival analysis and related ML approaches are covered in chapter 2 together with a literature review of scoring and ML approaches to the prediction of survival following liver transplantation. In chapter 3 we introduce the dataset used together with details of pre-processing steps applied and the results of initial data analysis. The methodology used to implement train and compare different models is covered in chapter 4. Finally results are presented in chapter 5 together with discussion and conclusions in chapter 6. A number of appendices are also included that provide supplementary materials and results that are not deemed essential to the core thesis.

[1]Timescales and types of model were established in conversations with the Centre for Evidence in Transplantation team

Chapter 2

Background

In this chapter we introduce the field of survival analysis and review some of the machine learning approaches that can be applied in such a context. This is followed by a review of liver transplantation scoring models and previous ML approaches to the prediction of survival post-transplant.

2.1 Survival analysis

Survival analysis is a branch of statistics concerned with the connection between one or more covariates and the expected time to a particular event of interest [4]. The event in question might be the death of a patient after surgery, the re-occurrence of a disease following treatment, the failure of a mechanical component or a customer defaulting on a loan. The time in question can relate to the expectation for a group of individuals or to a single individual. In such models the covariates in question are typically measured once at time $t = 0$.

A key way in which survival analysis approaches differ from other statistical modelling techniques is that they are able to account for data that can only be partially observed. This is known as *data censoring*. If we consider a clinical trial over a period of 5 years, we might expect to see a proportion of individuals experience an event of interest in that time. We say that these events are uncensored since we are able to record the precise times when they occurred. Other subjects might drop out of the study, die of an unrelated cause or not experience the event of interest before the end of the study. These events are all said to be right-censored in that we know that the event did not occur before a specific time but not precisely when it did occur (if at all). Formally, for right-censored data, if δ is an indicator of whether the event was censored or not

then the observed time y can be defined as equal to the event time t (when $\delta = 1$) or the censoring time c (when $\delta = 0$). Other types of censoring (e.g. left-censoring and interval-censoring) also exist but are outside the scope of this project.

A key quantity of interest in survival analysis is the survival function $S(t)$ which determines the probability that the event of interest has not occurred at time t. Here T is the random variable indicating the time the event occurred

$$S(t) = P(T > t). \tag{2.1}$$

Kaplan-Meier (KM) estimators [5] can be used to estimate $S(t)$ and create survival curve plots for right-censored datasets. The estimator is defined as

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - e_i}{n_i} \tag{2.2}$$

where e_i are the number of events that occurred at time t_i and n_i are the number of subjects at risk immediately prior to time t_i. KM estimates of the survival function are commonly used in conjunction with a log-rank test [6] to compare the survival function of two or more groups in a trial and determine whether any differences between them are statistically significant. Though useful for looking at group-level survival rates, KM estimators are based on aggregated observations and therefore are not able to account for the effect of multiple covariates on the survival function or to produce survival predictions for an individual within a group.

Another important quantity of interest in survival analysis is the hazard function $\lambda(t)$ which is the instantaneous event rate for an individual who has survived up until time t. It can be shown to be equal to the first derivative w.r.t. time of the log of the survival function $S(t)$ [4].

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t | T \geq t)}{\Delta t} = \frac{\partial \log(S(t))}{\partial t} \tag{2.3}$$

Cox proportional hazards (Cox PH) [7] is a regression model that allows the effect of 1 or more covariates on a baseline hazard function $\lambda_0(t)$ to be estimated. The model fitted is of the form

$$\lambda(t) = \lambda_0(t) \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right) \tag{2.4}$$

where $\lambda(t)$ is the hazard function adjusted for the model covariates $x_1, \ldots, x_k$. Here we refer to $\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)$ as the risk function and $\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

as the log-risk function. Cox PH doesn't estimate the form of $\lambda_0(t)$ itself, only that of the linear log-risk function and as such it is considered a semi-parametric method. In order to perform Cox PH regression, the coefficients β are tuned to optimize the Cox partial likelihood which is defined as the product of the probability at each event time T_i that the event has occurred to individual i, given the set of individuals still at risk at time T_i.

Inherent to the model assumptions are that i) covariates are independent of one another, ii) each covariate has a linear effect on the log-risk function iii) the ratio of the hazard function for any two individuals is constant over time

Despite these somewhat restrictive model assumption, Cox PH has in practice been shown to be a effective method for estimating the relationship between multiple covariates and the time to an event in the presence of right-censored data. As such, its use today in medical and other fields is widespread. As well as allowing the relative contribution of model covariates to the hazard function to be quantified, Cox PH can also be used to produce survival predictions for a single individual when combined with an estimate of the baseline hazard function $\lambda_0(t)$.

2.2 Machine learning approaches to survival analysis

The distinction between statistical modelling and machine learning is not always clear-cut but here we define statistical models to be those that assume an underlying probability distribution for the data generating process. Latent parameters associated with the distribution are then estimated using data samples. In contrast, machine learning approaches tend to be more algorithmic and allow complex interactions between covariates to be learnt directly from training data without any assumptions being made about their underlying distribution.

A simple way to apply machine learning to survival analysis tasks is to re-frame them as classification tasks. In order to do this survival times are converted to binary labels associated with survival at a specific future time (e.g. 'Did the event occur within 1 year?'). Handling of right-censored data must be considered with the most common approach being to drop any rows censored prior to the time of interest. Assuming that not too many rows are discarded as a result, this approach allows any type of machine learning classifier to be applied to the task.

A number of machine learning approaches have also been developed that directly target survival analysis tasks and which are able to make use of right-censored data

during training. These include random survival forests (RSFs) and a number of approached based on modified artificial neural networks (ANNs). In the next two sections we provide an overview of traditional random forest and ANN classifiers along with an overview of some related ML approaches that have be adapted specifically for survival analysis tasks.

2.2.1 Tree-based approaches

A decision tree [8] is a learning model based on a binary tree structure. Each internal node in the tree represents a yes/no question based on a single model covariate, with the data split based on the answer. The predicted outcome of each leaf node based based on the majority class of the associated samples. Decisions trees are typically learned by considering multiple candidate covariates and split points at each internal node and selecting the one that provides the biggest increase is homogeneity of classes in the resulting subsets. A number of metrics exist for assessing this homogeneity including entropy (or information gain) and Gini impurity (a measure of logical entropy), with the latter being most commonly used for classifiers.

Random forests [9] is an ensemble method based on training multiple decision trees in parallel and then taking the majority response as the model output. Sampling of data and features is used to reduce the overall generalisation error of the model. In particular, bagging [10] is used to select a subset of training data used by each tree and a randomly selected subset of variables is considered at each internal split node. Calculating the ratio of class predictions across trees also allows RF ensembles to produce probabilistic predictions to which a threshold can be applied, allowing model precision to be traded off against recall. Random forests allow the relative importance of each variable to be determined by considering how often a variable was selected to split on during training, and how much the squared error improved as a result. This provides a valuable degree of interpretability for such models.

The random survival forests (RSF) method [11] is an extension of random forests that supports the analysis of right-censored survival data. The implementation of an RSF follows the same pattern as for a random forest but with a modified splitting rule that is able to account for right-censored data by measuring the survival difference between the samples on either side of the split. RSF has become popular as a non-parametric alternative to Cox PH due to its less restrictive model assumptions.

2.2.2 ANN-based approaches

ANNs are machine learning models inspired by the biological neural networks found in human brains. In an ANN, a collection of units (neurons) are connected via sets of weight parameters that are learned in an iterative fashion using numerous labelled training examples. A typical ANN model has an input layer, one or more intermediate hidden layers, and an output layer that represent either the model's predicted probabilities for different class labels or the predicted value of a regression variable. ANNs have the ability to learn complex non-linear interactions between variables and in recent years have produced state-of-the-art performance in a wide range of fields including image classification and natural language processing [12, 13].

As previously discussed, conventional ANN classifiers can be applied to survival data by assigning point-in-time event labels and dropping previously censored rows. There are however also a number of ANN-based approaches that address survival analysis tasks with right-censored data directly. These are reviewed in the next part of this section.

An early approach [14] used a very simple ANN with no hidden layers to predict the risk function for discrete survival time intervals. Each time interval has a corresponding output node in the network and the model can be considered to be identical to fitting a 'grouped' version of Cox PH regression. Faraggi and Street [15], also use a simple ANN with a single output node and 1 hidden layer. Conceptually, the output node represents the log-risk function of the Cox PH model. The model is trained by maximising the partial likelihood in a fashion similar to that used to opimtise Cox PH regression models allowing right-censored data to be incorporated. Neither of these models however have been shown to outperform traditional Cox PH models [16].

PLANN [17] is another ANN-based approach with a single output node and 1 hidden layer. Here survival times are grouped into regular intervals and used as an input to the model along with the other covariates. The model output is used to estimate a 'smoothed' hazard function $\lambda(t)$. Partial likelihood is again used as a loss function to train the network. Street [18] also employed grouped survival times but in this case the model has multiple output nodes with one for each time interval considered. Right-censored data is incorporated into training by estimating survival probabilities for intervals beyond the censoring time using KM estimates. This results in 'soft' labels which are optimised using a cross entropy loss function. This approach is compared with PLANN directly by Sharaf [19] with the approach proposed by Street [18]

shown to have superior performance.

DeepSurv [20] is a recent approach that draws on the work of Faraggi and Street [15] in that the Cox PH log-risk function is also estimated. The ANN model used however is significantly more complex and employs modern deep learning techniques including dropout [21], an Adam optimiser [22] and Scaled Exponential Linear Units (SELU) activation functions [23]. When compared with a traditional Cox PH model, DeepSurv retains the same proportionality constraint but has the advantage of being able to learn non-linear interactions between covariates. In a series of test on simulated and real-world datasets, the model is shown to produce superior C-index scores (see section 4.2) compared to a Cox PH model and to be on par with an equivalent RSF model.

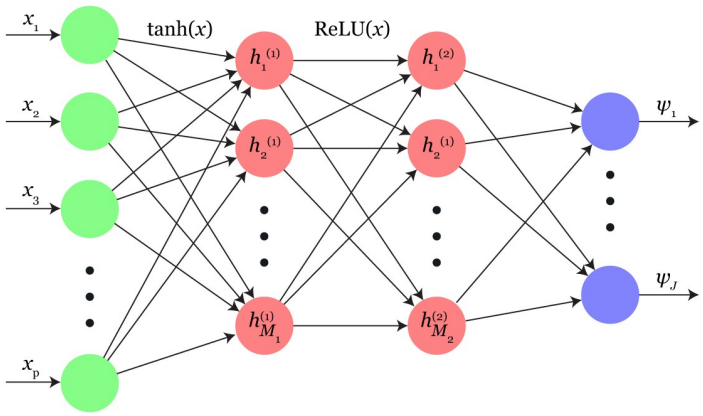

Figure 2.1: Representation of a 2-hidden layer N-MTLR model. Each output unit ($\psi_1 \dots \psi_J$) represents an probability of survival for a distinct time interval. Reproduced from Fotso's paper [24].

Yu et al. [25] proposes a method named multi-task logistic regression (MTLR) to learn individual survival distributions. The survival function is approximated by combining multiple logistic regression models where each predicts survival for a specific time interval. Survival labels are encoded in a binary sequence similar to those used by Street [18] and censored data is handled by considering the likelihood over all sequences that are consistent with the censoring time. Additional regularisation terms in the loss function ensure that probabilities vary smoothly across time intervals. Work by Fotso [24] replaces the linear logistic regression core of MTLR with an ANN allowing more complex models to be learnt (see Figure 2.1). The model, referred to as

Neural MTLR (N-MTLR), is show to outperform both MTLR and traditional Cox PH models on datasets where non-linear interactions between features are present.

2.3 Liver transplant scoring approaches

Scoring approaches can be broadly divided into those based on recipient factors only, those based on donor factors only and those based on a combination of donor, recipient and other factors. They generally take the form of a sum of a scaled set of factors where scaling is defined by weight coefficients established using regression models.

2.3.1 Recipient scores

Scores based on recipient factors were the first to be used in practice and were designed to address limitations with previous 'first-come, first-served' policies that were unable to account for the urgency associated with a specific patient. The model for end-stage liver disease (MELD) [26] is a score that is based on the following factors: recipient liver disease etiology (i.e. the cause of the disease), bilirubin, international normalised ratio (INR) and creatinine. The score has been shown to be predictive of 3 month waiting-list survival (i.e. patient survival if no transplant is carried out) and so can be used to assess the relative urgency of patients on the waiting list. It currently forms the basis of liver allocation policy in the US. A number of MELD variants have subsequently been developed to address specific scenarios or geographic variations. These include UKELD [27], a UK-specific version of MELD that formed the basis of allocation policy in the UK until recently.

2.3.2 Donor scores

The growing use of scores based on donor factors has been driven by the increase utilisation of extended criteria donor (ECD) livers (see section 1.1) and the need to qualitatively assess the increased risk of using such organs. The donor risk index (DRI) [28] was developed using Cox PH models and a US dataset to identify the most predictive donor factors for graft failure. These include: age, donor cause of death, split grafts, ethnicity, height and cold ischemic time. The donor liver index (DLI) [29] is a UK-specific version of DRI also developed using analysis based on Cox PH models. Two separate DLI indices were developed, the first covering survival at 1 year and the other overall survival up to 10 years post-transplant. The following factors were

collectively identified as being most predictive of failure: age, gender, donor cause of death, split or partial grafts, height, smoking history, bilirubin, ethnicity, presence of steatosis and history of cardiac disease (the interested reader is referred to section D.1).

2.3.3 Combined donor-recipient scores

In order to fully assess the suitability of a specific donor-recipient match in the allocation process, it is necessary to consider both donor and recipient factors in parallel. To this end a number of scoring approaches have been developed that combine donor and recipient factors. D-MELD [30] is a simple extension to MELD that additionally accounts for donor age. Survival outcomes following liver transplantation (SOFT) [31] was developed to predict post-transplant survival at 3 month, as a complement to the 3 month waiting-list survival prediction of the MELD score. The score uses 13 recipient factors, 4 donor factors and 2 operative factors.

Since 2017 liver allocation policy in the UK has been based on the National Offering Scheme[1]. Under this scheme, livers are allocated to recipients on the elective waiting list on the basis of a Transplant Benefit Score (TBS). This score, calculated using both donor and recipient factors, takes into account the difference between the predicted waiting list survival curve and the predicted post-transplant survival curve. A total of 21 recipient and 7 donor factors are integrated in the score including: age, gender, donor cause of death, recipient etiology and numerous biometric indicators (the interested reader is referred to section D.2). In some cases interactions between factors are also included. Separate scoring formulae are used for cancer and non-cancer patients to generate a Cox PH risk function value which is then combined with a baseline survival curve estimate to produce a predicted individual survival curve.

2.4 Machine learning approaches to liver transplant survival prediction

Except where otherwise stated, the studies reviewed in this section are limited to those concerned with the prediction of individual liver transplant survival using data that is available pre-transplant.

Haydon et al. [32] makes use of self-organising maps (SOM) to learn a low dimensional mapping of the input factors and demonstrates how such a model could be

[1] https://www.britishlivertrust.org.uk/new-system-liver-transplant/

used to predict suitable matches in a real-world setting. A Bayesian network [33] built using a large US dataset (around 12,000 patients) was able to predict 3 month survival with AUC ROC values (see section 4.2) that were competitive with those produced by a number of linear models using the same dataset (in the range 0.6-0.7). In 2007 Cucchetti et al. [34] used an Italian dataset of around 250 recipients to train an ANN to predict 3 month survival. The model was shown to have AUC ROC of 0.95 and outperform predictions based on the MELD score by a significant amount.

Zhang et al. looked at non-cancer and cancer patient survival seperately in 2 studies [35, 36]. Both studies involved ANNs trained using features selected using forward stepwise selection. When compared using C-index to a number of scoring approaches the ANN models were shown to be superior.

In 2 closely related works [37, 38] evolutionary algorithms were used to determine the optimal architecture for ANN models that predict 3 month graft survival and loss rates for a Spanish dataset containing around 1,000 patients. The best of these models [38] was found to outperform score-based approached by up to 34% with an absolute AUC ROC value of 0.82.

A study of around 1,200 Iranian transplants attempted to to predict survival up to 5 years post-transplant [39]. A 3 layer ANN model is compared with a traditional Cox Regression models and is shown to outperform the Cox Regression model by around 7% (measured using AUC ROC). It should be noted however that this model utilised a number of post-operative factors.

Dorado-Moreno et al. [40] use an ANN architectural variant that is able to model ordinal classification to predict 0-15, 15-90 and 90-365 and 365+ day survival probabilities respectively. A number of techniques including over-sampling are employed to address class imbalances in the dataset. This model is compared with a number of other approaches including support vector machines (SVMs), random forests and gradient boosted trees. The ANNs models are shown to have promising results, particularly with regards to the classification of minority classes.

In work by Lau et al. [41] models based on random forests and ANNs are used to predict 30 day post-transplant survival rates using a small Australian D-R dataset (around 200 patients). Both types of models are shown to increase predictive accuracy with the best ANN model increasing AUC ROC by as much as 30% compared to score-based approaches including MELD, DRI and SOFT. A best absolute AUC ROC of 0.84 was obtained here.

Chapter 3

Data Pre-processing and Analysis

3.1 National Transplant Register dataset

The dataset used throughout this project is an anonymised subset of the National Transplant Register (NTR) provided by NHSBT. Permission for the use of the dataset was obtained jointly by CET and Dr. Jacques Fleuriot of the University of Edinburgh specifically for the purpose of this project. The dataset contains details of D-R pairs for liver transplants that were carried out in the period between January 2000 to December 2016. For each D-R pair, observations about the donor and recipients (including demographics, medical history and biomedical indicators), details of the transplanted organ and the surgery carried out are recorded. Survival time data relating to both the the transplanted organ and the transplant recipient are also recorded, where available, up to 2018.

The dataset consists of 10,388 D-R pairs as well as supplementary metadata describing the columns present in the dataset and details of the categorical numeric codes used. The following categories of transplant were excluded at source from the dataset: transplants for recipients under the age of 16; auxiliary and heterotopic liver transplants and blood group incompatible transplants. Transplants, donors and recipients are identified only by anonymised IDs within the dataset which was supplied as an Excel spreadsheet.

3.2 Data pre-processing

Prior to analysis and modelling, a number of initial data cleaning steps were carried out on the dataset.

A total of 19 rows had no survival data (i.e. the transplant survival time was not recorded) and so were discarded. In 2016 the recording of donor biomedical indicators was changed to allow separate retrieval and referral values to be recorded. A increase in the percentage of missing values was also observed after this time. For these reasons, 809 post-2016 rows were also discarded. All anonymised identifier columns (e.g. transplant ID, donor ID etc.) were deemed irrelevant to the prediction task and were discarded.

Many of the categorical fields contained separate codes for explicit 'Unknown' and 'Not Reported' observations. Rows with these types of codes were combined together with null observations (i.e. where no value is recorded) into a single null group.

A number of the supplied categorical columns had high cardinality ($\geq$ 20 categories) with some of these categories having very small numbers of observations each. To avoid problems with uneven distributions of these values between training and test data and to improve the interpretability of models, a grouping exercise was undertaken using clinical expertise informed by CET and by groupings used in the TBS score formulae. This involved consolidating 2 or more related categorical values into a new group category. For example, the categories Alcohol poisoning, Paracetamol overdose, Other drug overdose and Self-poisoning were combined into a group labelled Overdose / poisoning. Raw categorical values that represented more than 2% of the total were retained ungrouped. The resulting reductions in feature cardinality are summarised in Table A.1 in the appendix.

CET also supplied a list of maximum acceptable clinical values for the biomedical indicators in the dataset and these were used to remove improbable outliers from the data. For example, the maximum permitted value for bilirubin was set to be 500 units/litre. Full detailed can be found in Table A.4 in the appendix. Values in excess of the maximum permitted were set to null. Figure 3.1 illustrates how this step is used to remove outliers from the data. A total of 861 rows were affected by this processing step (i.e. had at least 1 value in excess of the defined maximum).

3.3 Data analysis

An analysis of the NTR dataset was carried out to gain insight into the structure of the data and to assess its suitability for predictive modelling. Factors considered included: the distribution of predictive feature; correlations between predictive features; correlations between features and dependent variables and the extent to which observations

were missing from the dataset. The analysis was carried out using Jupyter Notebooks[1] together with standard python data analysis libraries (pandas, matplotlib). The finding are summarised below.

3.3.1 Predictive features

In total there are 37 feature relating to the donor, 29 relating to the recipient and 4 others that relate to the transplant procedure itself or to interactions between the donor and recipient features. Table A.2 and Table A.3 in the appendix provide a detailed summary of each of the categorical and continuous features present respectively.

The median donor was 48 years old with a BMI of 25.2. There were approximately equal numbers of male and female donors (51% male; 49% female). Around 96% of donors were of white ethnicity with Asian, black, Chinese and mixed-race donors making up the remainder. The frequencies of specific donor causes of death are summarised in Table 3.1. It can be seen that cerebrovascular causes (e.g. strokes) are by far the most common, with trauma (e.g. road traffic accidents) and hypoxic brain damage also common. Around 12% of donors were over 65 years, the age at which a donor is considered to be in the extended criteria category. Around 88% of organs were donated after brain death (DBD) compared with only 12% following cardiac death (DCD). It was also observed that the frequency of DCD donation increased over time with relatively few DCD donors before 2006 and the number per year increasing steadily from 2007 onwards (see Figure 3.2). Steatosis was observed in around 41% of donors.

The median transplant recipient was 53 years old with a BMI of 25.8. Around 60% of transplant recipients were male. Around 87% were of white ethnicity and around 8% of Asian ethnicity. Table 3.2 summarises the most common primary indications for liver transplant. It can be seen that the two most common indications are alcoholic liver disease and cirrhosis due to hepatitis C infection (collectively around 34%). Re-transplantation (i.e. transplantation following a previous failed transplant) represent around 6% of transplants. Approximately 15% of recipients were treated as urgent at the time of transplant. Around 29% were hospital inpatients and 10% were receiving ventilation. Some degree of encephalitis was observed in approximately 34% of patients.

Other features include the year in which the transplant occurred. The temporal

[1] `https://jupyter.org/`

Description	Count	%
Cerebrovascular	6,367	66.82
Trauma	1,142	11.98
Hypoxic brain damage	1,137	11.93
Other	360	3.78
Infective	173	1.82
Malignancy	138	1.45
Cardiovascular	88	0.92
Known or suspected suicide	60	0.63
Respiratory	37	0.39
Overdose / poisoning	27	0.28

Table 3.1: Donor causes of death. Categories in bold indicate grouped categories made up of multiple related sub-causes.

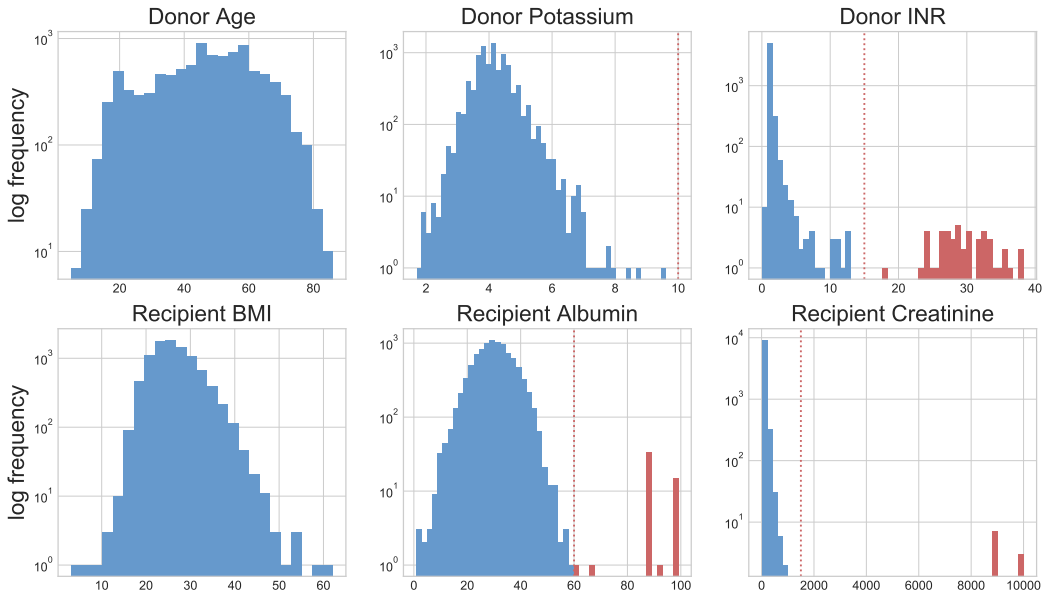

Figure 3.1: Example distributions of donor and recipient features. Red bars indicate observations in excess of defined biomedical indicator maxima (indicated by red dotted line) which were removed during data pre-processing. A log y-scale is used to highlight the existence of outliers.

trend can be seen in Figure 3.2 with an overall upward trend in the number of transplants carried out per year over time. The median cold ischemic time (the time between the chilling of the organ and the time it has blood supply restored.) was 543 minutes.

Description	Count	%
Alcoholic liver disease	1,999	20.96
Hepatitis C cirrhosis	1,219	12.78
Primary biliary cirrhosis	932	9.77
Primary sclerosing cholangitis	819	8.59
Other liver diseases	626	6.56
Retransplantation	594	6.23
Hepatocellular carcinoma - cirrhotic	513	5.38
Acute hepatic failure (serologically indeterminate)	434	4.55
Cryptogenic cirrhosis	414	4.34
Autoimmune chronic active liver disease	357	3.74
Acute Hepatic Failure (paracetamol hepatotoxicity)	317	3.32
Other metabolic liver disease	304	3.19
Non-alcoholic fatty liver disease	294	3.08
Hepatitis B cirrhosis	280	2.94
Other acute hepatic failure	267	2.80
Acute vascular occlusion (artery plus vein)	122	1.28
Other cancers	47	0.49

Table 3.2: Primary indications for liver transplant. Categories in bold indicate grouped categories made up of multiple related sub-indications.

Histogram plots of continuous factors revealed a number of different types of distributions. Figure 3.1 shows a representative sample of these. Some have a near-symmetrical distribution (e.g. age, BMI, potassium, albumin) while others have exponential-like falling distribution (e.g. INR, creatinine).

Most features had less than 10% of recorded values missing. Those with more than 10% missing are highlighted in Table A.2 and Table A.3 in the appendix. A number of donor biomedical indicators (INR, Gamma GT, AST, ALT) in particular were poorly recorded. In some cases (e.g. Donor family history of diabetes, Organ appearance), it was observed that the pattern of recording of data changed over time (see Figure 3.2). Approaches to the imputation of missing observations is addressed in section 4.1.

Correlations between donor and recipient continuous features were calculated to determine if there were significant linear dependencies. Spearman (rank-order) correlation [42] was used due to the skewed nature of some of the continuous feature dis-

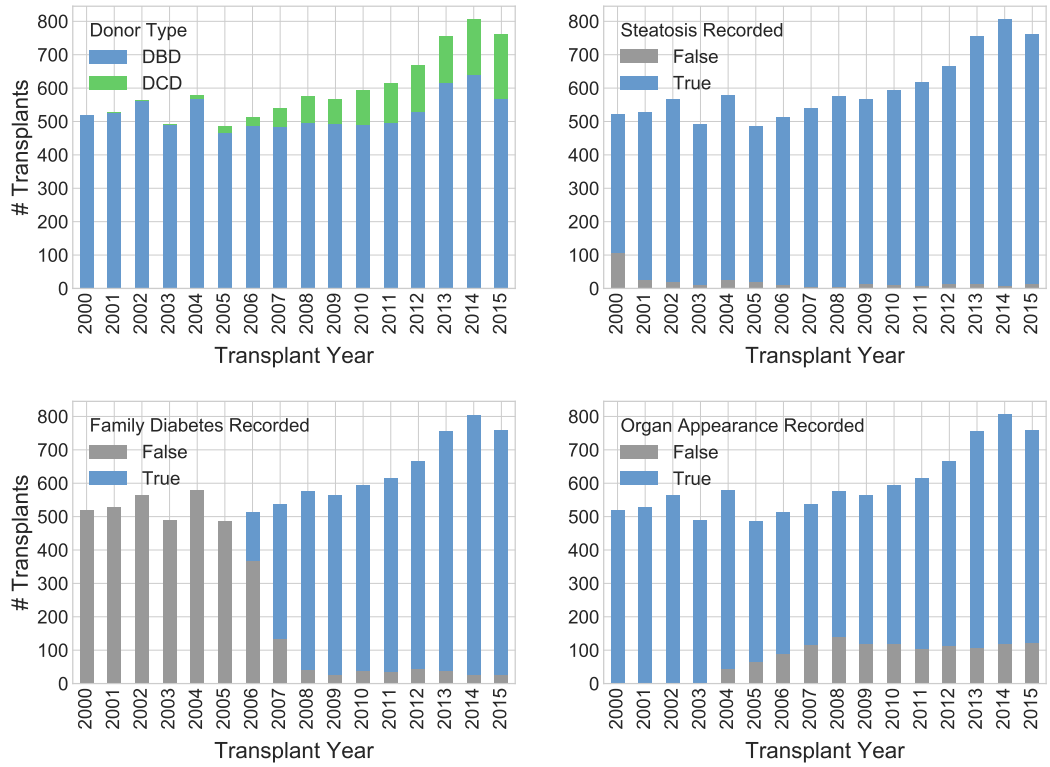

Figure 3.2: Temporal trends in recording of features. (Top left) transplants from DCD donors did not occur prior to 2004 and have increased in frequency since. (Top Right, Bottom Left, Bottom Right) changes in the frequency with which specific features are recorded over time (Steaotosis, History of Family Diabetes and Organ Appearance).

tributions and donor and recipient factors were examined separately. Table A.5 in the appendix summarises the feature pairs for which the correlation was determined to be greater than 0.25. The most significant correlations were between weight and BMI for both donors and recipients (0.7471 and 0.8477) which is to be expected. The decision was taken to retain both features however since BMI is also dependent on a 2nd-order polynomial term with respect to height i.e. $\text{BMI} = \text{weight}/(\text{height}^2)$. Weight and height were also correlated to some extent for both donors and recipients (0.4856 and 0.5461 respectively). Other correlations of note were between Donor blood urea and Donor creatinine (0.5229) and between Recipient INR, albumin and bilirubin (absolute correlation values between 0.3796-0.4734).

Correlations between categorical features were calculated using Cramér's V [43], a measure of the intercorrelation of two discrete variables. These are summarised in Table A.6 in the appendix. There is a strong correlation between Recipient urgency and Primary indication for liver transplant (0.8587) which would suggest that some

indications are considered more urgent that others. Recipient ventilation, encephalitis, urgency, inpatient status and lifestyle were also seen to be correlated with each other (0.7275-0.8101).

3.3.2 Survival data

For each D-R pair, survival data is recorded that relates to transplant and patient survival. For each survival category, the number of days is recorded along with a Boolean censoring flag that indicates whether the event occurred (e.g. the transplant failed or the patient died as a result of transplant failure) or that the recipient was censored out of the study at the recorded time (e.g. the patient could not be contacted at subsequent follow-up, died of an unrelated cause or survived beyond the end of the study). Patient survival is recorded only in cases not involving re-transplantation. Figure 3.3 (Left) shows the KM estimates of the transplant and patient survival functions $S(t)$. It can be seen that survival probability drops rapidly in the first few months so that after 1 year transplant survival is around 85% and patient survival around 90%. The curve then flattens out so that after 5 years transplant survival has dropped to around 73% and patient survival to around 79%. Figure 3.3 (Right) shows the breakdown of event status at 1 to 5 years post-transplant. In particular these figures highlights the effect of censoring on the dataset. At 1 year the effect is negligible, whilst at 5 years over 20% of the rows are censored.

Directly measuring the correlation between predictive features and observed survival time is complicated by the existence of right-censored data. Cox PH is the most commonly used regression method in this scenario and is covered in chapter 4. Another popular approach is to use KM estimators to estimate the survival curve for 2 distinct partitions within the data and then apply a log-rank test to determine the significance of any survival differences between the groups. A log-rank analysis of the transplant survival data was carried out using the lifelines[2] implementation of the KM estimator and log-rank test. For categorical features, differences in survival were compared between categories that represent at least 10% of all values and the remainder of the categories. For continuous features, the dataset was partitioned using the feature's median value. Significance was defined as a log-rank test with a p-value of 0.1 or less. A summary of all significant log-rank tests can be found in Table A.7 in the appendix.

In some cases the differences in survival between groups were in accordance with

[2] https://lifelines.readthedocs.io/

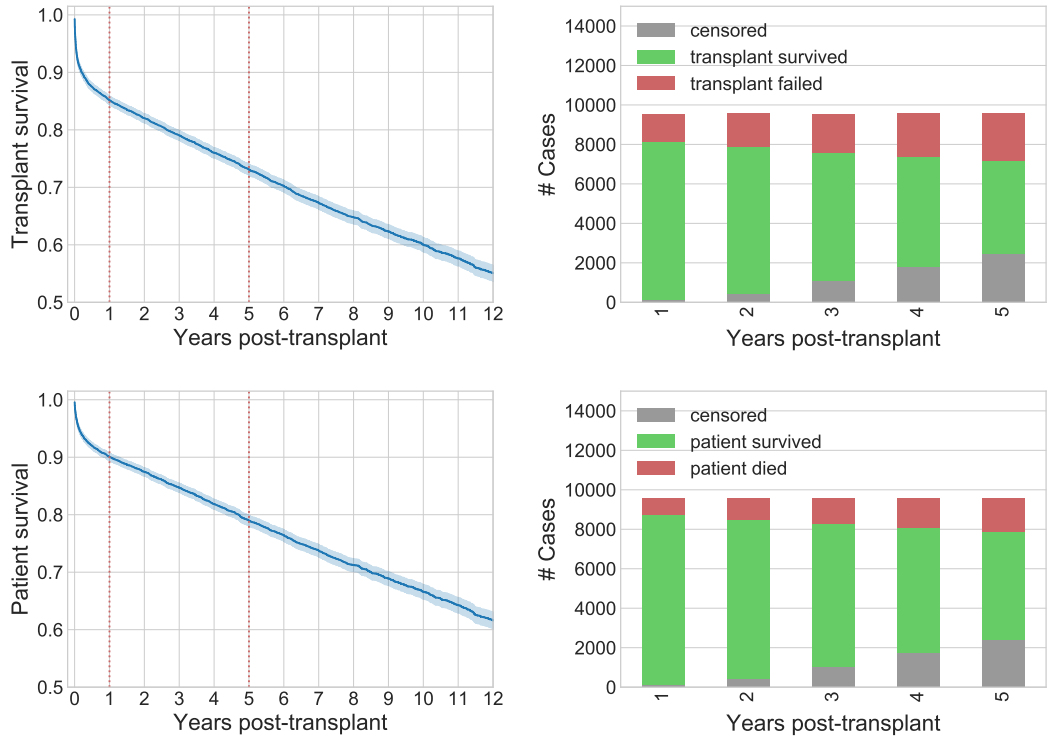

Figure 3.3: Summary of survival data. (Left) KM estimates of the transplant (Top) and patient (Bottom) survival functions $S(t)$. Red dotted lines indicate 1 and 5 years. (Right) transplant (Top) and patient (Bottom) failure event status at 1 to 5 years post-transplant.

clinical expectation. Donors with AST levels ≥ 1 (a measure of damage to liver cells) and donors above the median age were both correlated with poorer outcomes (see Figure 3.4). Similarly, recipient in-patient status, recipient lifestyle being severely limited and the presence of recipient sepsis or previous abdominal surgery were all indicative of poorer outcomes.

The reasons for other observed differences were less clear. This included the effect on outcome of transplantation year, with overall survival prospects improving significantly between 2000 and 2015 (see Figure 3.4). Despite its association with extended criteria, a donor type of DCD was also correlated with better outcomes, Similarly unexpected patterns of improved outcomes were seen for donors with histories of alcohol abuse and drug abuse and for recipients with raised levels of INR. These anomalies are addressed in section 6.1.

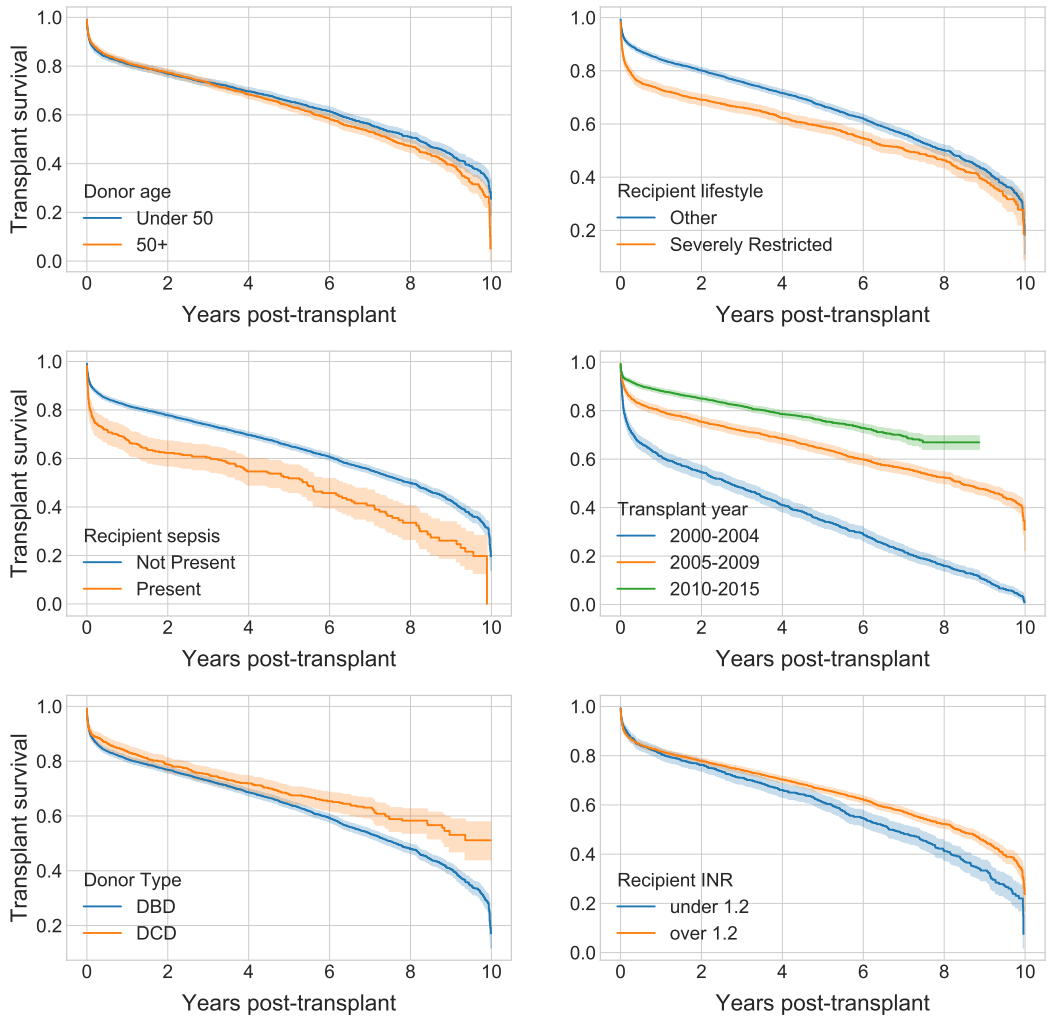

Figure 3.4: KM log-test plots showing features that reveal significant differences in survival. In some cases these are inline with clinical expectation whilst in others they are contrary to expectation. The figures also illustrate that survival differences can change over time (e.g. Donor age appears to affect long-term survival whereas recipient lifestyle is more important to early survival prospects.)

Chapter 4

Methodology

The primary objective of this project was to evaluate the effectiveness of a number of ML methods when applied to the task of predicting survival following liver transplantation. Most previous approaches have treated the task as one of binary classification at specific time points (see section 2.4). In this project, both random forests and ANN classifiers were evaluated. Random forests were selected due to their popularity and relative interpretability whilst ANNs were chosen as being the most commonly used approach in previous work and having demonstrated state-of-the-art performance in some studies reviewed. Two ML approaches that specifically target survival analysis tasks were also evaluated. These were Random survival forests (RSFs) and Neural Multi-task Logistic Regression (N-MTLR). Both of these approaches were introduced in section 2.2.

All experimentation was carried out using Jupyter Notebooks running on a single 8-core Linux compute instance hosted on the Google Cloud Platform (GCP). Implementation details of each of the model are provided inline.

4.1 Data preparation

The dataset used for this project and the initial preprocessing and data cleaning steps carried out prior to modelling are described in chapter 3. The cleaned dataset was split into separate training and test datasets with 80% of the data assigned to the training dataset and 20% to the test dataset which was held out for final model evaluation. Tuning of model hyperparameters and comparison of dataset variants (e.g. different imputation strategies) were carried out using a 5-fold cross-validation process across the training dataset. For classification tasks, stratified fold allocation [44] was used to

insure that similar numbers of each class label were present in the folds. In a number of cases it proved difficult to precisely determine optimal hyperparameter values in some cases due to the small differences in between values when compared to the standard deviation of the cross-validation estimates. In these cases, the mid-point of a range of similar values was typically selected. Based on the optimally selected hyperparameters, the final models were then re-trained using the full training dataset and evaluated using the held-out test dataset.

Where models were unable to handle categorical input features directly, one-hot encoding was used to create a separate binary feature for each category present. Where models were unable to handle null observations a number of imputation strategies were employed. For categorical variables, missing observations were assigned to an 'Unknown' category. For continuous variables, a number of imputation strategies were compared on a model-by-model basis. The simplest method used was to impute null values using either the mean or median value of the non-null values in the column. The k-nearest neighbours (k-NN) imputation approach [45] finds the nearest neighbours for a missing observation based on the mean squared distance of other non-null continuous features and then imputes the missing value from the average of the values obtained from those neighbours. Iterative imputation [46] is an approach that fits multiple Bayesian ridge regression models using non-null observations and uses these models to iteratively impute the missing values. In both of these cases, separate imputation steps were carried out for donor and recipient features due to a strong assumption of independence between the two sets of features. The k-NN and iterative implementations provided by the fancyimpute[1] python library were used.

To support binary classification approaches, target failure/survival labels were created at 1 and 5 years for both transplant and patient survival. Rows censored prior to the time of interest were not assigned a label and were excluded from training and evaluation. At 1 year, the amount of censoring was small (around 1%) but at 5 years, the amount of censoring was significant at around 25%. The failure classes of interest represent between 10% and 20% of labels at 1 and 5 years meaning the classes are imbalanced. This is known to be problematic for binary classifiers which can tend to over-classify the majority class so the effects of class rebalancing were evaluated for the classifier models.

[1] `https://github.com/iskandr/fancyimpute`

4.2 Evaluation metrics

The following evaluation metrics were used to compare the predictive ability of the models developed:

Area under the receiver operating characteristic curve (AUC ROC): The receiver operating characteristic curve is a plot of the true positive rate (TPR) versus the false positive rate (FPR) for the predictions of a binary classifier at multiple thresholds. The integrated area under the curve (AUC ROC) provides a summary measure of the discriminative ability of the model across all evaluated thresholds. Because performance is evaluated at multiple thresholds, AUC ROC is less sensitive to imbalanced class distributions than other commonly reported metrics (e.g. Accuracy, TPR, True Negative Rate (TNR)) making it suitable for this task [47]. The range of AUC ROC values is between 0.5 and 1.0 with a value of 0.5 representing a classifier that is no better that randomly guessing the class and a value of 1.0 signifying a classifier with perfect discriminative ability.

Concordance Index (C-index): The concordance index [48] measures a model's ability to provide a reliable ranking of survival time based on predicted risk. The value is calculated by considering all viable pairs and comparing their relative risk score with their actual survival time. Unlike the AUC ROC metric, it is able to account for right-censored data. The C-index can be calculated using the formula

$$\text{C-index} = \frac{\sum_{i,j} 1_{T_j<T_i} \cdot 1_{\eta_j<\eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j} \tag{4.1}$$

where η_i is the risk score for observation i, T_i is the survival time for observation i, $1_{T_j<T_i}$ is 1 if $T_j < T_i$ and 0 otherwise (similarly for $1_{\eta_j<\eta_i}$) and δ_i is the censoring indicator for observation i. As with AUC ROC the range of values is between 0.5 and 1.0 with 1.0 signifying perfectly ranked predictions.

Brier score: The Brier score [49] calculates the mean squared error between predicted probabilities and the target values. As such, it can be considered another measure of a model's discriminative ability. The score summarizes the magnitude of the error in the probability forecasts. The range of values is between 0.0 and 1.0 with a score of 0.0 representing perfect predictions.

4.3 Liver transplant scores

In order to compare the performance of the ML models developed with current clinical practice, a number of liver transplantation scores were implemented and the scores calculated using the test dataset were used to produce evaluation metrics. Both scores consist of weighted linear combinations of factors or factor interactions.

The donor liver index (DLI) scores (see section 2.3) were implemented using the formulae provided in the original paper (see section D.1). The DLI1 and DLI scores were used as comparison baseline for the performance of donor-only models at 1 year and 5 years respectively.

A set of formulae for the TBS were provided by NHSBT and additional details of the implementation were established from the implementation provided by Surgical Informatics[2]. The TBS model consists of separate cancer and non-cancer scores for both post-transplant and waiting list survival. For this study only the post-transplant parts of the score were implemented since these are the parts relating to post-transplant survival prediction. Specifically the log-risk parts of the post-transplant scores (which are indicative of relative risk) were implemented (see section D.2). Selection of the cancer or non-cancer formulae was based on whether cancer was amongst the indications for liver transplant. A number of recipient features present in the TBS score were missing from the supplied NRT dataset. These were waiting list time, diabetes history and a number of cancer-specific features (size of tumour, number of tumours, max Alpha-Fetoprotein (AFP)). These components of the score were omitted.

The log-risk TBS score was used as a comparison baseline for the performance of donor/recipient models at 1 year and 5 years respectively. Due to the large number of missing factors for the cancer score, the TBS score was only evaluated for non-cancer D-R pairs. Equivalent non-cancer metrics were also evaluated for the corresponding ML models.

4.4 Baseline Cox proportional hazards models

Cox PH models were also fitted to the NRT dataset to provide an alternative performance baseline for the ML models developed. The Cox PH implementation provided by the lifelines[3] python library was used. The fitting process requires numeric non-null

[2] http://transplantbenefit.org

[3] https://lifelines.readthedocs.io/

features so one-hot encoding of categorical features and imputation of missing continuous fields was used. Median, mean, k-NN and iterative imputation approaches were compared and models fitted using median imputation were determined to have the best performance. Though not strictly required by the model, continuous features were scaled to be between 0 and 1 to enhance the interpretability of the model coefficients. The Cox PH implementation used supports a L2 regularisation (a loss function penalty based on the squared magnitude of the model coefficients that favours less complex models). A search for L2 hyperparameter values between 0 and 0.1 was carried out but no significant difference in model performance was found and the default values of 0.0 was retained.

Using the optimal model hyperparameters Cox PH models were fitted using the transplant and patient survival data and the predicted survival probabilities at 1 and 5 years were computed for the test dataset respectively. These survival probabilities were then used to calculate performance metrics for the models.

Each feature used in the Cox PH model has an associated coefficient that provides an indication of the feature's importance in the model and whether it is associated with lesser or greater risk. The model also returns a standard error value for each coefficient allowing p-values to be calculated for the null hypothesis that the true coefficient value is 0 (which would indicate that the feature has no effect on the risk). Models were fitted to the entire dataset (i.e. training and test) for both transplant and patient survival data and model coefficients were recorded where the associated p-value was ≤ 0.025 (equivalent to a 95% confidence interval).

4.5 Random forest models

Two distinct implementations of random forests were initially considered for the classification task. The first is the implementation provided by the popular scikit-learn[4] python package. Though widely used the scikit-learn implementation is unable to directly handle categorical features without the application of one-hot encoding. In contrast, the implementation provided by the h2o[5] java and python package supports categorical features directly by using an implementation based on the construction of categorical bins that are used during splitting. Both implementations share the following common hyperparameters: the number of trees to be trained, the maximum depth

[4] `https://scikit-learn.org/`
[5] `https://www.h2o.ai/`

that a single tree can grow to and the minimum number of rows that a leaf node can contain. A initial grid search over these hyperparameters was performed for both implementations using the 1 year transplant survival labels. The test dataset performance of the optimal h2o model was found to be marginally better than that of the scikit-learn model and had additional benefits in terms of interpretability of the model due to its use of categorical features directly. On this basis, the h2o implementation was selected for all subsequent random forest modelling.

Separate hyperparameter tuning was carried out for each of the classification tasks. Though the h2o model is able to handle null values directly the effect of different imputation strategies were compared (median, mean, k-NN, iterative) but did not provide any performance benefit so no imputation was used. The h2o random forest implementation provides the ability to specify weights for each class allowing over-sampling of the minority failure class during the bootstrap sampling of data used to build each tree. The performance of models trained with over-sampling between $1\times$ and $3\times$ were evaluated. The selected hyperparameters are summarised in Table B.1. Using the optimised hyperparameters, models were fitted on the full training dataset and performance metrics were calculated for the model using the test dataset.

To calculate feature importance for the random forest models, separate models were trained using optimal hyperparameters and the full dataset (training and test dataset combined) and the individual feature importance scores for each model were recorded.

4.6 Artificial neural network (ANN) models

An ANN classifier model was implemented that consisted of a fully connected multilayer perception (MLP) architecture with ReLU activations between hidden layers. The final layer had 2 outputs representing the probabilities of the survival and failure classes. Dropout [21] and batch normalisation [50] were applied between layers to regularise the model and prevent overfitting. A cross-entropy loss function was used to train the model using back-propagation. ANNs are sensitive to input scale and are unable to handle missing values so continuous features were normalised to have 0 mean and a standard deviation of 1 and missing values were imputed (see below). Categorical features were mapped to lower-dimensional continuous representations using an embedding layer. In this approach, the weights of the embedding layer are learnt during training allowing similarities between categories to be captured in contrast to

one-hot encoding which assumes independence between categorical dimensions. The outputs from the embedding layer are concatenated with the continuous feature inputs and fed into the first hidden layer of the model. The model was implemented using the PyTorch[6] and fastai[7] ANN python libraries.

The model has many hyperparameters including hidden layer count, layer sizes, dropout probabilities, learning rate and batch size, making an exhaustive grid search impractical. Instead, single or small groups of parameters were tuned individually for each type of model, with the optimal values from each stage fed into the next. Using 5-fold cross validation, each model was trained for 50 epoch with the model from the epoch with the lowest validation loss being selected and the average of the AUC ROC scores for the best epochs being used to compare variants. An Adam [22] optimiser was used to train the model using mini-batch gradient descent.

An initial search was made of different model architectures and dropout probabilities. Models with 1 to 3 hidden layers were considered with 40 to 400 units in the initial hidden layer and the numbers of units in subsequent layers decreasing by approximately 50% each time. For the hidden layers dropout probabilities between 0.2 and 0.8 were evaluated. In the next stage, the effect of median, mean, k-NN and iterative imputation approaches were compared, with median imputation being selected. The default mini-batch size and learning rate parameters were also fine-tuned. Finally the effect of class rebalancing was assessed. The effect of both down-sampling of the majority survival class (i.e. discarding rows randomly) and over-sampling the minority failure class (randomly re-sampling some of the rows) were both evaluated. The selected hyperparameters are summarised in Table B.2. Optimised models were fitted on the full training dataset and performance metrics were calculated for the model using the test dataset.

4.7 Random survival forest (RSF) models

An implementation of RSF provided by the pysurvival library[8] was used. To reduce the number of times for the model to predict, survival times were mapped to the nearest 10th of a year. Categorical features were one-hot encoded and continuous features were imputed using the median strategy. As with the random forest model a grid

[6] https://pytorch.org/
[7] https://www.fast.ai/
[8] https://square.github.io/pysurvival/

search of the main model hyperparameter was carried out: the number of trees to be trained, the maximum depth that a single tree can grow to and the minimum number of rows that a leaf node can contain. The selected hyperparameters are summarised in Table B.3. RSF models were then fitted using the full transplant and patient survival training dataset and performance metrics for these models were calculated using the test dataset.

4.8 Neural multi-task logistic regression (N-MTLR) models

An implementation of N-MTLR provided by the pysurvival library[9] was used with SELU [23] activations between layers. Categorical features were one-hot encoded and continuous features were imputed using the median strategy. The model has many hyperparameters including the number of layers and their sizes, learning rate, dropout and L2 regularization. As with the ANN model, a series of grid searches were carried out on successive subsets of hyperparameters. Firstly initial learning rate, dropout value and number of time windows to predict were established. Next the number of and size of the hidden layers was optimised with models with 1-2 layers of up to 160 units considered. Finally the L2 regularization parameter was optimised. In general, optimisation of the N-MTLR model proved more difficult that with the other models due to unpredictable gradient explosion problems during training. The selected hyperparameters are summarised in Table B.4. N-MTLR models were then fitted using the full transplant and patient survival training data. Models were trained for up to 3000 epochs and the epoch with the best mean C-index value was selected. Performance metrics for these models were then calculated using the test dataset.

[9] https://square.github.io/pysurvival/

Chapter 5

Results and discussion

5.1 Feature importance and model interpretability

Log-rank tests (see section 3.3), Cox PH models and random forest models (see section 5.1) all provide means by which to rank the importance of different features in terms of their value in predicting survival outcomes. These measures can be used to interpret a model's behaviour to an extent and allow clinicians to better understand the causal factor at play. As such they represent a significant advantage for such models in a clinical context.

In the case of log-rank tests it is important to note that the analysis is univariate and therefore unable to account for confounding factors within the data (i.e. features that influence both the predictive feature and the dependent variable). For example, contrary to expectation, in a log-rank test, DCD donors were observed to have better survival outcomes (see Figure 3.4). As highlighted in Figure 3.2 though this is likely due to the greater prevalence of DCD donors in more recent years which in turn are associated with better overall survival. All 3 approaches identified the transplant year as being a highly important feature. The reasons for this strong correlation remain unclear.

For the Cox PH model, parameters greater than 0 are indicative of increased risk and consequently, reduced survival probability whereas parameters less than 0 are indicative of reduced risk and improved survival. The values of the parameters for which a p-value of ≤ 0.025 was obtained are recorded in the appendix in Table C.1 and Table C.2. The main recipient factors associated with increased risk are specific indications for liver transplant: retransplantation, cancer, non-alcoholic fatty liver disease, hepatitis C and alcoholic liver disease. Other positive recipient factors include the use

of ventilation, the presence of encephalitis and sepsis, Cytomegalovirus (CMV) infection and a severely restricted lifestyle. Donor factors associated with increased risk include DCD donors and donors who died as a results of suicide. Other significant positive factors include the transplantation of a reduced liver. Only a small number of factors were associated with reduced risk: The transplant year (i.e. risk reduces over time), the recipient being flagged as urgent, the presence of recipient oesophageal varices, and increased donor potassium.

For the random forest model, the mean of the % feature importance values taken from the 4 donor/recipient models (i.e. patient and transplant survival at 1 and 5 years) are recorded in in the appendix in Table C.3 where the value is greater than 1%. The importance value is indicative only of the absolute effect of the feature on the model and doesn't indicate whether the feature is negatively or positively correlated with survival. Important donor features include: cause of death (6%), donor age (3%), Gamma GT, albumin and potassium. The following recipient features were observed to be of most importance: Primary and secondary indication for liver transplant (14% and 4% respectively), age (3%), infection with CMV, recipient lifestyle, creatinine, bilirubin and potassium. Of the other features, transplant year is the 2nd most important feature overall (7%) and cold ischaemic time is also important (2%).

Both the Cox and the RF models identified the following donor features as important: cause of death, age, and INR. Similarly the following recipient features were identified as important by both models: age, primary and secondary indication for transplant, creatinine, bilirubin, INR, lifestyle, the use of renal support prior to transplant, and the presence of encephalitis or oesophageal varices. Transplant year and cold ischemic time were also identified as important by both models. Despite being present in the formula for either the DLI or TBS scores (or both), the following donor features were not identified by either of the models as being important: sex, ethnicity, smoking history, steatosis and bilirubin. Similarly for recipient factors sex, in-patient status and the presence of ascites were not identified as important by either model. Of these, the absence of steatosis is perhaps the most surprising given its status as an extended donor criteria.

5.2 Model performance

Four ML approaches were considered in this study. For each approach, separate models were trained using donor only and donor/recipient features. For each approach and

Model	Surv. Type	Year	AUC ROC	C-index	Brier score
DLI	Transplant	1	0.5450	0.5273	-
Cox PH	Transplant	1	0.5573	0.5397	0.1468
RF	Transplant	1	0.5826 ± 0.0057	0.5549 ± 0.0026	**0.1271 ± 0.0001**
ANN	Transplant	1	**0.5918 ± 0.0085**	0.5599 ± 0.0072	0.1267 ± 0.0004
RSF	Transplant	1	0.5879 ± 0.0101	**0.5627 ± 0.0046**	**0.1271 ± 0.0002**
N-MTLR	Transplant	1	0.5673 ± 0.0075	0.5476 ± 0.0049	0.1293 ± 0.0010
DLI	Transplant	5	0.5499	0.5439	-
Cox PH	Transplant	5	0.5470	0.5397	0.2847
RF	Transplant	5	0.5328 ± 0.0076	0.5194 ± 0.0054	0.2114 ± 0.0003
ANN	Transplant	5	0.5531 ± 0.0134	0.5395 ± 0.0121	**0.2104 ± 0.0007**
RSF	Transplant	5	**0.5569 ± 0.0097**	**0.5572 ± 0.0063**	0.2119 ± 0.0005
N-MTLR	Transplant	5	0.5509 ± 0.0146	0.5499 ± 0.0097	0.2145 ± 0.0018
DLI	Patient	1	0.5501	0.5211	-
Cox PH	Patient	1	0.5701	0.5405	0.0988
RF	Patient	1	0.6242 ± 0.0130	0.5698 ± 0.0082	0.0895 ± 0.0002
ANN	Patient	1	0.6177 ± 0.0201	0.5618 ± 0.0130	**0.0891 ± 0.0003**
RSF	Patient	1	**0.6569 ± 0.0044**	**0.5860 ± 0.0063**	**0.0891 ± 0.0001**
N-MTLR	Patient	1	0.5864 ± 0.0151	0.5493 ± 0.0115	0.0902 ± 0.0009
DLI	Patient	5	0.5537	0.5485	-
Cox PH	Patient	5	0.5404	0.5405	0.2308
RF	Patient	5	0.5515 ± 0.0151	0.5307 ± 0.0093	0.1811 ± 0.0008
ANN	Patient	5	0.5608 ± 0.0128	0.5492 ± 0.0094	**0.1802 ± 0.0005**
RSF	Patient	5	**0.5798 ± 0.0100**	**0.5827 ± 0.0077**	0.1808 ± 0.0003
N-MTLR	Patient	5	0.5615 ± 0.0153	0.5606 ± 0.0161	0.1836 ± 0.0023

Table 5.1: Performance metrics for baselines and models trained on donor only features. For each grouping and metric, the best model is highlighted in bold.

feature set, predictions of survival at 1 and 5 years was assessed for both transplant and patient survival data. For both donor only and donor/recipient feature sets, equivalent baseline Cox PH models were also trained and evaluated. For the donor only model the DLI and DLI1 scores were used as additional baselines and for donor/recipient models, the TBS score was used as an additional baseline. Due to limitations with the implementation of the cancer versions of the TBS score (see section 4.3), donor/recipient models were trained and evaluated only for non-cancer recipients. The results of these evaluations are summarised in Table 5.1 and Table 5.2 respectively. The AUC ROC and C-index metrics are also visualised in Figure 5.1 and Figure 5.2 respectively and

Model	Surv. Type	Year	AUC ROC	C-index	Brier score
TBS	Transplant	1	0.5698	0.5468	-
Cox PH	Transplant	1	0.6320	0.5920	0.1450
RF	Transplant	1	0.6303 ± 0.0064	0.5732 ± 0.0037	**0.1167 ± 0.0004**
ANN	Transplant	1	**0.6507 ± 0.0081**	0.5855 ± 0.0072	0.1226 ± 0.0005
RSF	Transplant	1	0.6387 ± 0.0061	**0.5927 ± 0.0053**	0.1239 ± 0.0003
N-MTLR	Transplant	1	0.6225 ± 0.0136	0.5660 ± 0.0117	0.1327 ± 0.0031
TBS	Transplant	5	0.5581	0.5478	-
Cox PH	Transplant	5	0.5534	0.5920	0.2722
RF	Transplant	5	**0.6151 ± 0.0042**	0.5599 ± 0.0031	**0.1955 ± 0.0005**
ANN	Transplant	5	0.5749 ± 0.0157	0.5711 ± 0.0115	0.2043 ± 0.0012
RSF	Transplant	5	0.5652 ± 0.0059	**0.5984 ± 0.0050**	0.2061 ± 0.0004
N-MTLR	Transplant	5	0.5234 ± 0.0114	0.5610 ± 0.0098	0.2237 ± 0.0072
TBS	Patient	1	0.5552	0.5347	-
Cox PH	Patient	1	0.5888	0.5718	0.1063
RF	Patient	1	0.6345 ± 0.0074	0.5581 ± 0.0056	**0.0832 ± 0.0003**
ANN	Patient	1	**0.6486 ± 0.0141**	**0.5819 ± 0.0082**	0.0940 ± 0.0005
RSF	Patient	1	0.6229 ± 0.0136	0.5797 ± 0.0079	0.0950 ± 0.0003
N-MTLR	Patient	1	0.6257 ± 0.0111	0.5678 ± 0.0087	0.0991 ± 0.0039
TBS	Patient	5	0.5520	0.5342	-
Cox PH	Patient	5	0.5354	0.5718	0.2224
RF	Patient	5	**0.6045 ± 0.0055**	0.5513 ± 0.0047	**0.1672 ± 0.0003**
ANN	Patient	5	0.5504 ± 0.0178	0.5496 ± 0.0202	0.1767 ± 0.0013
RSF	Patient	5	0.5509 ± 0.0101	**0.5881 ± 0.0096**	0.1782 ± 0.0005
N-MTLR	Patient	5	0.5322 ± 0.0155	0.5606 ± 0.0095	0.1888 ± 0.0022

Table 5.2: Performance metrics for baselines and models trained on donor and recipient features (non-cancer recipients only). For each grouping and metric, the best model is highlighted in bold.

the shapes of ROC curves for the transplant survival donor/recipient models and Cox PH baselines can be seen in Figure 5.3.

When assessing classifier models (random forest, ANN) the predicted class probabilities were used to calculate performance metrics. For the survival models (Cox PH, RSF and N-MTLR) predicted failure probabilities (i.e. $1 - S(t)$) at 1 and 5 years were calculated for the models and these were then used to derive performance metrics. For the DLI1, DLI and TBS scores only AUC ROC and C-index metrics were calculated since the Brier score is only defined for probability values between 0.0 and 1.0. Con-

fidence intervals were estimated for each of the ML models by training the model 10 times with different initial random state.

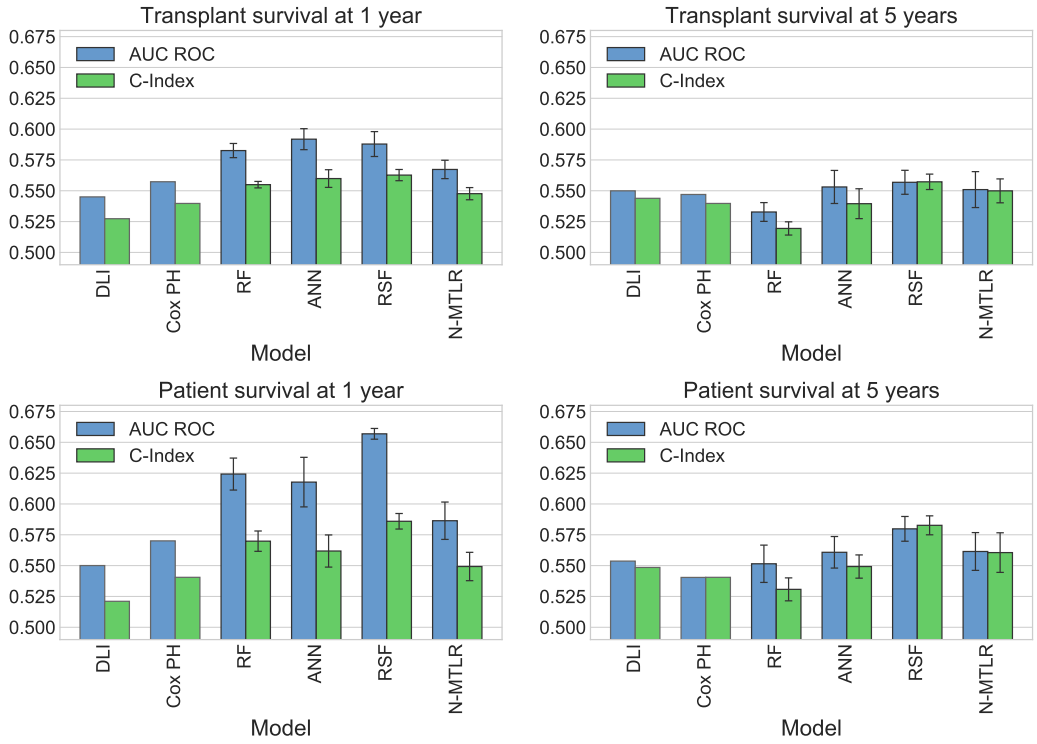

Figure 5.1: AUC ROC and C-index metrics for baselines and ML models trained using donor only features.

For the donor only models the RSF model proved to have the best overall relative performance both in terms of AUC ROC and C-index (the only exception being for transplant survival at 1 year where the ANN model outperformed it on the AUC ROC metric). Brier scores proved to be very similar for the random forest, ANN and RSF models respectively which all performed better on this metric than the Cox PH baseline. In all cases the best ML model outperformed the Cox PH model by between 2% and 15% for the AUC ROC metric and between 3% and 8% for the C-index metric. With respect to the DLI score baseline, the best ML model outperformed by between 1 and 20% for the AUC ROC metric and 2% and 12% for the C-index metric. The DLI score itself performed well and actually outperformed the baseline Cox PH model in several cases.

For the donor/recipient models the random forest and ANN models performed best in terms of the AUC ROC metric with the ANN model being best for predictions at 1 year and the random forest models best at 5 year survival prediction. In contrast the RSF models were largely superior when evaluated on the C-index metric (with the

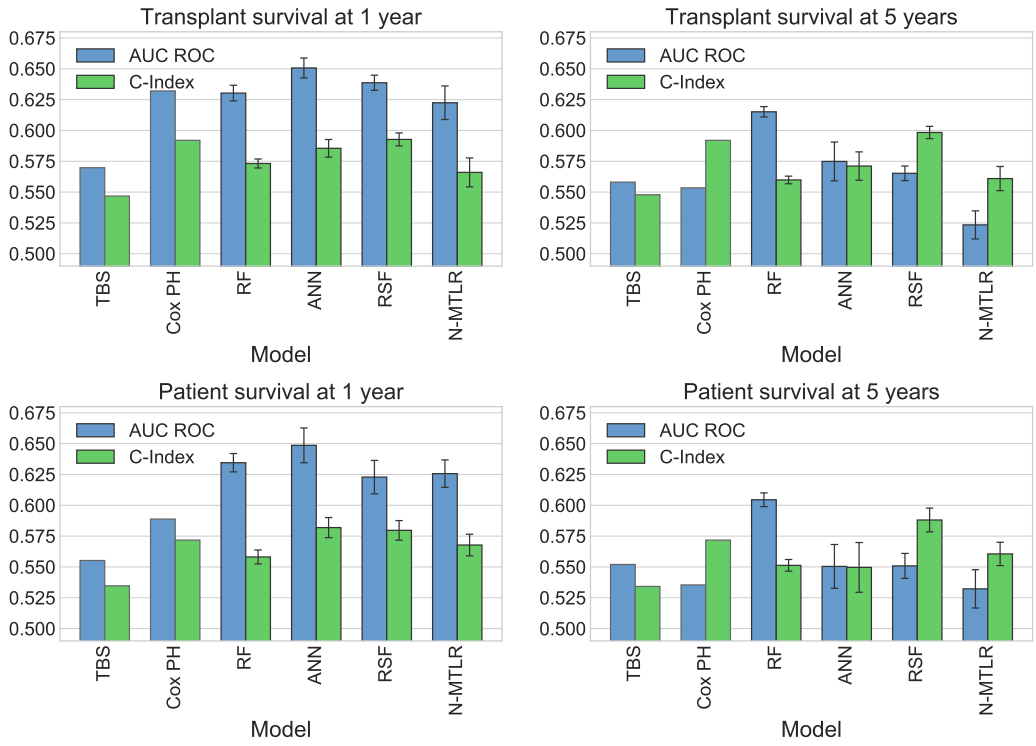

Figure 5.2: AUC ROC and C-index metrics for baselines and ML models trained using donor and recipient features (non-cancer recipients only).

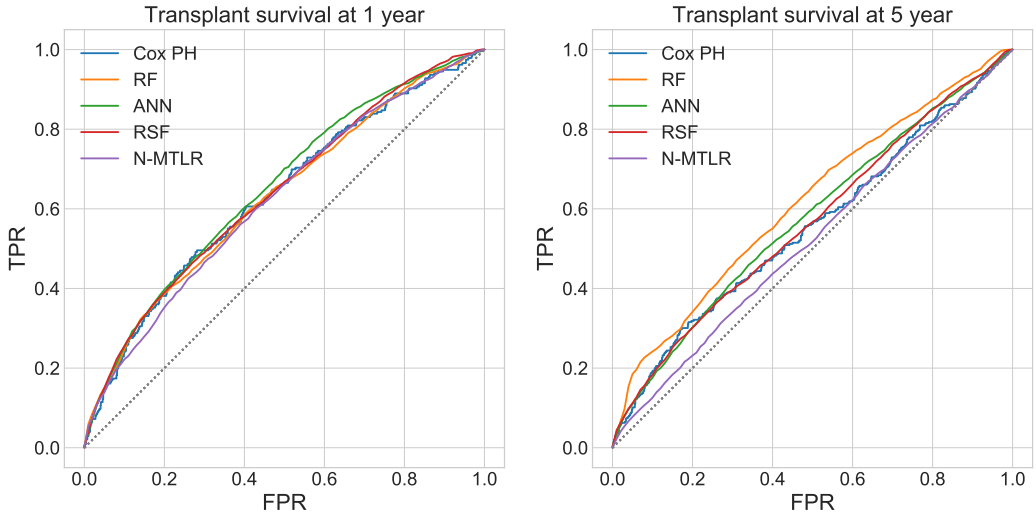

Figure 5.3: ROC curves for Cox PH baseline and ML models trained using donor and recipient features (transplant survival at 1 and 5 years, non-cancer recipients only).

exception of patient survival at 1 year where the ANN model was marginally better). Here, the random forest model consistently performed best in terms of Brier scores. Again, in all cases, the best ML model outperformed the Cox PH baseline model

(though in some cases only just). Here the relative performance improvements over the Cox PH baseline were between 3% and 13% for the AUC ROC metric and 0.1% and 10% for the C-index metric. With respect to the TBS score baseline, the best ML model outperformed it by between 10% and 17% for the AUC ROC metric and 8% and 10% for the C-index metric.

Looking at absolute performance, for the donor only models, the best AUC ROC values obtained were in the range 0.56-0.66 and for the donor/recipient models, the range was 0.60-0.65 suggesting that overall the models are relatively poor classifiers. These values are also low when compared to the results of some of the studies reviewed in section 2.4. It is important to note however that it is difficult to make direct comparisons between the predictive performance of models trained on different features and datasets and that many of the studies reviewed looked at survival over much shorter time scales (e.g. 3 months). Within these results it was observed that predictive accuracy at 1 year was better than at 5 years and that the relative performance of the ML models to their baselines was also better at 1 year. Looking at relative performance against baselines overall, the models evaluated here exhibited a fairly wide range of improvements over the score-based baselines of between 1% and 20%, less than the improvements of 30% or more seen in some studies.

Both of the score-based baselines performed better against the Cox PH baselines for 5 year prediction than at 1 year. In the case of the DLI scores, this may be due to the distinct formulae used for 1 and 10 year survival. It is also notable that the TBS score metrics were generally not much better than those for the DLI scores despite being derived using significantly more predictive features.

During model tuning none of the more complex imputation strategies evaluated (k-NN, iterative imputation) were observed to outperform simple median imputation or null imputation. This is somewhat surprising though is possibly explained by the relatively small correlation between continuous donor and recipient factors (see section 3.3).

5.3 Individual survival predictions

Both the Cox PH and RSF models have the ability to produce complete survival curve predictions for a single individual which can be used to estimate total survival benefits for an individual.

Figure 5.4 illustrate the predicted survival curves for 5 randomly selected individ-

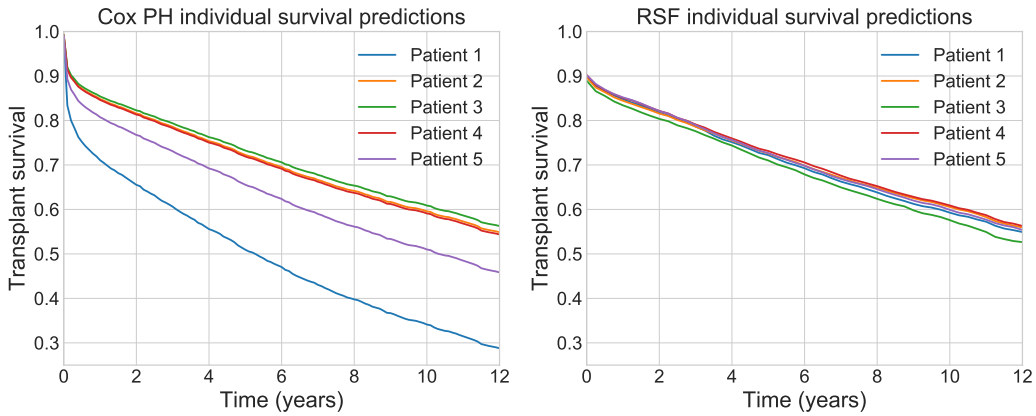

Figure 5.4: Individual survival function ($S(t)$) predictions for Cox PH and RSF models. Each line shows the predicted survival function for a randomly selected individual from the test dataset.

uals. The figure highlights the proportional nature of the Cox PH model in that the survival curves never cross. Conversely, we can see that for the RSF models, the survival curves are not proportional and can cross one another, allowing for more complex individual survival functions to be fitted.

Chapter 6

Conclusions

In this project the task of predicting survival following liver transplantation was investigated. In order to interpret the dataset provided, a working knowledge of the fields of liver disease and transplantation were first gained by reviewing relevant literature and as a result of discussions with domain experts at CET. Reviews of survival analysis literature in general and of literature relating specifically to liver transplantation survival were then carried out. This established how such survival prediction tasks are currently approached allowing baseline models to be established for the work. The review also identified a number of relevant ML approaches that formed the core of this study. For these models hyperparameters were extensively tuned along with a number of data processing variants including the imputation of missing data and the rebalancing of classes. In practice model tuning proved to be difficult due to high variance in the results of cross-validation comparisons.

We saw that ML approaches to predicting survival following a liver transplant can consistently outperform both liver transplant scores and traditional Cox PH models though the relative increases in performance were more modest than we might have initially hoped based on the results of previous studies. Both random forest and ANN classifiers frequently outperformed Cox PH baseline models suggesting that there is a benefit to be gained from learning the non-linear interactions between features. Overall though the most consistent performer was the RSF model which did particularly well when measured using the C-index metric. C-index is a measure of a model's ability to rank cases in order of eventual survival time and is therefore arguably the most relevant metric where the eventual goal is the ranking of different organ allocation scenarios. The superior performance of the RSF can potentially be attributed to its ability to learn from right censored data which are omitted when training the classifier

models. The models produced by the RSF are also richer than those produced by classifier models in that individual survival curve predictions (see Figure 5.4) can be made. Furthermore, since the RSF model is based on the architecture of a random forest, it also potentially has the ability to provide model interpretability in the form of feature importance rankings.

Despite being derived from a relatively large number of available factors (including 2nd order interactions between factors), the TBS score performed quite poorly when compared to equivalent Cox PH models. The 2 missing factors from the TBS score implementation used (see section 4.3) could be in part responsible to this underperformance (though neither factor was available to the equivalent Cox PH models either). Another possible explanation could be that the distribution of predictive factors and their relation to survival risk changes over time (a phenomenon known as *concept drift*) meaning that the weights defined are less accurate when used to predict survival in years other that those used to derive the model. This highlights an important limitation with static score based methods in terms of their applicability to populations and eras different to those for which they have been calibrated. In contrast, ML models can be relatively easily retrained using the most appropriate (e.g. most recent) data.

6.1 Future work

Some of the correlations observed between predictive features and survival time remain unexplained and further investigation here could be of value. In particular it would seem to be important to better understand the reasons for the strong correlation observed between transplant year and better survival outcomes.

Model tuning was complicated by the fact that metrics produced by cross-validation had high variance making the selection of optimal values difficult. In order to increase the effective size of training datasets and reduce variance, bootstrap sampling [51] might be used instead in future evaluations.

The success of the RSF model in this project suggests that ML approaches that are able to account for right censored data can improve predictive performance. The N-MTLR model used also potentially has this ability but for the most part performed relatively poorly. This may be due in part to problems encountered whilst tuning hyperparameters for this model. As such, it is anticipated that further time spent stabilising and tuning this model might result in improved overall performance. Evaluations of other ANN-based approaches to survival analysis may also be worthwhile (e.g. Deep-

Surv [20])

A number of features used in the TBS score were missing from the TBS implementation used in the project meaning that comparisons could only be made for non-cancer recipients and even then using an incomplete version of the score. As such, it would be useful to repeat the evaluation of the full TBS score using a more complete NRT dataset. It would also be interesting to contrast this with a variant TBS score based on the survival curve predictions produced by an RSF model.

Feature selection (i.e. building models using only a sub-set of the most important features) was used in a number of the studies reviewed in section 2.4 [35, 36, 41] and could be another worthwhile line of inquiry.

Appendix A

Supplementary data analysis tables

Feature	Pre-grouping	Post-grouping
Donor cause of death	41	12
Primary indication for liver transplant	57	17
Secondary indication for liver transplant	42	13
Tertiary indication for liver transplant	20	5

Table A.1: Summary of reduction in cardinality of high-cardinality features following grouping of related categories

Source	Feature	Cardinality	% null
D	Donor type	2	0.00
D	Donor gender	2	0.01
D	Donor blood group	4	0.00
D	Donor cause of death	12	0.25
D	Donor CMV status	2	1.71
D	Donor ethnicity	6	0.42
D	Donor history of alcohol abuse	2	4.08
D	Donor cardio disease	2	4.26
D	Donor diabetes	2	3.25
D	Donors family history of diabetes	2	**41.36**
D	Donor history of drug abuse	2	4.35
D	Donor hypertension	2	4.03
D	Donor liver disease	2	4.35
D	Donor smoker	2	3.53

D	Donor tumour	2	3.41
D	Donor received inotropes	2	0.00
D	Donor respiratory arrest	2	4.25
D	Donor cardiac arrest	2	2.92
D	Liver steatosis	2	3.24
D	Degree of steatosis	4	3.44
D	Capsular damage	2	3.92
D	Organ appearance	2	**13.20**
O	Organ transplanted	3	0.10
O	Donor to recipient blood group match	2	0.00
R	Recipient urgency status	2	0.00
R	Recipient gender	2	0.01
R	Recipient blood group	4	0.00
R	Recipient CMV test result	2	**14.40**
R	Recipient ethnicity	6	0.02
R	Primary indication for liver transplant	17	0.23
R	Secondary indication for liver transplant	13	0.00
R	Terriary indication for liver transplant	5	0.00
R	Lifestyle activity score p.t.t.	5	0.90
R	Inpatient p.t.t.	2	0.10
R	Ventilated p.t.t.	2	0.10
R	Renal support p.t.t.	3	0.21
R	Ascites p.t.t.	2	0.23
R	Diuretic therapy p.t.t.	2	0.47
R	Encephalopathy grade p.t.t.	5	1.07
R	Previous abdominal surgery p.t.t.	2	0.32
R	Oesophageal varices p.t.t.	3	0.66
R	Oesophageal varices shunt p.t.t.	3	**26.59**
R	Sepsis confirmed p.t.t.	2	0.44

Table A.2: Summary of categorical features. % null values $>$ 10% are highlighted in bold. (D = donor; O = other; R = recipient; ret. = retrieval; reg. = registration; p.t.r. = prior to retrival; p.t.t. = prior to transplant)

Source	Feature	Min	Median	Max	Mean	SD	% null
D	Donor age (years)	5.00	48.00	86.0	46.49	15.80	0.00

D	Donor height (cm)	101.00	170.00	209.0	170.50	10.49	1.20
D	Donor weight (kg)	22.00	75.00	200.0	74.76	14.58	0.19
D	Donor BMI (kg/m2)	10.50	25.20	67.6	25.73	4.66	1.24
D	Donor blood urea at ret.	0.10	5.10	49.0	6.08	3.99	6.81
D	Donor creatinine at ret.	5.00	77.00	1388.0	93.16	71.21	6.37
D	Donor potassium at ret.	1.70	4.10	9.6	4.17	0.62	8.10
D	Donor sodium at ret.	15.00	147.00	192.0	147.59	8.39	7.91
D	Donor ALT p.t.r.	0.00	29.00	4223.0	62.55	143.89	**19.69**
D	Donor AST p.t.r.	0.00	1.00	89.0	1.10	2.53	**30.69**
D	Donor bilirubin p.t.r.	0.00	9.00	158.0	11.15	8.39	5.84
D	Donor gamma GT p.t.r.	0.00	31.00	1650.0	64.34	96.88	**39.24**
D	Donor albumin p.t.r.	2.00	29.00	60.0	29.50	7.52	6.91
D	Donor alkaline phosphate p.t.r.	1.00	69.00	986.0	84.80	59.14	7.27
D	Donor INR p.t.r	0.00	1.10	13.0	1.24	0.64	**42.33**
O	Transplant year	2000.00	2008.00	2015.0	2008.12	4.71	0.00
O	Cold ischaemic time	0.00	543.00	1380.0	555.89	167.86	1.69
R	Recipient age (years)	16.00	53.00	74.0	50.12	12.66	0.00
R	Recipient weight at reg. (kg)	10.00	75.00	169.3	76.68	17.39	0.49
R	Recipient height at reg. (cm)	115.00	170.00	206.0	169.88	9.77	3.17
R	Recipient BMI	3.12	25.83	62.1	26.51	5.09	3.20
R	Recipient creatinine	11.00	87.00	858.0	99.64	54.25	0.10
R	Recipient albumin p.t.t.	1.00	30.00	58.0	30.25	7.33	0.53
R	Recipient INR p.t.t.	0.00	1.40	15.0	1.81	1.45	2.97
R	Recipient Bilirubin p.t.t.	2.00	54.00	500.0	95.45	105.48	4.61
R	Recipient Sodium p.t.t.	5.00	137.00	164.0	137.05	5.75	0.16

R	Recipient Potassium p.t.t.	2.00	4.20	9.9	4.21	0.59	0.23

Table A.3: Summary of continuous features. % null values $>$ 10% are highlighted in bold. (D = donor; O = other; R = recipient; ret. = retrieval; reg. = registration; p.t.r. = prior to retrival; p.t.t. = prior to transplant)

Biomedical Indicator	Maximum Value
ALT	5000 units/l
AST	5000 units/l
Bilirubin	500 units/l
Gamma GT	5000 units/l
Albumin	60 mmol/l
Phosphate	1.5 mmol/l
Alkaline Phosphate	1000 units/l
Blood Urea	50 mmol/l
Creatinine	1500 mmol/l
Potassium	10 mmol/l
Sodium	200 mmol/l
INR	15

Table A.4: Maximum permitted values for biomedical indicators

Feature 1	Feature 2	Source	ρ
Donor weight (kg)	Donor BMI (kg/m2)	D	0.7471
Donor blood urea at ret.	Donor creatinine at ret.	D	0.5229
Donor weight (kg)	Donor height (cm)	D	0.4856
Donor ALT p.t.r.	Donor gamma GT p.t.r.	D	0.4805
Donor alkaline phosphate p.t.r.	Donor gamma GT p.t.r.	D	0.4082
Donor ALT p.t.r.	Donor blood urea at ret.	D	0.3063
Donor height (cm)	Donor creatinine at ret.	D	0.2681
Donor age (years)	Donor BMI (kg/m2)	D	0.2581
Donor weight (kg)	Donor creatinine at ret.	D	0.2559
Recipients BMI	Recipient weight at reg. (kg)	R	0.8477
Recipient weight at reg. (kg)	Recipient height at reg. (cm)	R	0.5461

Bilirubin p.t.t.	INR p.t.t.	R	0.4734
Recipient albumin p.t.t.	INR p.t.t.	R	-0.4213
Recipient albumin p.t.t.	Bilirubin p.t.t.	R	-0.3796

Table A.5: Spearman correlations (ρ) for continuous feature where $|\rho| > 0.25$. (D = donor, R = recipient)

Feature 1	Feature 2	Source	ϕ_c
Donor cardiac arrest	Donor respiratory arrest	donor	0.7001
Donor cardiac arrest	Donor cause of death	donor	0.6038
Primary indication for liver transplant	Recipient urgency status	recipient	0.8587
Encephalopathy grade p.t.t.	Ventilated p.t.t.	recipient	0.8101
Inpatient p.t.t.	Lifestyle activity score p.t.t.	recipient	0.7926
Lifestyle activity score p.t.t.	Recipient urgency status	recipient	0.7886
Lifestyle activity score p.t.t.	Ventilated p.t.t.	recipient	0.7539
Ventilated p.t.t.	Recipient urgency status	recipient	0.7398
Encephalopathy grade p.t.t.	Recipient urgency status	recipient	0.7275
Ventilated p.t.t.	Primary indication for liver transplant	recipient	0.7079
Renal support p.t.t.	Ventilated p.t.t.	recipient	0.6919
Inpatient p.t.t.	Recipient urgency status	recipient	0.6350
Oesophageal varices shunt p.t.t.	Oesophageal varices p.t.t.	recipient	0.6226
Inpatient p.t.t.	Primary indication for liver transplant	recipient	0.6225
Renal support p.t.t.	Recipient urgency status	recipient	0.5837
Previous abdominal surgery p.t.t.	Primary indication for liver transplant	recipient	0.5787
Ascites p.t.t.	Diuretic therapy p.t.t.	recipient	0.5777
Inpatient p.t.t.	Encephalopathy grade p.t.t.	recipient	0.5503
Inpatient p.t.t.	Ventilated p.t.t.	recipient	0.5283

Table A.6: Cramer's V (ϕ_c) correlations for categorical donor and recipient features where $|\phi_c| > 0.5$

Feature	Split Criteria	p-value
Transplant year	$\geq$ 2010	1.72E-104
Cold ischaemic time	$\geq$ 525	4.68E-19
Previous abdominal surgery p.t.t.	True	5.06E-17
Recipient creatinine	$\geq$ 85.00	1.01E-15
Donor cause of death	Hypoxic brain damage	1.26E-12
Lifestyle activity score p.t.t.	Seriously restricted	1.36E-12
Sepsis confirmed p.t.t.	True	5.62E-12
Renal support p.t.t.	Filtration	6.73E-12
Donor AST p.t.r.	$\geq$ 1.00	7.74E-12
Ventilated p.t.t.	True	6.28E-10
Donor cardiac arrest	True	1.93E-09
INR p.t.t.	$\geq$ 1.40	2.07E-09
Inpatient p.t.t.	True	1.26E-08
Donor to recipient blood group match	True	1.03E-05
Primary indication for liver transplant	Alcoholic liver disease	1.47E-05
Donor creatinine at ret.	$\geq$ 76.00	1.81E-05
Recipient urgency status	True	5.78E-05
Donor respiratory arrest	True	9.76E-05
Donor potassium at ret.	$\geq$ 4.10	1.30E-04
Donor received inotropes	True	1.50E-04
Donor type	DCD	1.65E-04
Oesophageal varices shunt p.t.t.	Portosystemic shunt	4.73E-04
Recipients BMI	$\geq$ 25.89	9.17E-04
Donor history of drug abuse	True	1.22E-03
Donor CMV status	True	2.66E-03
Donor cause of death	Cerebrovascular	4.99E-03
Donor ALT p.t.r.	$\geq$ 30.00	5.25E-03
Donor history of alcohol abuse	True	7.76E-03
Recipient weight at reg. (kg)	$\geq$ 75.80	8.95E-03
Ascites p.t.t.	True	1.19E-02
Donor weight (kg)	$\geq$ 75.00	1.33E-02
Donor blood urea at ret.	$\geq$ 5.30	1.97E-02
Recipient CMV test result	True	2.09E-02
Donor age (years)	$\geq$ 49.00	3.03E-02
Organ appearance	Suboptimal	3.62E-02
Diuretic therapy p.t.t.	True	4.12E-02

Sodium p.t.t.	≥ 138.00	4.95E-02
Bilirubin p.t.t.	≥ 53.00	5.39E-02
Donor cause of death	Trauma	5.78E-02
Donor blood group	O	6.04E-02
Donor albumin p.t.r.	≥ 30.00	8.41E-02
Donor bilirubin p.t.r.	≥ 9.00	8.85E-02

Table A.7: Summary of feature splits exhibiting significant log-rank test differences. (ret. =retrieval; reg. = registration; p.t.r. = prior to retrival; p.t.t. = prior to transplant)

Appendix B

Model hyperparameters

Model	Surv. type	Surv. year	Num. trees	Max depth	Min rows	Class weights
D/R	trans.	1	600	32	4	1:1.5
D/R	trans.	5	600	32	32	1:1.5
D/R	pat.	1	600	32	4	1:1.5
D/R	pat.	5	600	32	16	1:1.5
D	trans.	1	600	16	32	1:1.5
D	trans.	5	600	4	4	1:1.5
D	pat.	1	600	4	32	1:1.5
D	pat.	5	600	16	1	1:1.5

Table B.1: Hyperparameters for random forest models. (D = donor only, D/R = donor/recipient, trans. = transplant, pat. = patient, surv. = survival)

Model	Surv. type	Surv. year	Layers	DO	Emb. DO	Upsamp.
D/R	trans.	1	[60, 30]	0.6	0.2	$\times$1.25
D/R	trans.	5	[60, 30]	0.6	0.2	$\times$1.25
D/R	pat.	1	[60, 30]	0.6	0.2	$\times$1.25
D/R	pat.	5	[60, 30]	0.6	0.2	$\times$1.25
D	trans.	1	[40]	0.6	0.2	$\times$1.25
D	trans.	5	[40]	0.6	0.2	$\times$1.25
D	pat.	1	[40]	0.6	0.2	$\times$1.25
D	pat.	5	[40]	0.6	0.2	$\times$1.25

Table B.2: Hyperparameters for ANN models. A learning rate of 0.025 and batch size of 96 was used for all models. (D = donor only, D/R = donor/recipient, trans. = transplant, pat. = patient, surv. = survival, DO = dropout, Emb. DO = embedding layer dropout, Upsamp. = upsampling of minority class)

Model	Surv. type	Surv. year	Num. trees	Max depth	Min rows
D/R	trans.	1	400	12	12
D/R	trans.	5	400	12	12
D/R	pat.	1	400	12	12
D/R	pat.	5	400	12	12
D	trans.	1	400	12	12
D	trans.	5	400	12	12
D	pat.	1	400	12	12
D	pat.	5	400	12	12

Table B.3: Hyperparameters for RSF models. (D = donor only, D/R = donor/recipient, trans. = transplant, pat. = patient, surv. = survival)

Model	Surv. type	Surv. year	Layers	DO	L2 reg.
D/R	trans.	1	[60]	0.3	0.001
D/R	trans.	5	[60]	0.3	0.001
D/R	pat.	1	[60]	0.3	0.001
D/R	pat.	5	[60]	0.3	0.001
D	trans.	1	[60]	0.3	0.001
D	trans.	5	[60]	0.3	0.001
D	pat.	1	[60]	0.3	0.001
D	pat.	5	[60]	0.3	0.001

Table B.4: Hyperparameters for N-MTLR models. A learning rate of 0.001 was used for all models. (D = donor only, D/R = donor/recipient, trans. = transplant, pat. = patient, surv. = survival, DO = dropout, L2 reg. = L2 regularization)

Appendix C

Supplementary feature importance tables

Feature	Feature Category	Coefficient
Organ transplanted	reduced liver	0.7593
Primary indication for LT	Other cancers	0.6442
Primary indication for LT	Retransplantation	0.5382
Primary indication for LT	Non-alcoholic fatty liver disease	0.5382
Primary indication for LT	Hepatocellular carcinoma - cirrhotic	0.4809
Donor cause of death	Known or suspected suicide	0.4744
Secondary indication for LT	Other cancers	0.4669
Primary indication for LT	Primary sclerosing cholangitis	0.4296
Donor type	DCD	0.3688
Primary indication for LT	Hepatitis C cirrhosis	0.3583
Secondary indication for LT	Retransplantation	0.3578
Lifestyle activity score p.t.t.	Completely reliant on nursing/medical care	0.3344
Encephalopathy grade p.t.t.	Stuporous but speaking and obeying simple commands	0.3303
Primary indication for LT	Acute vascular occlusion	0.3273
Renal support p.t.t.	Filtration	0.3184
Lifestyle activity score p.t.t.	Only capable of limited self care.	0.3174
Primary indication for LT	Autoimmune chronic active liver disease	0.2963
Primary indication for LT	Alcoholic liver disease	0.2677

Donor cause of death	Other	0.2560
Sepsis confirmed p.t.t.	Yes	0.2182
Lifestyle activity score p.t.t.	Can move freely. Capable of self care.	0.2130
Organ transplanted	split liver	0.2119
Primary indication for LT	Cryptogenic cirrhosis	0.2064
Secondary indication for LT	Cirrhotic hepatocellular carcinoma	0.1778
Donor diabetes	Yes	0.1769
Previous abdominal surgery p.t.t.	Yes	0.1634
Donor cardio disease	Yes	0.1526
Organ appearance	Suboptimal	0.1251
Donor CMV status	Positive	0.0742
Recipient age (years)	-	0.0065
Donor age (years)	-	0.0061
Cold ischaemic time	-	0.0002
Transplant year	-	-0.0401
Bilirubin p.t.t.	-	-0.0694
Oesophageal varices p.t.t.	Previous variceal bleed	-0.0927
Donor INR p.t.r	-	-0.1165
INR p.t.t.	-	-0.1527

Table C.1: Significant features for transplant survival Cox PH model. (LT = liver transplant, ret. = retrieval, reg. = registration, p.t.r. = prior to retrival, p.t.t. = prior to transplant)

Feature	Feature Category	Coefficient
Recipient ethnicity	Mixed-race	1.4199
Primary indication for LT	Other cancers	0.8390
Primary indication for LT	Cirrhotic hepatocellular carcinoma	0.6956
Primary indication for LT	Non-alcoholic fatty liver disease	0.6107
Organ transplanted	reduced liver	0.5576
Primary indication for LT	Paracetamol hepatotoxicity	0.4757
Primary indication for LT	Hepatitis C cirrhosis	0.4367
Encephalopathy grade p.t.t.	Stuporous but speaking and obeying simple commands	0.4252
Primary indication for LT	Alcoholic liver disease	0.3691
Renal support p.t.t.	Filtration	0.3312

Primary indication for LT	Other metabolic liver disease	0.3265
Lifestyle activity score p.t.t.	Only capable of limited self care.	0.3070
Primary indication for LT	Cryptogenic cirrhosis	0.3041
Primary indication for LT	Primary sclerosing cholangitis	0.2979
Primary indication for LT	Other liver diseases	0.2896
Recipient ethnicity	Black or Black/British	0.2750
Primary indication for LT	Autoimmune chronic active liver disease	0.2616
Secondary indication for LT	Cirrhotic hepatocellular carcinoma	0.2516
Sepsis confirmed p.t.t.	Yes	0.2346
Recipient creatinine	-	0.1360
Previous abdominal surgery p.t.t.	Yes	0.1197
Recipient age (years)	-	0.0183
Donor age (years)	-	0.0053
Transplant year	-	-0.0495
Bilirubin p.t.t.	-	-0.0744
INR p.t.t.	-	-0.1759

Table C.2: Significant features for patient survival Cox PH model. (LT = liver transplant, ret. = retrieval, reg. = registration, p.t.r. = prior to retrival, p.t.t. = prior to transplant)

Feature	Mean % Importance
Primary indication for liver transplant	14.36%
Transplant year	7.12%
Donor cause of death	6.18%
Secondary indication for liver transplant	4.50%
Recipient age (years)	2.98%
Donor age (years)	2.67%
Recipient creatinine	2.50%
Bilirubin p.t.t.	2.30%
Lifestyle activity score p.t.t.	2.29%
Cold ischaemic time	1.95%
Recipient weight at registration (kg)	1.94%
Potassium p.t.t.	1.78%
Donor albumin p.t.r.	1.71%
Donor gamma GT p.t.r.	1.68%
Recipients BMI	1.68%

Recipient albumin p.t.t.	1.66%
Encephalopathy grade p.t.t.	1.64%
Donor potassium at ret.	1.62%
Recipient blood group	1.54%
Donor blood urea at ret.	1.51%
Recipient height at registration (cm)	1.50%
Renal support p.t.t.	1.48%
Donor INR p.t.r	1.47%
Donor height (cm)	1.40%
Donor weight (kg)	1.40%
Donors family history of diabetes	1.31%
Donor blood group	1.27%
Oesophageal varices p.t.t.	1.26%
Donor BMI (kg/m2)	1.25%
Donor sodium at ret.	1.23%
INR p.t.t.	1.15%
Donor alkaline phosphate p.t.r.	1.12%
Sodium p.t.t.	1.10%
Oesophageal varices shunt p.t.t.	1.05%
Organ appearance	1.04%
Donor AST p.t.r.	1.00%

Table C.3: Random forest features with mean % importance across transplant and patient survival models of 1% or more. (ret. = retrieval, reg. = registration, p.t.r. = prior to retrival, p.t.t. = prior to transplant)

Appendix D

Liver transplant score formulae

D.1 DLI score

DLI1 formula (survival at 1 year)

```
2.3159
+ 0.9106 * (DCD donor)
+ 0.7140 * (liver meets split criteria)
- 0.01434 * (height)
+ 0.3058 * (history of cardiac disease)
+ 0.2545 * (steatosis present)
+ 0.01222 * (donor bilirubin)
+ 0.1736 * (history of smoking)
+ 0.6453 * (black ethnicity)
```

DLI formula (overall survival)

```
1.6775
+ 0.009179 * (donor age)
— 0.1948 * (female)
+ 0.6363 * (DCD donor)
+ 0.4697 * (liver meets split criteria)
— 0.01283 * (height)
+ 0.1570 * (history of smoking)
+ 0.009019 * (bilirubin)
```

D.2 TBS score

TBS non-cancer post-transplant formula

```
+ 0.00633 * (donor age - 46.49820)
+ 0.23837 * (donor cause of death of trauma)
- 0.33765 * (donor cause of death of other trauma)
- 0.07308 * (other donor cause of death)
+ 0.01234 * (donor BMI - 25.87324 )
- 0.23126 * (donor history of diabetes)
+ 0.10454 * (unknown history of diabetes)
+ 3.68574 * (DCD donor)
+ 0.84086 * (recipient HCV and donor history of diabetes interaction)
+ 0.90577 * (recipient HCV and unknown donor history of diabetes interaction)
+ 0.01472 * (recipient HCV and donor age interaction)
+ 0.00730 * (DCD and recipient age interaction)
- 0.79467 * (DCD and ln(creatinine) interaction)
- 0.26206 * (DCD and HCV interaction)
- 0.39010 * (DCD and HBV interaction)
+ 0.08220 * (DCD and PSC interaction)
- 0.24764 * (DCD and PBC interaction)
- 0.73028 * (DCD and AID interaction)
+ 0.41517 * (DCD and metabolic disease interaction)
- 0.68482 * (DCD and other disease interaction)
- 0.41539 * (DCD and previous transplant interaction)
+ 0.16495 * (blood group compatibility)
+ 0.36079 * (liver meets split criteria)
- 0.09830 * (recipient age - 50.49098)
- 0.00502 * (female recipient)
- 0.29677 * (if Hepatitis C)
- 0.24214 * (Hepatitis C (HCV) disease)
+ 2.57881 * (Hepatitis B (HBV) disease)
- 0.10269 * (Primary sclerosing cholangitis (PSC) disease)
- 1.45184 * (Primary biliary cirrhosis (PBC) disease)
- 0.08155 * (Auto-immune and cryptogenic disease (AID))
+ 0.81764 * (Metabolic liver disease)
+ 0.14073 * (Other liver disease)
+ 0.58202 * (one or more previous transplants)
```

```
- 0.97180 * (ln(creatinine) - 4.47103)
+ 0.03579 * (ln(bilirubin) - 4.21900)
- 0.43956 * (ln(INR) - 0.39632)
- 0.00818 * (sodium - 136.38798)
+ 0.01296 * (potassium - 4.19735)
- 0.01464 * (albumin - 30.59920)
+ 0.59629 * (renal replacement therapy)
+ 0.27483 * (in-patient)
+ 0.14433 * (previous abdominal surgery)
+ 0.00928 * (encephalopathy)
+ 0.02489 * (ascites)
- 0.00315 * (ln(waiting time) - 4.34687 )
+ 0.14190 * (diabetes)
+ 0.02117 * (recipient age and ln(creatinine) interaction - 226.25081  )
+ 0.00099 * (age and HCV interaction)
- 0.05865 * (age and HBV interaction)
+ 0.00743 * (age and PSC interaction)
+ 0.02608 * (age and PBC interaction)
+ 0.00128 * (age and AID interaction)
- 0.01243 * (age and metabolic disease interaction)
+ 0.00117 * (age and other disease interaction)
- 0.00042 * (age and previous transplant interaction)
```

TBS cancer post-transplant formula

```
+ 0.02000 * (donor age - 49.19070 )
+ 0.83592 * (donor cause of death of trauma)
- 0.86246 * (donor cause of death of other trauma)
- 0.00300 * (other donor cause of death)
- 0.01591 * (donor BMI - 26.42063 )
+ 0.16656 * (donor history of diabetes)
- 0.06450 * (unknown history of diabetes)
+ 1.74717 * (DCD donor)
+ 1.50486 * (recipient HCV and donor history of diabetes interaction)
+ 0.94683 * (recipient HCV and unknown donor history of diabetes interaction)
+ 0.02709 * (recipient HCV and donor age interaction)
+ 0.03521 * (DCD and recipient age interaction)
- 0.74070 * (DCD and ln(creatinine) interaction)
```

```
+ 1.26327 * (blood group compatibility)
+ 0.42794 * (liver meets split criteria)
+ 0.35880 * (age - 57.10233)
- 0.34232 * (female)
- 1.06678 * (if Hepatitis C)
+ 4.84495 * (ln(creatinine) - 4.36459)
+ 0.02605 * (ln(bilirubin) - 3.08049)
- 0.08014 * (ln(INR) - 0.26144)
+ 0.04974 * (sodium - 139.06512)
+ 0.30323 * (potassium - 4.19023)
+ 0.02115 * (albumin - 34.8)
- 1.21662 * (renal replacement therapy)
- 0.22755 * (in-patient)
+ 0.05661 * (previous abdominal surgery)
+ 0.45698 * (encephalopathy)
+ 0.52255 * (ascites)
+ 0.04573 * (ln(waiting time) - 4.22328)
+ 0.24128 * (diabetes)
- 0.08481 * (recipient age and ln(creatinine) interaction - 249.39181)
+ 0.05055 * (ln(maximum AFP) - 3.01254)
+ 0.24722 * (maximum tumour size - 2.63209)
+ 0.01906 * (if two tumours)
- 0.35068 * (if three or more tumours)
```

Bibliography

[1] James Neuberger. Liver transplantation in the united kingdom. *Liver Transplantation*, 22(8):1129–1135, 2016.

[2] Eric S Orman, Maria E Mayorga, Stephanie B Wheeler, Rachel M Townsley, Hector H Toro-Diaz, Paul H Hayashi, and A Sidney Barritt. Declining liver graft quality threatens the future of liver transplantation in the united states. *Liver Transplantation*, 21(8):1040–1050, 2015.

[3] Irine Vodkin and Alexander Kuo. Extended criteria donors in liver transplantation. *Clinics in liver disease*, 21(2):289–301, 2017.

[4] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[5] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[6] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50:163–170, 1966.

[7] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[8] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[9] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[10] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[11] Hemant Ishwaran and Min Lu. Random survival forests. *Wiley StatsRef: Statistics Reference Online*, pages 1–13, 2008.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[14] Knut Liestøl, Per Kragh Andersen, and Ulrich Andersen. Survival analysis and neural nets. *Statistics in medicine*, 13(12):1189–1200, 1994.

[15] David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

[16] Anny Xiang, Pablo Lapuerta, Alex Ryutov, Jonathan Buckley, and Stanley Azen. Comparison of the performance of neural network methods and cox regression for censored survival data. *Computational statistics & data analysis*, 34(2):243–257, 2000.

[17] Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

[18] W Nick Street. A neural network model for prognostic prediction. In *ICML*, pages 540–546, 1998.

[19] Taysseer Sharaf and Chris P Tsokos. Two artificial neural networks for modeling discrete survival time of censored data. *Advances in Artificial Intelligence*, 2015:1, 2015.

[20] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in neural information processing systems*, pages 971–980, 2017.

[24] Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

[25] Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems*, pages 1845–1853, 2011.

[26] Patrick S. Kamath, Russell H. Wiesner, Michael Malinchoc, Walter Kremers, Terry M. Therneau, Catherine L. Kosberg, Gennaro D'amico, E. Rolland Dickson, and W. Ray Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464–470, 2001.

[27] J Neuberger, A Gimson, M Davies, M Akyol, J O'Grady, A Burroughs, M Hudson, Advisory Group Liver, et al. Selection of patients for liver transplantation and allocation of donated livers in the uk. *Gut*, 57(2):252, 2008.

[28] S. Feng, S.M. Greenstein, J.D. Punch, M.A. DebRoy, R.M. Merion, J.L. Bragg-Gresham, N.P. Goodrich, and D.M. Dykstra. Characteristics Associated with Liver Graft Failure: The Concept of a Donor Risk Index. *American Journal of Transplantation*, 6(4):783–790, 2006.

[29] David Collett, Peter J. Friend, and Christopher J. E. Watson. Factors Associated With Short- and Long-term Liver Graft Survival in the United Kingdom. *Transplantation*, 101(4):786–792, 2016.

[30] O. Fix, R. Bakthavatsalam, J. B. Halldorson, J. D. Perkins, and J. D. Reyes. D-MELD, a Simple Predictor of Post Liver Transplant Mortality for Optimization of Donor/Recipient Matching. *American Journal of Transplantation*, 9(2):318–326, 2008.

[31] M. A. Hardy, L. E. Ratner, A. Rana, K. J. Halazun, B. Samstein, J. C. Emond, D. C. Woodland, J. V. Guarrera, and R. S. Brown Jr. Survival Outcomes Following Liver Transplantation (SOFT) Score: A Novel Method to Predict Patient Survival Following Liver Transplantation. *American Journal of Transplantation*, 8(12):2537–2546, 2008.

[32] Geoffrey H Haydon, Yrjo Hiltunen, Michael R Lucey, David Collett, Bridget Gunson, Nick Murphy, Peter G Nightingale, and James Neuberger. Self-organizing maps can determine outcome and match recipients and donors at orthotopic liver transplantation. *Transplantation*, 79(2):213–218, 2005.

[33] N Hoot and D Aronsky. Using Bayesian networks to predict survival of liver transplant patients. Technical report, 2005.

[34] Alessandro Cucchetti, Marco Vivarelli, Nigel D Heaton, Simon Phillips, Fabio Piscaglia, Luigi Bolondi, Giuliano La Barba, Matthew R Foxton, Mohamed Rela, John O'Grady,

et al. Artificial neural network is superior to meld in predicting mortality of patients with end-stage liver disease. *Gut*, 56(2):253–258, 2007.

[35] Ming Zhang, Fei Yin, Bo Chen, You Ping Li, Lu Nan Yan, Tian Fu Wen, and Bo Li. Pretransplant prediction of posttransplant survival for liver recipients with benign end-stage liver diseases: a nonlinear model. *PLoS One*, 7(3):e31256, 2012.

[36] Ming Zhang, Fei Yin, Bo Chen, Bo Li, You Ping Li, Lu Nan Yan, and Tian Fu Wen. Mortality risk after liver transplantation in hepatocellular carcinoma recipients: a nonlinear predictive model. *Surgery*, 151(6):889–897, 2012.

[37] Manuel Cruz-Ramirez, Cesar Hervas-Martinez, Juan Carlos Fernandez, Javier Briceno, and Manuel De La Mata. Predicting patient survival after liver transplantation using evolutionary multi-objective artificial neural networks. *Artificial intelligence in medicine*, 58(1):37–49, 2013.

[38] Javier Briceño, Manuel Cruz-Ramírez, Martín Prieto, Miguel Navasa, Jorge Ortiz De Urbina, Rafael Orti, Miguel Ángel Gómez-Bravo, Alejandra Otero, Evaristo Varo, Santiago Tomé, Gerardo Clemente, Rafael Bañares, Rafael Bárcena, Valentín Cuervas-Mons, Guillermo Solórzano, Carmen Vinaixa, Ángel Rubín, Jordi Colmenero, Andrés Valdivieso, Rubén Ciria, César Hervás-Martínez, and Manuel De La Mata. Use of artificial intelligence as an innovative donor-recipient matching model for liver transplantation: Results from a multicenter Spanish study. *Journal of Hepatology*, 61(5):1020–1028, 2014.

[39] Bahareh Khosravi, Saeedeh Pourahmad, Amin Bahreini, Saman Nikeghbalian, and Goli Mehrdad. Five Years Survival of Patients After Liver Transplantation and Its Effective Factors by Neural Network and Cox Proportional Hazard Regression Models. 15(9):25164, 2015.

[40] Manuel Dorado-Moreno, María Pérez-Ortiz, Pedro A. Gutiérrez, Rubén Ciria, Javier Briceño, and César Hervás-Martínez. Dynamically weighted evolutionary ordinal neural network for solving an imbalanced liver transplantation problem. *Artificial Intelligence in Medicine*, 77:1–11, mar 2017.

[41] L Lau, Y Kankanige, B Rubinstein, R Jones, C Christophi, V Muralidharan, and J Bailey. Machine-learning algorithms predict graft failure after liver transplantation. *Transplantation*, 101(4):e125–e132, 2017.

[42] Gerald J Glasser and Robert F Winter. Critical values of the coefficient of rank correlation for testing the hypothesis of independence. *Biometrika*, 48(3/4):444–448, 1961.

[43] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

[44] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[45] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, 16(3):74, 2016.

[46] S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[47] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[48] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[49] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[50] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[51] Bradley Efron and C Stein. The annals of statistics. *Bootstrap method: another look at the jackknife*, 7:1–26, 1979.