Analysis of sarcoma and non-sarcoma clinical data with statistical methods and machine learning techniques

Kantidakis, G.

6

Survival prediction models since liver transplantation - comparisons between Cox models and machine learning techniques

This chapter is based on: G. Kantidakis, H. Putter, C. Lancia, J. Boer, A. E. Braat, and M. Fiocco. Survival prediction models since liver transplantation - comparisons between Cox models and machine learning techniques. *BMC Medical Research Methodology*, 20(1):1-14, 2020.

Abstract

Background: Predicting survival of recipients after liver transplantation is regarded as one of the most important challenges in contemporary medicine. Hence, improving on current prediction models is of great interest. Nowadays, there is a strong discussion in the medical field about machine learning (ML) and whether it has greater potential than traditional regression models when dealing with complex data. Criticism to ML is related to unsuitable performance measures and lack of interpretability which is important for clinicians.

Methods: In this paper, ML techniques such as random forests and neural networks are applied to large data of 62294 patients from the United States with 97 predictors selected on clinical/statistical grounds, over more than 600, to predict survival from transplantation. Of particular interest is also the identification of potential risk factors. A comparison is performed between 3 different Cox models (with all variables, backward selection and LASSO) and 3 machine learning techniques: a random survival forest and 2 partial logistic artificial neural networks (PLANNs). For PLANNs, novel extensions to their original specification are tested. Emphasis is given on the advantages and pitfalls of each method and on the interpretability of the ML techniques.

Results: Well-established predictive measures are employed from the survival field (C-index, Brier score and Integrated Brier Score) and the strongest prognostic factors are identified for each model. Clinical endpoint is overall graft-survival defined as the time between transplantation and the date of graft-failure or death. The random survival forest shows slightly better predictive performance than Cox models based on the C-index. Neural networks show better performance than both Cox models and random survival forest based on the Integrated Brier Score at 10 years.

Conclusions: In this work, it is shown that machine learning techniques can be a useful tool for both prediction and interpretation in the survival context. From the ML techniques examined here, PLANN with 1 hidden layer predicts survival probabilities the most accurately, being as calibrated as the Cox model with all variables.

6.1 Introduction

Liver transplantation (LT) is the second most common type of transplant surgery in the United States after kidney [1]. Over the last decades, the success of liver transplants has improved survival outcome for a large number of patients suffering from chronic liver disease everywhere on earth [2]. Availability of donor organs is a major limitation especially when compared with the growing demand of liver candidates due to the enlargement of age limits. Therefore, improvement on current prediction models for survival since LT is important.

There is an open discussion about the value of machine learning (ML) versus statistical models (SM) within clinical and healthcare practice [3–7]. For survival data, the most commonly applied statistical model is the Cox proportional hazards regression model [8]. This model allows a straightforward interpretation, but is at the same time restricted to the proportional hazards assumption. On the other hand, ML techniques are assumption-free and data adaptive which means that they can be effectively employed for modelling complex data. In this article, the results between SM and ML techniques are assessed based on a 3-stage comparison: predictive performance for large sample size/large number of covariates, calibration (absolute accuracy) which is often neglected, and interpretability in terms of the most prognostic factors identified. Advantages and disadvantages for each method are detailed.

ML techniques need a precise set of operating conditions to perform well. It is important that a) the data have been adequately processed so that the inputs allow for good learning, b) modern method is applied using state-of-the-art programming software and c) proper tuning of the parameters is performed to avoid sub-optimal or default choices for parameters which downgrade the algorithm's performance. Danger of overfitting is associated with ML approaches (as they employ complex algorithms). A note of caution is required during model training to prevent from overfitting, e.g. the selection of suitable hyper-parameters. Needless to say, overfitting might also occur with a traditional model if it is too complex (estimation of too many parameters) thus limiting generalizability outside training instances.

Neural networks have been commonly applied in healthcare. Consequently, different approaches for time-to-event endpoints are present in the literature. Biganzoli *et al.* proposed a partial logistic regression approach of feed forward neural networks (PLANN) for flexible modelling of survival data [9]. By using the time interval as an input in a longitudinally transformed feed forward network with logistic activation and entropy error function, they estimated smoothed discrete hazards at each time interval in the output layer. This is a well known approach for modelling survival neural networks [10]. In 2000, Xiang *et al.* [11] compared the performance of 3 existing neural network methods for right censored data (the Faraggi-Simon [12], the Liestol-Andersen-Andersen [13] and a modification of the Buckley-James method [14]) with Cox models in a Monte Carlo simulation study. None of the networks outperformed the Cox models and they only performed as good as Cox for some scenarios. Lisboa *et al.* extended the PLANN approach introducing a Bayesian framework which can perform Automatic Relevance Determination for survival data (PLANN-ARD) [15]. Several applications of the PLANN and the PLANN-ARD methods can be found in the literature [16–19]. They show potential for neural networks in systems with non-linearity and complex interactions between factors. Here extensions of the PLANN approach for big LT data are examined.

The clinical endpoint of interest for this study is overall graft-survival defined as the time between LT and graft-failure or death. Predicting survival after LT is hard as it depends on many factors and is associated with donor, transplant and recipient characteristics whose importance changes over time and per outcome measure [20]. Models that combine donor and recipient characteristics have usually better performance for predicting overall graft-survival and particularly those that include sufficient donor risk factors have better performance for long-term graft survival [21]. The aims of this manuscript can be summarised as: i) potential role of ML as a competitor of traditional methods when complexity of the data is high (large sample size, high dimensional setting), ii) identification of potential risk factors using 2 ML methods (random survival forest, survival neural networks) complementary

to the Cox model, iii) use of variable selection methods to compare predictive ability with the models including the non-reduced set of variables, iv) evaluation of predictions and goodness of fit, and v) clinical relevance of the findings (potential for medical applications).

The paper is organized as follows. Section "Methods" presents details about data collection and the imputation technique, SMs and ML. Further sections discuss model training, predictive performance assessment on test data, and details about interpretability of the models. Comparisons between models based on global performance measures, prediction error curves, variable importance and calibration plots are discussed in the section "Results". The article is concluded by the "Discussion" about findings, limitations of this work and future perspectives. All analyses were performed in R programming language version 3.5.3 [22]. Preliminary results were presented at 40th Annual Conference of the International Society for Clinical Biostatistics [23].

6.2 Methods

An analysis is presented on survival data after LT based on 62294 patients from the United States. Information was collected from the United Network of Organ Sharing (UNOS) [1]. After extensive pre-processing from a set of more than 600 covariates, 97 variables were included in the final dataset based on clinical and statistical considerations (see Additional file 1); 52 donor and 45 liver recipient characteristics (missing values were imputed). As the UNOS data is large in both number of observations and covariates, it is of interest to see how ML algorithms - which are able to capture naturally multi-way interactions between variables and can deal with big datasets - will perform compared to Cox models. The clinical endpoint is overall graft-survival (OGS) the time between LT and graft-failure or death). The choice for this endpoint was made for two reasons 1) it is of primary interest for clinicians and 2) it is the most appropriate outcome measure to evaluate the efficacy of LT, because it incorporates both patient mortality and survival of the graft [21].

This section is divided into different subsections including the necessary components of analyses for OGS (provided in "Results" section). We discuss in detail both Cox models and ML techniques (Random Survival Forest, Survival Neural Networks). Elements of how the models were trained and how the predictive performance was assessed on the test data are presented. More technical details are provided in the supplementary material. We conclude this extensive section with a focus on methods to extract interpretation for the ML approaches.

6.2.1 Data collection and imputation technique

UNOS manages the Organ Procurement and Transplantation Network (OPTN) and together they collect, organise and maintain data of statistical information regarding organ transplants in the Scientific Registry of Transplant Recipients (SRTR) database [2]. SRTR gathers data from local Organ Procurement Organisations (OPO) and from OPTN (primary source). It includes data from transplantations performed in the United States from 1988 onwards. This information is used to set priorities and seek improvements in the organ donation process.

The data provided by UNOS included 62294 patients who underwent LT surgery from 2005 to 2015 (project under DUA number 9477). Standard analysis files contained 657 variables for both donors and patients (candidates and recipients). Among these, 97 candidate risk factors - 52 donor and 45 patient characteristics - were pre-selected before carrying out analysis. This resulted in a final dataset with 76 categorical and 21 continuous variables amounting to 2.2% missing data overall. The percentage of missing values for each covariate varied from 0 to 26.61% (no missing values for 26 covariates, up to 1% missingness for 51 covariates, 1 to 10% for 11 variables, 10 to 25% for 7 variables and 25 to 26.61% for only 2 variables). Analysis on the complete case would reduce

[1] UNOS is a non-profit and scientific organisation in the United States which arranges organ donation and transplantation. For more information visit its website https://unos.org.

[2] Dictionary for variables details is provided at: https://www.srtr.org/requesting-srtr-data/saf-data-dictionary/.

the available sample size from 62294 to 33394 patients leading to a huge waste of data. Furthermore, this could lead to invalid results (underestimation or overestimation of survival) if the excluded group of patients represents a subgroup from the entire sample [24]. To reconstruct the missing values the `missForest` algorithm [25] was applied for both continuous and categorical variables. This is a non-parametric imputation method that does not make explicit assumptions about the functional form of the data and builds a random forest model for each variable (500 trees were used). It specifies the model to predict missing values by using information based on the observed values. It is the most exhaustive and accurate of all random forests algorithms used for missing data imputation, because all possible variable combinations are checked as responses.

6.2.2 Cox proportional hazard regression models

In survival analysis, the focus is on the time till the occurrence of the event of interest (here graft-failure or death). The Cox proportional hazards model is usually employed to estimate the effect of risk factors on the outcome of interest [8].

Data with sample size n consist of the independent observations from the triple (T, D, X) i.e. $(t_1, d_1, x_1), \cdots, (t_n, d_n, x_n)$. For the i^{th} individual, t_i is the survival time, d_i the indicator ($d_i = 1$ if the event occurred and $d_i = 0$ if the observation is right censored) and x_i is the vector of predictors $(x_1, \cdots, x_p)$. The hazard function of the Cox model with time-fixed covariates is as follows:

$$h(t|X) = h_0(t) \exp(X^T \boldsymbol{\beta}), \tag{6.1}$$

where $h(t|X)$ is the hazard at time t given predictor values X, $h_0(t)$ is an arbitrary baseline hazard and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ is a parameter vector.

The corresponding partial likelihood can be written as:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p} \beta_k X_{ik}\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)}, \tag{6.2}$$

where D is the set of failures, and $R(t_i)$ is the risk set at time t_i of all individuals who are still in the study at the time just before time t_i. This function is then maximised over $\boldsymbol{\beta}$ to estimate the model parameters.

Two other Cox models were employed 1) a Cox model with a backward elimination and 2) a penalised Cox regression with the Least Angle and Selection Operator (LASSO). Both models have been widely used for variable selection. We aim to compare these more parsimonious models versus a Cox model with all variables in terms of predictive performance. For the first, a numerically stable version of the backward elimination on factors was applied using a method based on Lawless and Singhal (1978) [26]. This method estimates the full model and computes approximate Wald statistics by computing conditional maximum likelihood estimates - assuming multivariate normality of estimates. Factors that require multiple degrees of freedom are dropped or retained as a group. The latter approach uses a combination of selection and regularisation [27]. Denote the log-partial likelihood by $\ell(\boldsymbol{\beta}) = logL(\boldsymbol{\beta})$. The vector $\boldsymbol{\beta}$ is estimated via the criterion:

$$\hat{\boldsymbol{\beta}} = \text{argmin}[\ell(\boldsymbol{\beta})], \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq s \tag{6.3}$$

with s a user specified positive parameter.

Equation (6.3) can also be rewritten as

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\text{argmin}} \left(\ell(\beta) + \lambda_{LASSO} \sum_{j=1}^{p} |\beta_j| \right). \tag{6.4}$$

The quantity $\sum_{j=1}^{p} |\beta_j|$ is also known as the L_1-norm and performs regularisation to the log-partial likelihood. The term λ_{LASSO} is a non-negative constant that assigns the amount of penalisation. Larger values for the parameter mean larger penalty to the β_j coefficients and enlarged shrinkage towards zero.

The tuning parameter s in equation (6.3) or equivalently parameter λ_{Lasso} in equation (6.4) is the controlling mechanism for the variance of the model. Higher values reduce further the variance but introduce at the same time more bias (variance-bias trade off). To find a suitable value for this parameter 5-fold cross-validation was performed to minimise the prediction error; here in terms of the cross-validated log-partial likelihood (CVPL) [28]

$$CVPL(s) = \sum_{i=1}^{n} (\ell(\hat{\beta}_{(-i)}(s)) - \ell_{(-i)}(\hat{\beta}_{(-i)}(s))), \tag{6.5}$$

where $\ell_{(-i)}(\beta)$ is the partial log-likelihood of equation (6.2) when individual i is excluded. Therefore, the term $\ell(\hat{\beta}_{(-i)}) - \ell_{(-i)}(\hat{\beta}_{(-i)})$ represents the contribution of observation i. The value that maximizes $\ell_{(-i)}(\beta_{(-i)})$ is denoted by $\hat{\beta}_{(-i)}$.

6.2.3 Random forests for survival analysis

Random Survival Forests (RSFs) are an ensemble tree method for survival analysis of right censored data [29] adapted from random forests [30]. The main idea of random forests is to get a series of decision trees - which can capture complex interactions but are notorious for their high variance - and obtain a collection averaging their characteristics. In this way weak learners (the individual trees) are turned into strong learners (the ensemble) [31].

For RSFs, randomness is introduced in two ways: bootstrapping a number of patients at each tree $\mathcal{B}$ times and selecting a subset of variables for growing each node. During growing each survival tree, a recursive application of binary splitting is performed per region (called node) on a specific predictor in such a way that survival difference between daughter nodes is maximised and difference within them is minimised. Splitting is terminated when a certain criterion is reached (these nodes are called terminal). The most commonly used splitting criteria are the log-rank test by Segal [32] and the log-rank score test by Hothorn and Lausen [33]. Each terminal node should have at least a pre-specified number of unique events. Combining information from the $\mathcal{B}$ trees, survival probabilities and ensemble cumulative hazard estimate can be calculated using the Kaplan-Meier and Nelson-Aalen methodology, respectively.

The fundamental principle behind each survival tree is the conservation of events. It is used to define ensemble mortality, a new type of predicted outcome for survival data derived from the ensemble cumulative hazard function (comparable to the prognostic index based on the Cox model). This principle asserts that the sum of estimated cumulative hazard estimate over time is equal to the total number of deaths, therefore the total number of deaths is conserved within each terminal node $\mathcal{H}$ [29]. RSFs can handle both data with large sample size and vast number of predictors. Moreover, they can reach remarkable stability combining the results of many trees. However, combining an ensemble of trees downgrades significantly the intuitive interpretation of a single tree.

6.2.4 Survival neural networks

Artificial neural networks (NNs) are a machine learning method able to model non-linear relationships between prognostic factors with great flexibility. These systems are inspired from biological neural networks that aimed at imitating the human brain activity [34]. A NN has a layered structure and is based on a collection of connected units called nodes or neurons which comprise a layer. The input layer picks up the signals and passes them through transformation functions to the next layer which is called ”hidden”. A network may have more than one hidden layer that connects with the previous and transmit signals towards the output layer. Connections between artificial neurons are called edges. Artificial neurons and edges have a weight (connection strength) which adjusts as learning proceeds. It increases or decreases the strength of the signal of each connection according to its sign. For the

purpose of training, a target is defined, which is the observed outcome. The simplest form of a NN is the single layer feed-forward perceptron with the input layer, one hidden layer and the output layer [35].

The application of NNs has been extended to survival analysis over the years [13]. Different approaches have been considered; some model the survival probability $\mathcal{S}(t)$ directly or the unconditional probability of death $\mathcal{F}(t)$ whereas other approaches estimate the conditional hazard $h(t)$ [10]. They can be distinguished according to the method used to deal with the censoring mechanism. Some networks have k output nodes [36] - where k denotes k separate time intervals - while others have a single output node.

In this research, the method of Biganzoli was applied, which specifies a partial logistic feed-forward artificial neural network (PLANN) with a single output node [9]. This method uses as inputs the prognostic factors and the survival times to increase the predictive ability of the model. Data have to be transformed into a longitudinal format with the survival times being divided into a set of k non-overlapping intervals (months or years) $I_k = (\tau_{k-1}, \tau_k]$, with $0 = \tau_o < \tau_1 < \cdots < \tau_k$ a set of pre-defined time points. In this way, the time component of survival data is taken into consideration. On the training data, each individual is repeated for the number of intervals he/she was observed in the study and on the test data for all time intervals. PLANN provides the discrete conditional probability of dying $\mathcal{P}(T \in I_k \mid T > \tau_{k-1})$ using as transformation function of both input and output layers the logistic (sigmoid) function:

$$f(\eta) = \frac{1}{1 + e^{-\eta}}, \tag{6.6}$$

where $\eta = \sum_{i=1}^{p} w_i X_i$ is the summed linear combination of the weights w_i of input-hidden layer and the input variables X_i $(i = 1, 2, \cdots, p)$.

The contribution to the log-likelihood for each individual is calculated all over the intervals one is at risk. The output node is one large target vector with 0 if the event did not occur and 1 if the event occurred in a specific time interval. Therefore, such a network first estimates the hazard for each interval $h_k = P(\tau_{k-1} < T \leq \tau_k | T > \tau_{k-1})$ and then $S(t) = \prod_{k:t_k \leq t}(1 - h_k)$.

In this work, novel extensions in the specification of the PLANN are tested. Two new transformation functions were investigated for the input-hidden layer the rectified linear unit (ReLU)

$$f(\eta) = \eta^+ = \max(0, \eta), \tag{6.7}$$

which is the most used activation function for NNs and the hyperbolic tangent (tanh)

$$f(\eta) = \frac{1 - e^{-2\eta}}{1 + e^{-2\eta}}. \tag{6.8}$$

These functions can be seen as different modulators of the degree of non-linearity implied by the input and the hidden layer.

The PLANN was expanded in 2 hidden layers with same node size and identical activation functions for input-hidden 1 and hidden 1 - hidden 2 layers. The k non-overlapping intervals of the survival times were treated as k separate variables. In this way, the contribution of each interval to the predictions of the model using the relative importance method by Garson [37] and its extension for 2 hidden layers can be obtained (see subsection Interpretability of the models below and Additional file 1).

6.2.5 Model training

The split sample approach was employed; data was split randomly into two complementary parts, a training set (2/3) and a test set (1/3) under the same event/censoring proportions. To tune a model, 5-fold cross validation was performed in the training set for the machine learning techniques (and for Cox LASSO). Training data was divided into 5 folds. Each time 4 folds were used to train a model and the remaining fold was used to validate

its performance and the procedure was repeated for all combination of folds. Tuning of the hyper-parameters was done using grid search and performance of final models was assessed on the test set. Analyses were performed in R programming language version 3.5.3 [22]. Package of implementation for RSFs and NNs as well as technical details regarding the choice of tuning parameters and the cross-validation procedure for each method are provided in Additional file 2.

6.2.6 Assessing predictive performance on test data

To assess the final predictive performance of the models the concordance index, the Brier score, and the Integrated Brier Score (IBS) were applied.

The most popular measure of model performance in a survival context is the concordance index [38] which computes the proportion of pairs of observations for which the survival times and model predictions order are concordant taking into account censoring. It takes values typically in the range 0.5 - 1 with higher values denoting higher ability of the model to discriminate and 0.5 indicating no discrimination. The C-index cannot be defined for neural network models since it relies on the ordering of individuals according to prognosis and there is no unique ordering between the subjects. At one year individual i may have better survival probability than individual j, but this could be reversed for a different time point.

The C-index provides a rank statistic between the observations that is not time-dependent. Following van Houwelingen and le Cessie [39] a time-dependent prediction error is defined as

$$\textit{Brier}(y, \hat{S}(t_0|x)) = (y - \hat{S}(t_0|x))^2, \tag{6.9}$$

where $\hat{S}(t_0|x)$ is the model-based probabilistic prediction for the survival of an individual beyond t_0 given the predictor x, and $y = 1\{t > t_0\}$ is the actual observation ignoring censoring. The expected value with respect to a new observation Y_{new} under the true model $S(t_0|x)$ can be written as:

$$E[\textit{Brier}(Y_{new}, \hat{S}(t_0|x))] = S(t_0|x)(1 - S(t_0|x)) + (S(t_0|x) - \hat{S}(t_0|x))^2. \tag{6.10}$$

The Brier Score consists of two components: the "true variation" $S(t_0|x)(1 - S(t_0|x))$ and the error due to the model $(S(t_0|x) - \hat{S}(t_0|x))^2$. A perfect prediction is only possible if $S(t_0|x) = 0$ or $S(t_0|x) = 1$. In practice the two components cannot be separated since the true $S(t_0|x)$ is unknown.

To assess the performance of a prediction rule in actual data, censored observations before time t_0 must be considered. To calculate Brier Score when censored observations are present, Graf proposed the use of inverse probability of censoring weighting [40]. Then an estimate of the average prediction error of the prediction model $\hat{S}(t|x)$ at time $t = t_0$ is

$$Err_{Score}(\hat{S}, t_0) = \frac{1}{n} \sum_i 1\{d_i = 1 \vee t_i > t_0\} \frac{Score(1\{t_i > t_0\}, \hat{S}(t_0|x_i))}{\hat{C}(\min(t_i -, t_0)|x_i)} \tag{6.11}$$

In (6.11), $\frac{1}{\hat{C}(\min(t_i -, t_0)|x_i)}$ is a weighting scheme known as inverse probability of censoring weighting (IPCW) and $Score$ is the Brier Score for the prediction model. It ranges typically from 0 to 0.25 with a lower value meaning smaller prediction error.

Brier score is calculated at different time-points. An overall measure of prediction error is the Integrated Brier Score (IBS) which can be used to summarise the prediction error over the whole range up to the time horizon $\int_0^{t_{hor}} Err_{Score}(\hat{S}, t_0) dt_0$ (here t_{hor} = 10 years) [41]. IBS provides the cumulative prediction error up to t_{hor} at all available times ($t^* = 1, 2, \cdots, 10$ years) and takes values in the same range as the Brier score. In this study, we use IBS as the main criterion to evaluate the predictive ability of all models up to 10 years.

6.2.7 Interpretability of the models

Interpretation of models is of great importance for the medical community. It is well known that Cox models offer a straightforward interpretation through hazard ratios.

For neural networks with one hidden layer the connection weights algorithm by Garson [37] – later modified by Goh [42] – can provide information about the mechanism of the weights. The idea behind this algorithm is that inputs with larger connection weights produce greater intensities of signal transfer. As a result, these inputs will be more important for the model. Garson's algorithm can be used to determine relative importance of each input variable, partitioning the weights in the network. Their absolute values are used to specify percentage of importance. Note that the algorithm does not provide the direction of relationships, so it remains uncertain whether the relative importance indicates a positive or a negative effect. For details about the algorithm see [43]. During this work, the algorithm was extended for 2 hidden layers to obtain the relative importance of each variable (for the implementation see algorithm 1 on Additional file 1).

Random survival forest relies on two methods which can provide interpretability: variable importance (VIMP) and minimal depth [44]. The former is associated with the prediction error before and after the permutation of a prognostic factor. Large importance values indicate variables with strong predictive ability. The latter is related to the forest topology as it assesses the predictive value of a variable by computing its depth compared to the root node of a tree. VIMP is more frequently reported than minimal depth in the literature [45]. For both methods interpretation is available only for variable entities and not for each variable level.

6.3 Results

Administrative censoring was applied to the UNOS data at 10 years. Median follow-up is equal to 5.36 years (95% CI: 5.19 - 5.59 years) and it was estimated with reverse Kaplan-Meier [46]. Clinical endpoint is overall graft-survival (OGS). From the total number of patients, 69.1% was alive/censored and 30.9% experienced the event of interest (graft-failure or death). 3 models were used from the Cox family to predict survival outcome: a) a model with all 97 prognostic factors, b) a model with backward selection and c) a model based on the LASSO method for variable selection. Furthermore, 3 machine learning methods were employed: a) a random survival forest, b) a NN with one hidden layer and c) a NN with two hidden layers.

6.3.1 Comparisons between models

In this section a direct comparison of the 6 models is illustrated in terms of variable importance on the training set and predictive performance on the test set. Specification of the variables with dummy coding included 119 variable levels from the 97 potentially prognostic factors. For NNs - to apply and extend the methodology of Biganzoli - follow-up time was divided into 10 time intervals $(0, 1], (1, 2], \cdots, (9, 10]$ denoting years since transplantation. For Cox models and RSF exact time points were used.

Cox model assumes that each covariate has a multiplicative effect in the hazard function (which is constant over time). Estimating a model with 97 prognostic factors leads inevitably to a violation of the proportional hazards assumption for some covariates (17 out of 97 here). This means that hazard ratios for those risk factors are the mean effects on the outcome which is still a valuable information for the clinicians. To consider all possible non-linear effects on interactions leads to a complex model where too many parameters need to be estimated and the interpretability becomes very difficult. On the other hand, ML techniques do not make any assumptions about the data structure and therefore their performance is not affected by the violation of PH. The backward and the LASSO methods selected 28 (out of 97) and 45 predictors (out of 119 dummy coded), respectively. Selection of a smaller set of variables by Cox backward was expected, since it is a greedier (heuristic) method than LASSO penalized regression. The 12 most influential variables for the Cox model with all variables were selected by both methods

(see table 6.2). 5 of these variables: *re-transplantation*, *donor type*, *log(Total cold ischemic time)*, *diabetes* and *pre-treatment status* violated the PH assumption.

5-fold cross-validation in the training data resulted in the following optimal hyper-parameters combinations for the machine learning techniques:

- For the Random Survival Forest `nodesize` = 50, `mtry` = 12, `nsplit` = 5 and `ntree` = 300. Stratified bootstrap sub-sampling of half the patients was used per tree (due to the large training time required).
- For the neural network with 1 hidden layer `activation function` = "sigmoid" (for the input-hidden layer), `node size` = 85, `dropout rate` = 0.2, `learning rate` = 0.2, `momentum` = 0.9 and `weak class weight` = 1.
- For the neural network with 2 hidden layers `activation function` = "sigmoid" (for the input-hidden 1 and the hidden 1-hidden 2 layers), `node size` = 110, `dropout rate` = 0.1, `learning rate` = 0.2, `momentum` = 0.9 and `weak class weight` = 1.

6.3.2 Global performance measures

The global performance measures on test data are provided in Table 6.1. Examining the Integrated Brier Score (IBS), the NNs with 1 and with 2 hidden layers have the lowest (IBS = 0.180) followed by the RSF (IBS = 0.182). Cox models have a comparable performance (IBS = 0.183). Therefore, the predictive ability of Cox backward and Cox LASSO is the same as the less parsimonious Cox model with all variables in terms of IBS. The best model in terms of C-index is the Random Survival Forest (0.622) while the Cox models with all variables has slightly worse performance. C-index for Cox backward and Cox LASSO are respectively 0.615 and 0.614.

	IBS	C-index
Cox all variables	0.183	0.620
Cox backward	0.183	0.615
Cox LASSO	0.183	0.614
RSF	0.182	**0.622**
Neural Network 1h	**0.180**	-
Neural Network 2h	**0.180**	-

Table 6.1: Integrated Brier Score (IBS) and C-index on the test data. Neural network 1h and 2h refer to a neural network with one and two hidden layers respectively.

Stability of the networks was investigated by rerunning the same models on the test data, and showed that the NN with 1 hidden layer had stable predictive performance and variable importance. In contrast, the NN with 2 hidden layers was quite unstable regarding variable importance. This behavior might be related to the vast amount of weights that had to be trained for this model which can lead to overfitting (in total 26621 connection weights were estimated for a sample size of 41530 patients in long format; whereas for the NN with 1 hidden layer 11136 connection weights). For the RSF, model obtained remarkable stability in terms of performance error after a particular number of trees (`ntree` = 300 was selected).

6.3.3 Prediction error curves

Figure 6.1 shows the average prediction Brier error over time for all models. Small differences can be observed between Cox models and RSF. The NNs with 1 hidden and with 2 hidden layers have almost identical evolution over time achieving better performance than the Cox models and the RSF.

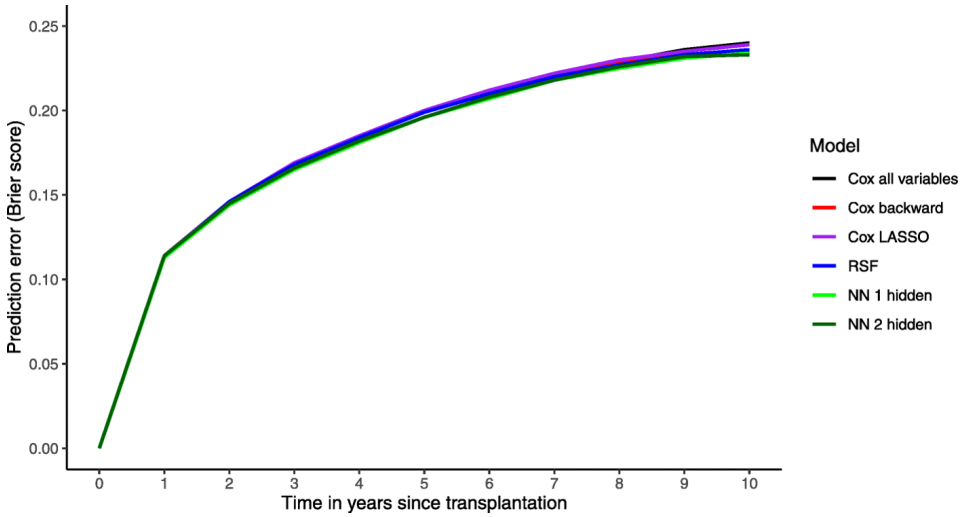

Figure 6.1: Prediction error curves for all models.

6.3.4 Variable importance

	Cox all variables HR (95% CI)	Cox backward HR (95% CI)	Cox LASSO HR
Re-transplantation	1.602 (1.491-1.721)	1.608 (1.501-1.722)	1.558
Donor age	1.010 (1.008-1.011)	1.011 (1.009-1.012)	1.009
Donor type DCD$^{(a)}$	1.483 (1.362-1.616)	1.443 (1.338-1.556)	1.298
log(Total cold ischemic time)	1.258 (1.192-1.327)	1.285 (1.221-1.353)	1.191
Diabetes	1.173 (1.125-1.225)	1.176 (1.128-1.226)	1.136
Race Black$^{(b)}$	1.240 (1.171, 1.314)	1.261 (1.193-1.332)	1.186
Life support	1.343 (1.240-1.454)	1.375 (1.272-1.487)	1.304
Recipient age	1.007 (1.005-1.009)	1.008 (1.006-1.010)	1.006
Incidental tumour	1.314 (1.202, 1.437)	1.315 (1.203-1.437)	1.203
Hypertensive bleeding	1.296 (1.185, 1.418)	1.301 (1.190-1.423)	1.214
HCV$^{(c)}$ serology status	1.147 (1.091-1.206)	1.148 (1.094-1.205)	1.166
Pre-treatment status ICU$^{(d)}$	1.240 (1.143, 1.346)	1.253 (1.160-1.354)	1.164

(a): Donor type DCD (Donor Circulatory Dead) vs DBD (Donor after Brain-Dead), (b): Race Black vs White,

(c): Chronic hepatitis C virus, (d): Intense Care Unit vs Non-hospitalised/Hospitalised

Table 6.2: Hazard ratios along with their 95% confidence intervals for the 12 most influential variables for the Cox models. Variables are presented in decreasing order according to the absolute z-score values (12.90 to 5.16) for the Cox model with all variables. Predictors shown are the most prognostic as their z-scores values correspond to low and very significant p-values. These variables were also selected by both Cox backward and Cox LASSO model which verifies their prognostic ability for Cox models.

In this section, the models are compared based on the most prognostic variables identified from the set of 97 predictors - 52 donor and 45 recipient characteristics. Hazard ratios of the 12 most prognostic variables for the Cox models are shown in Table 6.2, based on the absolute z-score values for the Cox model with all variables. The strongest predictor is *re-transplantation*. Having been transplanted before increases the hazard of graft-failure or death by more than 55%. The other most detrimental variables are *donor age* and *donor type circulatory dead*. One unit increase for donor age rises the hazard by around 1% while having received the graft from a donor circulatory versus brain-dead increases the hazard by more than 29% for all models. The rest of the factors which have an adverse effect are: *cold ischemic time*, *diabetes*, *race*, *life-support*, *recipient age*, *incidental tumour*, *spontaneous hypertensive bleeding*, *serology status of HCV* and *intense care unit before the operation*.

Neural network 1h	Rel-Imp	Neural network 2h	Rel-Imp	RSF	VIMP
Re-transplantation	0.035	Re-transplantation	0.028	Donor age	0.010
Life-support	0.025	HCV[(d)] serology status	0.025	Re-transplantation	0.009
Pre-treatment status ICU[(a)]	0.023	Life-support	0.024	Life support	0.007
Donor type DCD[(b)]	0.023	Donor age	0.023	HCV[(d)] serology status	0.007
Race Black[(c)]	0.022	Diabetes	0.021	Pre-treatment status	0.006
HCV[(d)] serology status	0.022	Pre-treatment status ICU[(a)]	0.020	Recipient age	0.004
Diabetes	0.020	Working income	0.020	Aetiology	0.003
Donor age	0.020	Race Black[(c)]	0.019	log(Last serum creatinine)	0.003
Working income	0.018	Previous abdominal surgery	0.015	Functional status	0.002
Functional status Total assistance[(e)]	0.017	Donor pre-recovery diuretics	0.015	log(Total cold ischemic time)	0.002
Aetiology HCV	0.017	Aetiology Cholestatic	0.011	Race	0.002
Hypertensive bleeding	0.017	Functional status Total assistance[(e)]	0.015	Diabetes	0.002

(a): Intense Care Unit vs Non-hospitalised/Hospitalised (b): Donor type DCD (Donor Circulatory Dead) vs DBD (Donor after Brain-Dead),

(c): Race Black vs White, (d): Chronic hepatitis C virus, (e): Total assistance vs No assistance

Table 6.3: The 12 most prognostic factors for the neural networks with 1 and 2 hidden layers (Rel-Imp: relative importance) and for the Random Survival Forest (VIMP: variable importance). Note that the NN utilises time intervals as one of the input variables (check the contribution of time intervals in Table 1 of Additional file 1). For RSF importance is measured for each variable without distinction for each level.

In Table 6.3 the most prognostic factors for the machine learning techniques are presented. The top predictors are provided in terms of relative importance (Rel-Imp) for the PLANN models and in terms of variable importance (VIMP) for the RSF. For the NNs, the strongest predictor is *re-transplantation* (Rel-Imp 0.035 for 1 hidden and 0.028 for 2 hidden layers), which is the second strongest for the RSF (VIMP 0.009). According to the tuned RSF, the most prognostic factor for the overall graft-survival of the patient is *donor age* (VIMP 0.010).

Other strong prognostic variables for the NN with 1 hidden layer are *life support* (Rel-Imp 0.025), *intense care unit before the operation* (Rel-Imp 0.023) and *donor type circulatory dead versus brain-dead* (Rel-Imp 0.023). For the NN with 2 hidden layers other very prognostic variables are *serology status for HCV* (Rel-Imp 0.025), *life support* (Rel-Imp 0.024) and donor age (Rel-Imp 0.023).

For the RSF *life support* (VIMP 0.007), *serology status for HCV* (VIMP 0.007) and *intense care unit before the operation* (VIMP 0.006). Note that variable *total cold ischemic time* which was identified as the 4th most prognostic for the Cox model with all variables and the 10th most prognostic for random survival forest is not in the list of the 12 most prognostic for both NNs.

6.3.5 Individual predictions

In this section, the predicted survival probabilities are compared for 3 new hypothetical patients and 3 patients from the test data.

In Figure 6.2a) the patient with reference characteristics shows the best survival. The highest probabilities are predicted by the RSF and the lowest by the Cox model. The same pattern occurs for the patient that suffers from diabetes (orange lines). The patient with diabetes who has been transplanted before has the worst survival predictions. In this case the NN predicts the highest survival probabilities and the Cox model built using all the prognostic factors the lowest.

In Figure 6.2b) the estimated survival probabilities are showed by the Cox model with all variables, the tuned RSF and the tuned PLANN with 1 hidden layer for 3 patients from the test set. The first patient shows the highest survival predictions by the 3 models. The RSF provides the highest survival probabilities and the NN the lowest. The second patient experiences lower survival probabilities (orange lines) whereas the third patient shows the lowest survival probabilities overall. For the second patient the NN predicts the lowest survival probabilities over time and for the third the Cox model.

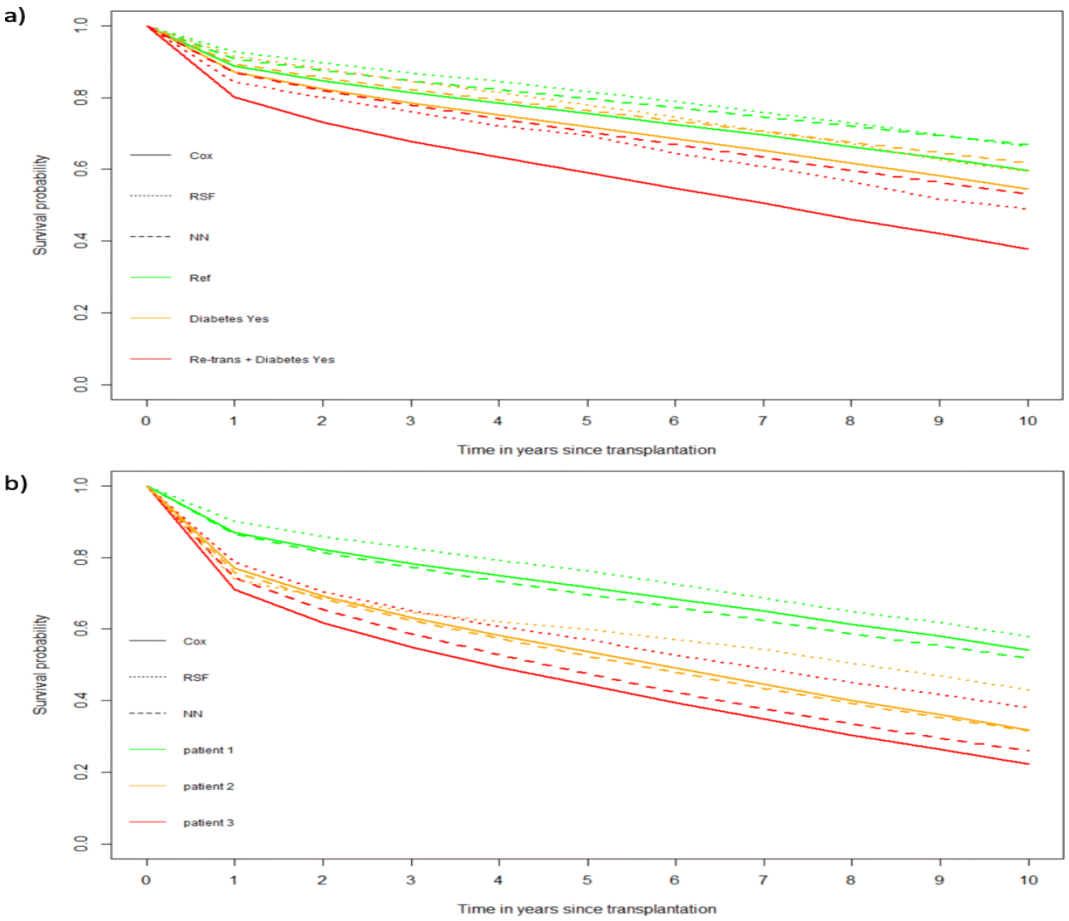

Figure 6.2: **a)** Predicted survival probabilities for 3 new hypothetical patients using the Cox model with all variables (solid lines), the tuned RSF (short dashed lines) and the tuned NN with 1 hidden layer (long dashed lines). The green lines correspond to a reference patient with the median values for the continuous and the mode value for categorical variables. The patient in the orange line has diabetes (the other covariates as in reference patient). The patient in the red line has been transplanted before and has diabetes simultaneously (the other covariates as in reference patient). Values for 10 prognostic variables for the reference patient are provided in Table 2 of Additional file 1.
b) Predicted survival probabilities for 3 patients selected from the test data based on the Cox model with all variables (solid lines), the tuned RSF (short dashed lines) and the tuned NN with 1 hidden layer (long dashed lines). Green lines correspond to a patient censored at 1.12 years. Patient in the orange line was censored at 6.86 years. Patient in the red line died at 0.12 years. Values for 10 prognostic variables for the patients are provided in Tables 3-5 of Additional file 1.

In general, the random survival forest provides the most optimistic survival probabilities whereas the most pessimistic survival probabilities are predicted by either the Cox model or the NN (more often by the Cox model). This may be related to the characteristics of the methods as RSF relies on recursive binary partitioning of predictors, whereas Cox models imply linearity, and NNs fit non-linear relationships.

6.3.6 Calibration

Here 4 methods are compared: Cox model with all variables, RSF, PLANN 1 hidden and 2 hidden layers based on the calibration on the test data. For each method, the predicted survival probabilities at each year are estimated and the patient data are split into 10 equally sized groups based on the deciles of the probabilities. Then the survival probabilities along with their 95% confidence intervals are calculated using the Kaplan-Meier methodology [47].

In figure 6.3 the results are showed at 2 years since LT. The Cox model with all variables and the PLANN with 1 hidden layer are both well calibrated. The RSF and the PLANN with 2 hidden layers tend to overestimate the

survival probabilities for the patients at higher risk. Survival neural network with 1 hidden layer seems to be the most reliable for predictions between the ML techniques. Calibration plots at 5 and 10 years can be found in Additional file 3.

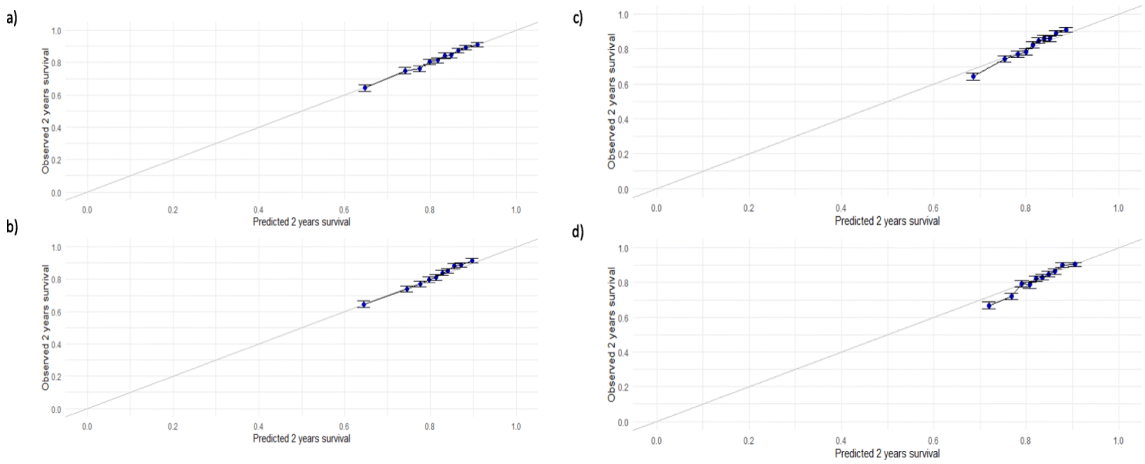

Figure 6.3: Calibration plots at 2 years on the test data: **a)** Cox model with all variables, **b)** Random Survival Forest, **c)** Partial Logistic Artificial Neural Network with 1 hidden layer, **d)** Partial Logistic Artificial Neural Network with 2 hidden layers.

6.4 Discussion

With the rise of computational power and technology on the 21^{st} century, more and more data have been collected in the medical field to identify trends and patterns which will allow building better allocation systems for patients, provide more accurate prognosis and diagnosis as well as more accurate identification of risk factors. During the past few years, machine learning (ML) has received increased attention in the medical area. For instance, in the area of LTs graft failure or primary non-function might be predicted at decision time with ML methodology [48]. Briceño *et al.* created a NN process for donor-recipient matching specifying a binary classification survival output (recipient or graft survival) to predict 3-month graft mortality [49].

In this study statistical and ML models were estimated for patients from the US post-transplantation. Random survival forest performed better than Cox models with respect to the C-index. This shows the ability of the model to discriminate between low and high risk groups of patients. The C-index was not estimated for NN because a natural ordering of subjects is not feasible. Therefore, the Brier score was measured each year for all methods. The RSF showed similar results to the Cox models having slightly smaller total prediction error (in terms of IBS). The NNs performed in general better than the Cox models or the RSF and had very similar performance over time. RSF and survival NN are ML techniques which have a different learning approach and model non-linear relationships between variables automatically. Both methods may be used in medical application but should be applied at present as additional analysis for comparison.

Special emphasis was given on the interpretation of the models. An indirect comparison was performed to examine which are the most prognostic variables for a Cox model with all variables, a RSF and NNs. Results showed that Cox model with all variables (via absolute z-score values) and the NNs with one/two hidden layer(s) (via relative importance) identified similar predictors. Both methods identified *re-transplantation* as the strongest predictor and *donor age*, *diabetes*, *life support* and *race* as relatively strong predictors. According to RSF, the most prognostic variables were *donor age*, *re-transplantation*, *life support* and *serology status of HCV*. *Aetiology* and *last serum creatinine* were selected as the 7^{th} and the 8^{th} most prognostic. This raises a known concern about the RSF bias towards continuous variables and categorical variables with multiple levels [50] (*aetiology* has 9 levels: metabolic, acute, alcoholic, cholestatic, HBV, HCV, malignant, other cirrhosis, other unknown). As continuous and multilevel variables incorporate larger amount of information than categorical, they tend to be favoured by the splitting rule of the forest during binary partitioning. Such bias was reflected in the variable importance results.

When comparing statistical models with machine learning techniques with respect to interpretability, Cox models offer a straightforward interpretation through the hazard ratios. On the contrary, for both neural networks and random survival forests the sign of the prediction is not provided (if the effect is positive or negative). Additionally, for NNs interpretation is possible for different variable levels (with the method of Garson and its extension), whereas for RSF only the total effect of a variable is shown. There is no common metric to directly compare Cox models with ML techniques in terms of interpretation. Future research in this direction is needed.

ML techniques are inherently based on mechanisms introducing randomisation and therefore very small changes are expected between different iterations of the same algorithm. To evaluate stability of performance, ML models were run several times under the same parametrisation. RSF were consistently stable after a certain number of trees (300 were selected). This was not the case for the NNs where instability is a common problem. It is challenging to tune a NN due to many hyper-parameter combinations available and the lack of a consistent global performance measure for survival data. IBS was used to tune the novel NNs, which may be the reason of instability for the NN with 2 hidden layers together with the large number of weights. Note also that the NN with 1 hidden layer is well calibrated whereas the NN with 2 hidden layers is less calibrated on the test data.

This is the first study where ML techniques are applied to transplant data where a comparison with the traditional Cox model was investigated. To construct the survival NN, the original problem had to be converted into a classification problem where exact survival times were transformed into (maximum) 10 time intervals denoting years since transplantation. On the other hand, for the Cox models and the RSF exact time to event was used. Recently, a new feed forward NN has been proposed for omics data which calculates directly a proportional hazards model as part of the output node using exact time information [51]. A survival NN with exact times may lead to better predictive performance. For UNOS data, 69.1% of the recipients were alive/censored and 30.9% had the event of interest. Results above were based on these particular percentages for censoring and events (for the NNs the percentages varied because of the reformulation of the problem).

It might be useful to investigate how the number of variables affects the performance of the models. Here 97 variables were pre-selected supported by clinical and statistical reasons (e.g. variables available before or during LT). It might be interesting to repeat the analyses on a smaller group of predictors, implementation time can be drastically reduced as the calculation complexity depends on sample size and predictors multiplicity. Alongside, predictive accuracy might be increased as some noisy factors will be removed from the dataset increasing the signal of potentially prognostic variables.

Both traditional Cox models and PLANNs allow for the inclusion of time-dependent covariates. For PLANNs, each patient is replicated multiple times during the transformation of exact times into a set of k non-overlapping intervals in long format. Thus, different values of a covariate can be naturally incorporated to increase the predictive ability of the networks. It would be interesting to apply and compare the predictive ability of time-dependent Cox models and PLANNs to liver transplantation data including explanatory variables whose values change over time. Such extension to more dynamic methods may increase predictive performance and help in decision making.

6.5 Conclusions

There is an increased attention to ML techniques beyond SM in the medical field with methods and applications being more necessary than ever. Utilization of these algorithmic approaches can lead to pattern discovery in the data promoting fast and accurate decision making. For time-to-event data, more ML techniques may be applied for prediction such as Support Vector Machines and Bayesian Networks. Moreover, deep learning with NN is gaining more and more attention and will likely be another trend in the future for these complex data.

In this work two alternatives to the Cox model from machine learning for medical data with large total sample size (62294 patients) and many predictors (97 in total) were discussed. RSF showed better performance than the Cox models with respect to C-index so it can be a useful tool for prioritisation of particular high risk patients. NNs showed better prediction performance in terms of Integrated Brier score. However, both ML techniques required

a non-trivial implementation time. Cox models are preferable in terms of straightforward interpretation and fast implementation. Our study suggests that some caution is required when ML methods are applied to survival data. Both approaches can be used for exploratory and analysis purposes as long as the advantages and the disadvantages of the methods are presented.

List of abbreviations

BS, Brier score; CVPL, cross-validated log-partial likelihood; DCD, Donor Circulatory Dead; HBV, Chronic hepatitis B virus; HCV, Chronic hepatitis C virus; IBS, Integrated Brier score; IPCW, Inverse Probability of Censoring Weighting; LASSO, least angle and selection operator; LT, liver transplantation; LUMC, Leiden University Medical Center; ML, machine learning; NN(s), artificial neural network(s); OGS, overall graft-survival; OPO, Organ Procurement Organisations; OPTN, Organ Procurement and Transplantation Network; PLANN, partial logistic artificial neural network; PLANN-ARD, partial logistic artificial neural network - automatic relevance determination; PH, proportional hazards; Rel-Imp, relative importance; RSF, random survival forest; SM, statistical model; SRTR, Scientific Registry of Transplant Recipients; UNOS, United Network of Organ Sharing; VIMP, variable importance.

Declarations

Availability of data and materials

The research data for this project is private. Unauthorized use is a violation of the terms of the Data Use Agreement with the U.S. Department of Health and Human Services. More information and instructions for researchers to request UNOS data can be found at https://unos.org/data/. R-code developed to perform the analysis is available at https://github.com/GKantidakis/Survival-prediction-models-since-liver-transplantation.

Competing interests

The data reported here have been supplied by the Minneapolis Medical Research Foundation (MMRF) as the contractor for the Scientific Registry of Transplant Recipients (SRTR). The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

This study used data from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procurement and Transplantation Network (OPTN). The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors.

Funding statement

Georgios Kantidakis's work as a Fellow at EORTC Headquarters was supported by a grant from the EORTC Soft Tissue and Bone Sarcoma Group and the Leiden University Medical Center (LUMC) Department of Medical Oncology as well as from the EORTC Cancer Research Fund (ECRF). The funding sources had no role in the design of the study and collection, analysis, and interpretation of data or preparation of the manuscript.

Acknowledgements

The authors would like to thank the United Network of Organ Sharing (UNOS) and Scientific Registry of Transplant Recipients (SRTR) for providing the data about liver transplantation to Leiden University Medical Center (LUMC) under DUA number 9477.

Online supplementary materials

The Additional files of this Chapter are available online at https://github.com/GKantidakis/Thesis_supplementary_materials/tree/main/Chapter6.

References

[1] J. M. Grinyó. Why is organ transplantation clinically important? *Cold Spring Harbor Perspectives in Medicine*, 3(6), 2013. doi: 10.1101/cshperspect.a014985.

[2] R. M. Merion, D. E. Schaubel, D. M. Dykstra, R. B. Freeman, F. K. Port, and R. A. Wolfe. The survival benefit of liver transplantation. *American Journal of Transplantation*, 5(2):307–313, 2005. doi: 10.1111/j.1600-6143.2004.00703.x.

[3] X. Song, A. Mitnitski, J. Cox, and K. Rockwood. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. *Stud Health Technol Inform.*, 107(Pt 1):736–740, 2004.

[4] R. C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, nov 2015. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.115.001593.

[5] K. Shailaja, B. Seetharamulu, and M. A Jabbar. Machine Learning in Healthcare: A Review. In *Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 910–914, Coimbatore, 2018. doi: 10.1109/ICECA.2018.8474918.

[6] I. A. Scott, D. Cook, E. W. Coiera, and B. Richards. Machine learning in clinical practice: prospects and pitfalls. *Medical Journal of Australia*, 211(5):203–205, sep 2019. ISSN 13265377. doi: 10.5694/mja2.50294.

[7] R. J. Desai, S. V. Wang, M. Vaduganathan, T. Evers, and S. Schneeweiss. Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. *JAMA network open*, 3(1):e1918962, 2020. doi: 10.1001/jamanetworkopen.2019.18962.

[8] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. URL http://www.jstor.org/stable/2985181.

[9] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–86, 1998. doi: 10.1002/(sici)1097-0258(19980530)17:10<1169::aid-sim796>3.0.co;2-d.

[10] P. Wang, Y. Li, and C. K. Reddy. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys*, 51(6), 2019. doi: 10.1145/3214306.

[11] A. Xiang, P. Lapuerta, A. Ryutov, J. Buckley, and S. 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. doi: https://doi.org/10.1016/S0167-9473(99)00098-5. URL www.elsevier.com/locate/csda.

[12] D. Faraggi and R. Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1):73–82, 1995. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780140108.

[13] Liestøl K., Andersen P. K., and Andersen U. Survival analysis and neural nets. *Statistics in Medicine*, 13 (12):1189–1200, 1994. doi: 10.1002/sim.4780131202.

[14] J. Buckley and I. James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979. doi: https://doi.org/10.1093/biomet/66.3.429.

[15] P. J. G. Lisboa, H. Wong, P. Harris, and R. Swindell. A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial Intelligence in Medicine*, 28(1):1–25, 2003. ISSN 09333657. doi: 10.1016/S0933-3657(03)00033-2.

[16] E. Biganzoli, P. Boracchi, and E. Marubini. A general framework for neural network models on censored survival data. *Neural Networks*, 15(2):209–18, 2002. doi: 10.1016/s0893-6080(01)00131-9. URL www.elsevier.com/locate/neunet.

[17] A. Biglarian, E. Bakhshi, A. R. Baghestani, M. R. Gohari, M. Rahgozar, and M. Karimloo. Nonlinear survival regression using artificial neural network. *Journal of Probability and Statistics*, 2013, 2013. doi: https://doi.org/10.1155/2013/753930.

[18] A. S. Jones, A. G. F. Taktak, T. R. Helliwell, J. E. Fenton, M. A. Birchall, D. J. Husband, and A. C. Fisher. An artificial neural network improves prediction of observed survival in patients with laryngeal squamous carcinoma. *European Archives of Oto-Rhino-Laryngology*, 263(6):541–547, jun 2006. doi: 10.1007/s00405-006-0021-2.

[19] A. Taktak, L. Antolini, M. Aung, P. Boracchi, I. Campbell, B. Damato, E. Ifeachor, N. Lama, P. Lisboa, C. Setzkorn, V. Stalbovskaya, and E. Biganzoli. Double-blind evaluation and benchmarking of survival models in a multi-centre study. *Computers in Biology and Medicine*, 37(8):1108–1120, 2007. doi: 10.1016/j.compbiomed.2006.10.001.

[20] J. J. Blok, H. Putter, H. J. Metselaar, R. J. Porte, F. Gonella, J. De Jonge, A. P. Van den Berg, J. Van Der Zande, J. D. De Boer, B. Van Hoek, and A. E. Braat. Identification and validation of the predictive capacity of risk factors and models in liver transplantation over time. *Transplantation Direct*, 4(9), 2018. doi: 10.1097/TXD.0000000000000822.

[21] J. D. de Boer, H. Putter, J. J. Blok, I. P. J. Alwayn, B. van Hoek, and A. E. Braat. Predictive Capacity of Risk Models in Liver Transplantation. *Transplantation Direct*, 5(6):e457, 2019. doi: 10.1097/TXD.0000000000000896.

[22] R Core Team. R: A Language and Environment for Statistical Computing, 2014. URL http://www.r-project.org/.

[23] G. Kantidakis, C. Lancia, and M. Fiocco. *Prediction models for liver transplantation – comparisons between Cox models and machine learning techniques [abstract OC30-4]*. 40th Annual Conference of the International Society for Clinical Biostatistics, 2019. URL https://kuleuvencongres.be/iscb40/images/iscb40-2019-e-versie.pdf.

[24] S. Van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18(6):681–694, 1999. doi: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R.

[25] D. J. Stekhoven and P. Bühlmann. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. doi: 10.1093/bioinformatics/btr597.

[26] J. F. Lawless and K. Singhal. Efficient Screening of Nonnormal Regression Models. *Biometrics*, 34(2): 318–327, jun 1978. doi: 10.2307/2530022. URL https://www.jstor.org/stable/2530022?origin=crossref.

[27] R. Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4): 385–395, 1997. URL https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4{%}3C385::AID-SIM380{%}3E3.0.CO;2-3.

[28] P. J. M. Verweij and H. C. Van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12(24):2305–2314, dec 1993. doi: 10.1002/sim.4780122407. URL http://doi.wiley.com/10.1002/sim.4780122407.

[29] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. *Annals of Applied Statistics*, 2(3):841–860, 2008. doi: 10.1214/08-AOAS169.

[30] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL http://link.springer.com/10.1023/A:1010933404324.

[31] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/978-0-387-84858-7. URL http://link.springer.com/10.1007/978-0-387-84858-7.

[32] M. R. Segal. Regression Trees for Censored Data. *Biometrics*, 44(1):35–47, 1988. URL http://www.jstor.org/stable/2531894.

[33] T. Hothorn and B. Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137, 2003. doi: 10.1016/S0167-9473(02)00225-6.

[34] M. van Gerven and S. Bohte. Editorial: Artificial Neural Networks as Models of Neural Information Processing. *Frontiers in Computational Neuroscience*, 11:114–114, 2017. doi: 10.3389/fncom.2017.00114. URL http://journal.frontiersin.org/article/10.3389/fncom.2017.00114/full.

[35] M. Minsky and S. Papert. *Perceptrons; an introduction to computational geometry*. MIT Press, Cambridge, MA, 1 edition, 1969. ISBN 9780262130431.

[36] P. Lapuerta, Azen S. P., and LaBree L. Use of neural networks in predicting the risk of coronary artery disease. *Computers and Biomedical Research*, 28(1):38–52, 1995. doi: 10.1006/cbmr.1995.1004.

[37] G. D. Garson. Interpreting Neural Network Connection Weights. *AI Expert*, 6(4):46–51, 1991.

[38] F. E. Harrell, K. L. Lee, and D. 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. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[39] J. C. Van Houwelingen and S. Le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 9(11): 1303–1325, 1990. doi: https://doi.org/10.1002/sim.4780091109.

[40] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–45, 1999. doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5. URL http://www.ncbi.nlm.nih.gov/pubmed/10474158.

[41] J. C. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2012. ISBN 9781439835333. URL https://www.crcpress.com/Dynamic-Prediction-in-Clinical-Survival-Analysis/van-Houwelingen-Putter/p/book/9781439835333.

[42] A. T. C. Goh. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, 9(3):143–151, jan 1995. ISSN 0954-1810. doi: 10.1016/0954-1810(94)00011-S. URL https://www.sciencedirect.com/science/article/pii/095418109400011S.

[43] J. D. Olden and D. A. Jackson. Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, 154(1-2):135–150, 2002.

[44] H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105(489):205–217, 2010. doi: 10.1198/jasa.2009.tm08622.

[45] H. Ishwaran and M. Lu. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine*, 38(4):558–582, 2019. doi: 10.1002/sim.7803.

[46] M. Schemper and T. L. Smith. A Note on Quantifying Follow-up in Studies of Failure Time. *Control Clin Trials*, 17(4):343–6, 1996. doi: 10.1016/0197-2456(96)00075-x.

[47] E. L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi: 10.2307/2281868.

[48] 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, apr 2017. ISSN 0041-1337. doi: 10.1097/TP.0000000000001600. URL http://insights.ovid.com/crossref?an=00007890-201704000-00025.

[49] J. Briceño, M. Cruz-Ramírez, M. Prieto, M. Navasa, J. O. De Urbina, R. Orti, M. Á. Gómez-Bravo, A. Otero, E. Varo, S. Tomé, G. Clemente, R. Bañares, R. Bárcena, V. Cuervas-Mons, G. Solórzano, C. Vinaixa, Á. Rubín, J. Colmenero, A. Valdivieso, R. Ciria, C. Hervás-Martínez, and M. 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. ISSN 16000641. doi: 10.1016/j.jhep.2014.05.039.

[50] W. Y. Loh and Y. S. Shih. Split selection methods for classification trees. *Statistica Sinica*, 7:815–840, 1997. URL https://www.jstor.org/stable/24306157.

[51] T. Ching, X. Zhu, and L. X. Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS computational biology*, 14(4), 2018. doi: 10.1371/journal.pcbi.1006076.