

Enhanced survival prediction using explainable artificial intelligence in heart transplantation

Paulo Lisboa

Liverpool John Moores University

Manoj Jayabalan

Liverpool John Moores University

Sandra Ortega-Martorell

Liverpool John Moores University

Ivan Olier

Liverpool John Moores University

Dennis Medved

Lund University

Johan Nilsson (✉ johan.nilsson@med.lu.se)

Lund University

Article

Keywords: Interpretability, Generalised Additive Models, Self-explaining Neural Networks, Machine explanation

Posted Date: April 7th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1529645/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The most limiting factor in heart transplantation is lack of donor organs. With enhanced prediction of outcome, it may be possible to increase the life-years from the organs that become available. Applications of machine learning to tabular data, typical of clinical decision support, pose the practical question of interpretation which has technical and potential ethical implications. In particular, there is an issue of principle about the predictability of complex data and whether this is inherent in the data or strongly dependent on the choice of machine learning model, leading to the so-called accuracy-interpretability trade-off. We model one-year mortality in heart transplantation data with a self-explaining neural network, which is benchmarked against a deep learning model on the same development data, in an external validation study with two data sets: 1) UNOS transplants in 2017-2018 (n= 4,750) for which the self-explaining and deep learning models are comparable in their AUROC 0.628 [0.602,0.654] cf. 0.635 [0.609,0.662] and 2) Scandinavian transplants during 1997-2018 (n= 2,293) showing good calibration with AUROC of 0.626 [0.588,0.665] and 0.634 [0.570, 0.698] respectively with and without missing data (n=982). This shows that for tabular data predictive models can be transparent and capture important non-linearities, retaining full predictive performance.

Introduction

Heart transplantation is life-saving for patients with end-stage heart disease. A limiting factor is the lack of donor organs. Recently the United Network for Organ Sharing (UNOS) organ allocation system has been changed to a more granular distinction of medical urgency resulting in lower waitlist mortality. However, recent analysis has shown worse post-transplantation survival¹. To reduce post-transplantation mortality, the factors affecting post-transplantation survival must be better understood².

Artificial Intelligence is experiencing explosive growth in clinical applications from specialist data processing such as radiology and text mining to diagnostic and prognostic risk stratification with tabular data typically comprised of clinical factors, demographics and multiple indicators derived from physiological measurements. This opens the potential for more precise predictions that add to the evidence base used for clinical decision making. Since the value of quantitative outcome models is highest where data are complex and noisy, this raises the question of a trade-off between model complexity and predictive performance, particularly in real-world settings with high stakes outcomes.

This study is about the prioritisation of patients to receive a scarce and highly valuable resource, a donor's heart. For the clinician, the first priority is to have access to the most accurate predictive model. Linked to this is the ethical question of justifying the choice of the recipient based on all of the available evidence, now augmented with an AI-based prediction.

A recent review indicates that there is no appreciable performance loss for sparse, linear-in-the-parameters models compared with deep learning applied to tabular data, that is to say flat, unstructured data with a mixture of largely independent continuous and discrete variables, widely used in clinical

decision making³. This is especially the case when the signal-to-noise ratio is low. This finding echoes the view that machine learning models for high stakes decisions should be interpretable by design as part of a drive to encourage responsible machine learning governance⁴. Model transparency brings additional advantages when seeking to find and fix problems in data sets, such as control of bias, which is a key consideration for observational data frequently used in cross-sectional and cohort studies⁵.

Among the algorithmic challenges for interpretable machine learning Generalised Additive Models (GAM) are attractive as they are inherently interpretable, certainly to the same level of understanding as popular risk stratification algorithms such as logistic regression⁴. Additive models have generic value for interpretability⁶ and meet accepted criteria for robust interpretability laid out in the following desiderata⁷ - explicitness/intelligibility: *"Are the explanations immediate and understandable?"*; faithfulness: *"Are relevance scores indicative of "true" importance?"*; and stability: *"How consistent are the explanations for similar/neighbouring examples?"*. In addition, GAMs can achieve also parsimony/sparsity *"Do the explanatory variables comprise a minimal set?"* and consistency: *"How robust are the explanations to perturbations in the data?"*

The challenge is to maintain the well-known data fitting capabilities of neural networks without sacrificing predictive performance⁸.

The main novel contribution of the paper is to show, with a substantive external validation study, that this question can be addressed by a method to efficiently derive a self-explaining neural network model while preserving the predictive performance of a much larger deep learning model. We derive an accurate predictor in the form of a sparse model with meaningful features constructed through an iterative process. The method is generic for tabular data in real-world applications where non-linear effects are estimated in the presence of significant levels of noise.

The motivation for the Partial Response Network (PRN) is that multivariate functions can be written as a summation of functions of fewer variables, with interesting theoretical properties. This is done within the formal framework of functional Analysis of Variance (ANOVA) decompositions⁹ where the component functions are orthogonal with respect to a given metric. If the metric is a Dirac function, then the component functions in the ANOVA decomposition are derived from cuts of the multivariate function at a given point, which we choose to be the overall median of the data. This is known as an anchored ANOVA decomposition. In application to medical data, the key principle is that the derived functions of one or two variables represent main effects and pairwise interactions that, between them, can reasonably comprise much of the discriminant information that is required of the model. The methodology to derive and optimize these functions is explained in more detail in the Method section.

The first novel contribution of this paper is to benchmark the classification performance of the PRN model with that a previously published deep learning model (International Heart Transplant Survival Algorithm [IHTSA]) and a traditional scoring model (Index for Mortality Prediction After Cardiac Transplantation [IMPACT])^{10,11}. This study aims to assess the relative performance of the interpretable

model and a deep neural network with two external cohorts, one comprising recent heart transplants from the UNOS data and the other with transplant data from a regional database in Scandinavia. This is a gold standard benchmarking study since it uses three data cohorts: 1997-2013 to train the model and 2014-16 to find the hyperparameters that optimise model performance, but a subsequent cohort of patients operated in 2017-18 to calculate the generalisation performance and compare it between the different models.

The second novelty is to verify consistency between the independent effects identified by the PRN for individual variables and pairs of variables, and clinical expertise. Our study seeks to move the focus of discussion of machine learning models of clinical risk from mainly classification performance, where different models often have comparable performance, to clinical insights coupled with performance, meaning interpretability without compromising predictive accuracy.

A further benchmark included for completeness is to compare against an alternative interpretable method from the machine learning literature, Explainable Boosting Machines (EBMs)¹². In common with the PRN, this uses the structure of a GAM. However, it models the univariate and bivariate component functions using rule ensembles, resulting staggered, rather than smooth functions. Moreover, there is no constraint of orthogonality between the component functions, as is the case in the ANOVA decomposition that underpins the PRN. The model results are compared in the Discussion section.

Results

Patient demographics in all cohorts/Population

Clinically relevant information about transplanted patients collected from the UNOS database between 1997 and 2016 were used for model development (derivation cohort [DC], $n = 31,315$) and test cohort ([TC], $n = 6,120$). The third cohort of patients transplanted between 2017 and 2018 was used as an external blinded validation cohort (VC, $n = 4,750$), Fig. 1. The derivation and validation cohorts comprised 296,451 patient years (median survival time 12.1 years, IQR 5.3–19.2). The one-year mortality was 11.8% ($n = 4,978$). The three study cohorts, DC, TC, and the blinded external VC have a different distribution of most of the included variables. As shown in Table 1, the patients were older, heavier, and present with more comorbidities such as diabetes mellitus and need for dialysis in the VC compared with the DC. Furthermore, the presence of critical state variables such as ECMO, inotropic support, and mechanical assist were more common in the VC. Despite this, the one-year mortality was significantly lower in the VC (9.8%) compared with TC (10.8%) and DC (12.3%), $p < 0.001$ (Kruskal-Wallis test). On the other hand, the duration of donor heart ischemia, infection and need for a ventilator before transplantation were lower in the VC compared with the DC.

The starting pool of inputs for model selection by the PRN consists of variables commonly used to predict 1-year mortality (recipient age, donor age, creatinine, ischemic time, ventilator at orthotopic heart

transplantation (OHT), diagnosis of non-ischemic disease in recipients, recipient female gender, infection within 2 weeks, history of prior transplantation).

Supplementary Figs. 1-4 shows the deviance and values of beta for both Lasso models, one derived from the original MLP and the other from the PRN model derived from it. The duration of ischemia came in early on as part of a core set of five variables (donor age, ischemic time, creatinine, ventilator at OHT, transplant year) selected even for very high regularisation. This indicates that the duration of ischemia, together with the other variables above, are consistently important.

The final PRN-Lasso model includes nine univariate effects: recipient age, donor age, creatinine, ischemic time, ventilator at OHT, diagnosis of non-ischemic disease in recipients, recipient female gender, infection within 2 weeks, and transplant year, together with a 2-way interaction involving recipient age and diagnosis of ischemic cardiomyopathy (ICM) in recipients. The AUROC for the prediction of one-year mortality was 0.653 (95% CI: 0.643–0.662) in DC and 0.605 (95% CI: 0.582–0.628) in the TC. As illustrated in Fig. 2a, the recipient age contributes to the model most at younger and older ages so has a U-shape, whereas the donor age contribution is monotonically increasing and close to linear (Fig. 2b). The duration of ischemia starts to contribute significantly after 3 and a half hours (Fig. 2d). The 2-way interaction is of interest as it suggests a protective effect for ischemic disease in younger recipients, in contrast to the expected increase in mortality for older age. A crude calculation of the hazard ratio for one-year mortality by filtering recipients in the age groups 18-40 and 60-70 shows values of 0.117 and 0.124, respectively lower than and marginally higher than the overall prevalence of 0.123 in the DC. This is consistent with the 2-way effect shown in Fig. 2j which was found to be statistically significant by the Lasso.

Predictive performance

As shown in Table 2 and Fig. 3, the discrimination (AUROC) for the PRN-Lasso including 10 partial response was similar for the IHTSA and IHTSA recalibrated model in the blinded external VC, 0.628 (CI 95%: 0.602–0.654) vs 0.635 (CI 95%: 0.609–0.662), $p=0.488$, and 0.643 (CI 95%: 0.617–0.669), $p=0.197$. The Hosmer-Lemeshow (HL) chi-square in VC was 15.01 for the PRN model ($p=0.135$), suggesting that there was good calibration (Fig. 4a). The calibration for the IHTSA and IMPACT models was poor ($p<0.001$), Fig. 4d. The discrimination for the IHTSA and IHTSA recalibrated model compared with the IMPACT model was significantly superior, $p=0.023$ and $p=0.004$, respectively.

The machine learning benchmarking method applied to the external cohort 2017-18, EBM, has a high estimated ROC area of 0.634 (CI 95%: 0.607–0.660), selecting the following variables: donor age, ischemic time, recipient age, creatinine, infection within 2 weeks, ventilator at OHT, female gender of the recipient, transplant era and diagnosis of ICM. These variables are entirely consistent with those selected by the PRN, lacking only the interaction between the diagnosis of ICM and NICM.

Further external validation of the PRN-LASSO model was carried out using a regional data set from Scandia Thoracic Transplantation Database from Scandinavia ($n=2,293$). This time the timeframe

overlaps as the transplants took place over the period 1997-2018. There is also a significant proportion of missing data in all but 982 rows of data. Missing values were imputed using the median value for that column, as in data standardization this maps onto zero so that missing values do not contribute to the model prediction. The AUROC for the full data set imputed was 0.626 (CI 95%: 0.588–0.665), and for observed values only it was 0.634 (CI 95%: 0.570–0.698). It is of interest to note that the calibration of the model developed with UNOS data, when applied to the regional data, is almost perfect (supplementary Fig. 5 and 6).

Test cases in the external validation cohort

To exemplify the interpretability of the PRN model to predict one-year mortality, we examined random cases in the external validation cohort. The predictions for a patient are made out of the following contributions, used in each of the plots (Fig. 3). Note that the contributions to the logit are the partial responses. They are referenced to the median value of the variable, that is to say, the partial response is 0 when the variables are at their median value. $P(\text{death in year } 1|x) = \text{sigmoid function of } (f1(\text{donor_age}) + f2(\text{ischemic_time}) + \dots + f9(\text{diagnosis:NICM}) + f10((\text{recipient_age} \times \text{diagnosis:ICM})) + \text{intercept})$. In the PRN-Lasso model, there are nine univariate effects and one 2-way interaction (f10). For this patient, $P(\text{death in year } 1|x) = \text{sigmoid function of the contributing logits listed in Supplementary Table 1 plus the intercept } (-2.335) = -2.833$. $\text{Sigmoid } (-2.833) = 0.056$. For this patient, the main effect is protective, diagnosis: NICM=1 (non-ischemic disease). The only factor that increases risk is donor age 35, but this is small in value. In summary, this patient has protective factors that far outweigh the risk factors so, not surprisingly, the prediction is a good outcome, and this is also the observation. The second patient was 66 years old, had long ischemic time, infection, renal failure and was supported by ventilator pre-transplant. This is a high-risk case where 5 risk factors contribute to the mortality risk. Here the donor age was a protective together with the diagnosis, Supplementary Table 2. A nomogram has been developed that can be used to predict the one-year probability of death after heart transplantation, Fig. 5.

Discussion

The development of a new risk calculator for decision support to be used in the clinical care of patients is complex. It is not only about achieving better performance compared to previous decision support tools but also about increasing the understanding and importance of the risk factors involved. To achieve this, the new decision support must be interpretable, which can be difficult when it is based on a machine learning. The results from this study show that an interpretable machine learning model is competitive in performance compared to previously developed deep learning models tested on the same data. Although the two IHTSA models have the best overall score in terms of AUCROC, the interpretable PRN-Lasso model was within their confidence intervals, unlike the classical interpretable model, IMPACT. Furthermore, the PRN-Lasso model was better calibrated on the external validation data compared to IHTSA and IMPACT.

Considering the confidence intervals, all the machine learning models in Table 2, PRN-Lasso, EBM, IHTSA and IHTSA recalibrated, are all comparable in performance. This supports the view that the limiting factor in overall predictive power is noise, therefore any model capable of fitting the structure in the data should result in similar performance to the optimum achievable. A particular strength of the interpretable machine learning models is the clear link between the input variables and the model prediction.

Why is it important to be able to risk-stratify a patient before a transplant? One reason may be to identify the best combination of risk factors for recipients/donors to optimize the outcome. This also includes avoiding combinations that result in a poor result, to make the best use of the available organs¹³. It is also important to understand how the input factors on which the decision support is based arrive at the result. A potential disadvantage of using risk scoring models is that if the model focuses on clinical criteria and is difficult to interpret, any ethically negative effects that are built into the model might not be identified¹⁴. Furthermore, it can be difficult for a model to identify risk factors from small subgroups. Such negative effects are important to identify early in the development phase¹⁵. After the model is implemented, it can be difficult to change it. This emphasises the importance of a model being interpretable but also the importance of evaluation before clinical implementation.

There are currently about 15 different algorithms that predict survival after heart transplantation. They all have a relatively poor ability to discriminate i.e. identify an individual patient with increased risk when used on external data¹⁶. The most cited risk stratification model is the IMPACT score - an interpretable algorithm published in 2011 by Weiss et al¹¹. The scoring model is developed using logistic regression analysis where only recipient characteristics are included. Compared with the IHTSA model developed with artificial neural networks (ANNs), the ability to discriminate of the IMPACT score is inferior. On the other hand, IMPACT is easier to interpret than IHTSA. The PRN-Lasso model, which is based on both recipient and donor variables, has the same discriminatory ability as IHTSA and is interpretable as IMPACT. Tree of prediction (TOP), an interpretable model developed using regression trees, showed similar discrimination results in an internal validation such as PRN-Lasso¹⁷. However, no external validation is available for this model.

In order to assess how well decision support can work in clinical practice, it should be appropriately validated. Simple cross-validation is usually not sufficient and a separate test cohort should be used. This was recently demonstrated in a study evaluating the predictive power of popular ML and statistical algorithms. The authors use different validation techniques to assess the accuracy of the prediction of one-year mortality. The results showed that temporal validation similar to that used in this study is important due to the temporal changes in the sample of patients and donors¹⁸.

In medical fields where the patient population is not very large, clinical databases are often required, which gather information from many different centres. Thus, external validation alone is not sufficient, but local validation is equally important. This is due to differences in local processes and protocols that may change the measurement of individual covariates, as well as factors that change over time. The external validation on the Scandia data set shows how well our model trained on UNOS data applies.

This indicates that there is a high level of consistency between the measurement of the predictive variables and also for the overall mortality trend over time.

In the PRN-Lasso and IHTSA models, the year of transplantation was one of the most significant variables. It is well known that one-year survival has improved continuously over the last two decades but also the temporal changes in the sample of patients and donors^{18,19}. Interaction between different risk factors may also vary over time, as Hsieh et al. recently demonstrated using Random forest². If the model does not take this into account, performance will be impaired, as we can see for the IMPACT model in this study. IMPACT was significantly less well calibrated in the external validation cohort compared to the PRN-Lasso model. The fact that the calibration of the IHTSA models also does not work so well is because the variable is divided into time eras instead of years and the last era in the development cohort was 2010. The PRN-Lasso model treats transplant years as a continuous variable and can therefore extrapolate risk forward in time. The transplant year acts as a recalibration to take into account the gradual decrease in mortality from year to year.

The five most significant variables in the PRN model are also among the top 10 most important variables in the IHTSA model besides ischemia¹⁰. The IMPACT model lacks donor-related variables, which are two of the top five¹¹. Both donor age and ischemia are known predictors of one-year survival. Their efficacy persisted despite strong regularization at the LASSO selection in this study, especially donor age, which has proven to be the most important variable in other prediction models. In the early 1970s, Griep et al. argued that an ideal heart donor should be younger than 30 years, as confirmed in the present study²⁰. A donor age above 30 years carries an increased risk of mortality. Unlike a non-interpretable model, the PRN-Lasso model clearly shows that the risk increase for donor age is not linear but is protective at low ages. Our model also shows that the risk increase caused by ischemia comes already at 3 hours and not at 4 hours which is an old clinical rule¹⁹. The fact that ischemia is not ranked as highly in the IHTSA model is probably due to the fact that this model was not primarily optimised for one year's survival but for long-term survival time.

When quantifying the univariate and bivariate effects modelled, the input factors whose weights add together to make the final prediction may be checked against clinical expertise. The importance of the recipient age for survival varies in the literature. The biological difficulty for the algorithm to model is that the risk of total mortality increases with age, as for other diseases. The risk of transplant-related complications increases with age^{2,10}. For example, the ISHLT report shows that mortality risk for one year increased after the age of 55²¹. At the same time, the risk of rejection is reduced. The immune system is at its most active at younger ages. The findings from this study show an increased mortality rate for older recipients but a protective effect for those between 35 and 55 years of age (figure 2a). In contrast to the IHTSA model, we see that the risk again increases for the youngest recipients. However, it is interesting to note that the predicted risk is reduced for the younger recipients if the patient is diagnosed with ICM. While this bivariate effect is consistent with crude filtering of the data as noted in the results section, it may be difficult to explain biologically. ICM is traditionally a diagnosis associated with higher mortality,

especially compared to NICM. An explanation might be that the aetiology of ICM is often different for younger patients compared to older ones. For an older patient, ICM is usually associated with general atherosclerotic disease whereas for a younger patient the cause is different²². It may be in the form of a coronary artery anomaly, meaning that the problem is completely bypassed when the heart is replaced.

The results of this study have limitations associated with the retrospective analysis of a registry database, the quality of source data, and the lack of standardization associated with multi-centre studies (such as various immunosuppressive regimens and various matching criteria), as has been described previously²³. The existence of missing values is another problem that can affect the result. We used a multiple imputation technique to be able to use the entire material and avoid selection bias if one would instead choose to remove patients or variables with missing values. However, this means that the importance of variables with many missing values is more difficult to quantify.

In this study, we present an interpretable sparse algorithm with the same classification performance as today's most well-known deep learning models. In particular, the PRN model is considered to be self-explanatory because the impact of the input variables on the output is transparent. By dividing a complex multivariate algorithm into elements with lower dimensionality, the elements of this additive model are easy to read and can be interpreted by clinicians. Although interpretability is better, the low level of discrimination still persists. However, this is a known problem when modelling tabular data where the data is largely pre-processed. Classification performance measured by the AUROC is maintained for the interpretable mode compared with others on the same data. However, the clear weights associated with individual effects enables a detailed discussion to be had about the clinical plausibility of the inputs to the model, which is where the discussion of risk models for high-stakes applications needs to go. This level of transparency provides both a rigorous explanation for predictions made in respect of individual decisions and also a diagnostic route for potential failure modes present in the model. Both are essential elements in deciding when the model is and is not safe to use.

Methods

Data Source

The data set of heart transplant patients was obtained from the United Network for Organ Sharing (UNOS) database, SRTR (Scientific Registry of Transplant Recipients). UNOS (www.unos.org) is a non-profit organisation that administers the Organ Procurement and Transplantation Network (OPTN) in the United States of America. The SRTR includes data on all donors, waitlisted candidates, and transplant recipients in the United States submitted by the members of the OPTN. Human error in data entry is minimized by error checks at time of data entry and internal verification of outliers. The database contains data from October 1, 1987, onwards and includes almost 500 variables that encompass recipient, donor, and transplant information. It consists of both deceased- and living-recipient transplants.

The Ethics Committee for Clinical Research at Lund University, Sweden approved the study protocol. The data was anonymized and de-identified prior to analysis and the institutional review board waived the need for written informed consent from the participants.

Study Population

Data on heart donors and the corresponding recipient who were transplanted between January 1, 1987, and June 30, 2020, were collected from the UNOS registry (n = 76,844). Pediatric cases (recipients younger than 18 years, n = 7,577), donor age < 15 years and those with incomplete mandatory data (age, gender, duration of follow-up, and/or vital status) were excluded, Fig. 1. The development data set was divided into two temporal cohorts: transplantation done between 1997- 2013 (derivation cohort) and after or during 2014 until 2016 (test cohort). The number of variables extracted from the database was 111 in total, where IHTSA uses 43 of them and IMPACT 18. The primary endpoint was one-year mortality.

To evaluate and calculate the metrics for the models, an external validation cohort was created from SRTR using the same inclusion and exclusion criteria. Here, the endpoint was masked from the development team. The blinded validation set contained patients transplanted between 2017 and 2018 and the latest follow-up was 31st August 2020.

The Scandinavian Thoracic Transplantation Database is a registry within the Scandia Transplant organization, incorporating all thoracic transplantations performed in Norway, Denmark, Sweden, Finland, and Estonia (<http://www.scandiatransplant.org>). The registry has been in existence since 1983 and consists of more than 400 variables, and it is mandatory for all centres performing transplantations. The total number of registered heart transplants was 3,930, on December 31, 2020.

The IMPACT model

IMPACT was created with a data set of heart transplant patients between 1997 to 2008 that were collected from the UNOS database¹¹. IMPACT only utilises recipient variables. By apportioning points according to the relative importance of the variables for the one-year mortality, a risk index was created. The minimum number of scoring points a patient can have is 0 and the maximum is 50. The points are then converted to a predicted probability of one-year mortality by a formula derived from logistic regression.

The IHTSA model

The data set used in developing IHTSA was extracted from the ISHLT registry containing HT patients who were transplanted between 1994 and 2010¹⁰. IHTSA utilises both recipient and donor variables. The survival model consists of a flexible nonlinear generalisation of the standard Cox proportional hazard model. Instead of using a single prediction model, this model integrates ensembles (10 sub-models) of ANNs. Each ANN consisted of an input layer (43 inputs), one hidden layer (18 nodes) and an output layer. In addition, its prediction capability is not limited to one year. The variables hypertension and

antiarrhythmic drugs are not recorded in the UNOS database from 2007 and onward. To handle this when evaluating the IHTSA model, we imputed those two variables with random values taken from the earlier time era. In the recalibrated IHTSA model, these two variables were excluded, and the neural network retrained (calibrated), utilizing a 5-fold cross-validation of the patients between 1997 and 2008 in UNOS. The same training procedure was used as described in the original IHTSA article, but we did not carry out any new variable selection. We called this model the recalibrated IHTSA model²⁴.

Development of the Partial Response Network (PRN) and PRN-Lasso model

The PRN is probabilistic, hence it readily manages data imbalance by adjusting the threshold for class assignment, and it models the logarithm of the predicted odds-ratio using a linear combination of terms. This is similar to logistic regression, except that the additive elements are semi-parametric, hence they are non-linear functions of the original variables. The modelling methodology is explained in the Methods section and in the supplementary Methods. Formally, the PRN has the structure of a GAM, so the interpretation of the model is the model itself. Each input contributes an amount to the model prediction, given by explicit functions of only one or two variables.

An overview of the stages involved in developing the model is as follows. First, a Multi-Layer Perceptron (MLP) is fitted to the data. This pre-trained network provides the initial estimate of the posterior probability of class membership, $P(\text{Class}|x)$. However, it has the form of a black box. Second, we carry out a functional ANOVA decomposition of the $\text{logit}(P(\text{Class}|x))$ anchored at the median of the data. This results in a set of partial responses of one and two variables, together with orthogonal functions involving more than two variables which, together, add up to exactly the predictions of the logit of the MLP.

The third step is to take the univariate and bivariate partial responses only, hence truncating the ANOVA decomposition. In order to recalibrate the model, these partial responses form the inputs for a GAM, which is the logistic regression Lasso¹⁰. Lasso optimisation is particularly efficient for model selection, as it applies L1 regularization in order to collapse to zero the coefficients of the less informative inputs. The result of this step is a fully interpretable model in the form of a GAM with main effects and pairwise interactions, derived from the original black box, the MLP.

However, the presence of uninformative variables in the original MLP may reduce performance, since they will affect the response of the variables retained in the final model selection. For this reason, the three steps are repeated with a new model structure to optimise performance. In step 4 the previously derived GAM is mapped onto the structure of a Generalised Additive Neural Network (GANN). The GANN is then further trained by back-propagation and so its classification performance is optimised, resulting in what we call the PRN. This new network is shown in Fig. 6.

Finally, the same process of ANOVA decomposition followed by Lasso modeling can be applied to the output of the PRN rather than the MLP. The result is the PRN-Lasso. In all of the figures, partial responses

derived from the MLP are shown as dashed lines and those from the PRN using solid lines. The performance of the PRN and PRN-Lasso models are listed in the Results section.

Statistical analysis

Statistical analyses were performed using the Stata MP statistical package version 16.1 (2021) (StataCorp LP, College Station, TX). Data are presented as means with standard deviation, median with interquartile range (IQR) and frequency as appropriate. Unpaired Mann-Whitney U-tests were used to compare continuous variables and χ^2 tests were used to compare categorical variables among groups. The Hosmer-Lemeshow goodness-of-fit test, was used to assess predictive accuracy. The discriminatory power for one-year mortality was assessed by calculating the area under the receiver-operating curve (AUROC). To compare different areas, the non-parametric approach described by DeLong was used²⁵.

We used a multiple imputation technique consisting of probability imputation with stratification of the era. Each missing value was imputed 10 times with a random existing data point from another patient, resulting in 10 study cohorts with a variation in variables that had missing data. We combined all imputed datasets into the derivation cohort [DC], consisting of the 10 x imputed datasets (n=313150), resulting in a counterweight of random fluctuations. Note that the overall proportion missing of all data points for our PRN-Lasso model was <1% (0.69%), which is significantly less than the Impact and IHTSA models. The number of complete cases is noted in Table 1a.

Declarations

Author contributions statement

All authors contributed to the study design and data interpretation. PL and JN, undertook the analysis and validation of the data. PL wrote the initial code for the PRN model and JN the IHTSA model. PL, MJ, SOM, IO, DM, and JN performed the computer programming. PL and JN drafted the initial report, and all authors revised the manuscript for intellectual content and have seen and approved the final version. Data was provided from the UNOS registry by staff at the US, United Network for Organ Sharing, and compiled by JN.

Acknowledgments

The study was supported by the Swedish Research Council (2019-00487), Vinnova (2017-04689), Swedish Heart-Lung Foundation (20190623), a government grant for clinical research, region Skane research funds, donation funds from Skane University Hospital, and the Anna-Lisa and Sven Eric Lundgrens Foundation. The supporting sources had no involvement in the study.

Competing interest

The authors declare no competing financial or non-financial interests.

Data availability

The data that support the findings of this study are available from the SRTR, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Code availability

The PRN Lasso model was developed in MATLAB 2020b (MathWorks, Natick, MA). The code files are available on request from the authors.

Material and Correspondence

Correspondence and requests for material should be addressed to PL or JN.

Supplementary information

Supplementary Methods.

Supplementary Figures.

Supplementary Tables.

References

- 1 Cogswell, R. et al. An early investigation of outcomes with the new 2018 donor heart allocation system in the United States. *J. Heart Lung Transplant.* 39, 1-4, doi:10.1016/j.healun.2019.11.002 (2020).
- 2 Hsieh, E. M. et al. Heart Transplantation: An In-Depth Survival Analysis. *JACC Heart Fail* 8, 557-568, doi:10.1016/j.jchf.2020.03.014 (2020).
- 3 Christodoulou, E. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* 110, 12-22, doi:10.1016/j.jclinepi.2019.02.004 (2019).
- 4 Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 206-215, doi:10.1038/s42256-019-0048-x (2019).
- 5 Caruana, R. et al. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1721–1730 (Association for Computing Machinery).
- 6 Lundberg, S. M. & Lee, S.-I. in *Advances in Neural Information Processing Systems 30* (eds I. Guyon et al.) 4765–4774 (Curran Associates, Inc., 2017).

- 7 Alvarez Melis, D. & Jaakkola, T. in 32nd Conf. Neural Inf. Process. Syst Vol. 31 (eds S. Bengio et al.) (Curran Associates, Inc., Montréal, Canada, 2018).
- 8 White, H. Artificial neural networks. (Blackwell Cambridge, Mass., 1992).
- 9 Hooker, G. Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics* 16, 709-732, doi:10.1198/106186007x237892 (2007).
- 10 Nilsson, J. et al. The International Heart Transplant Survival Algorithm (IHTSA): a new model to improve organ sharing and survival. *PLoS One* 10, e0118644, doi:10.1371/journal.pone.0118644 (2015).
- 11 Weiss, E. S. et al. Creation of a quantitative recipient risk index for mortality prediction after cardiac transplantation (IMPACT). *Ann. Thorac. Surg.* 92, 914-921; discussion 921-912, doi:10.1016/j.athoracsur.2011.04.030 (2011).
- 12 Lou, Y., Caruana, R. & Gehrke, J. in *KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* 150-158.
- 13 Dorent, R. et al. New French heart allocation system: Comparison with Eurotransplant and US allocation systems. *Am. J. Transplant.* 20, 1236-1243, doi:10.1111/ajt.15816 (2020).
- 14 Babic, B., Cohen, G., Evgeniou, T., Gerke, S. & Trichakis, N. *Can AI Fairly Decide Who Gets an Organ Transplant?*, 2020).
- 15 Martin, K. Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics* 160, 835-850, doi:10.1007/s10551-018-3921-3 (2018).
- 16 Aleksova, N. et al. Risk prediction models for survival after heart transplantation: A systematic review. *Am. J. Transplant.* 20, 1137-1151, doi:10.1111/ajt.15708 (2020).
- 17 Yoon, J. et al. Personalized survival predictions via Trees of Predictors: An application to cardiac transplantation. *PLoS One* 13, e0194985, doi:10.1371/journal.pone.0194985 (2018).
- 18 Miller, R. J. H. et al. Temporal Shift and Predictive Performance of Machine Learning for Heart Transplant Outcomes. *The Journal of Heart and Lung Transplantation*, doi:10.1016/j.healun.2022.03.019 (2022).
- 19 Lund, L. H. et al. The Registry of the International Society for Heart and Lung Transplantation: Thirty-fourth Adult Heart Transplantation Report-2017; Focus Theme: Allograft ischemic time. *J. Heart Lung Transplant.* 36, 1037-1046, doi:10.1016/j.healun.2017.07.019 (2017).
- 20 Griepp, R. B., Stinson, E. B., Clark, D. A., Dong, E., Jr. & Shumway, N. E. The cardiac donor. *Surg. Gynecol. Obstet.* 133, 792-798 (1971).

- 21 Stehlik, J. et al. The Registry of the International Society for Heart and Lung Transplantation: 29th official adult heart transplant report–2012. *J. Heart Lung Transplant.* 31, 1052-1064, doi:10.1016/j.healun.2012.08.002 (2012).
- 22 Lipshultz, S. E. et al. Cardiomyopathy in Children: Classification and Diagnosis: A Scientific Statement From the American Heart Association. *Circulation* 140, e9-e68, doi:10.1161/CIR.0000000000000682 (2019).
- 23 Dickinson, D. M. et al. Transplant data: sources, collection, and caveats. *Am. J. Transplant.* 4 Suppl 9, 13-26, doi:10.1111/j.1600-6135.2004.00395.x (2004).
- 24 Medved, D. et al. Improving prediction of heart transplantation outcome using deep learning techniques. *Sci. Rep.* 8, 3613, doi:10.1038/s41598-018-21417-7 (2018).
- 25 DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837-845 (1988).

Tables

Table 1a. Recipient baseline characteristics for the study cohorts.

	N	Ext. validation 2017-2018 N=4,750	Test 2014-2016 N=6,120	Train/validation 1997-2013 N=31,315	Test Statistic
Age (yrs) ^{*,#,†}	42185	47 57 64 54 ± 13	47 56 63 54 ± 13	46 55 61 52 ± 12	F_2 42182=91, P<0.001 ²
Female gender ^{*,†}	42185	26% ¹²⁴⁴ / ₄₇₅₀	26% ¹⁵⁶² / ₆₁₂₀	24% ⁷⁴²³ / ₃₁₃₁₅	$\chi^2_2=20$, P<0.001 ¹
Height (cm) *	42076	167.6 175.0 180.3 173.8 ± 10.1	167.6 175.0 180.3 173.8 ± 10.0	167.6 175.2 180.3 173.9 ± 9.7	F_2 42073=1.7, P=0.19 ²
Weight (kg) *	42091	71 84 97 84 ± 18	70 83 95 83 ± 18	69 80 92 81 ± 17	F_2 42088=109, P<0.001 ²
Race [#]	42185				$\chi^2_8=377$, P<0.001 ¹
Asian		3.6% ¹⁶⁹ / ₄₇₅₀	3.6% ²²² / ₆₁₂₀	2.3% ⁷⁰⁹ / ₃₁₃₁₅	
Black		22.8% ¹⁰⁸² / ₄₇₅₀	22.5% ¹³⁷⁴ / ₆₁₂₀	16.5% ⁵¹⁶⁸ / ₃₁₃₁₅	
Hispanic		8.8% ⁴²⁰ / ₄₇₅₀	8.7% ⁵³² / ₆₁₂₀	6.8% ²¹¹⁵ / ₃₁₃₁₅	
Other		1.2% ⁵⁷ / ₄₇₅₀	1.3% ⁷⁸ / ₆₁₂₀	1.0% ³¹⁰ / ₃₁₃₁₅	
White		63.6% ³⁰²² / ₄₇₅₀	64.0% ³⁹¹⁴ / ₆₁₂₀	73.5% ²³⁰¹³ / ₃₁₃₁₅	
Diagnosis ^{*,#,†}	42185				$\chi^2_{10}=531$, P<0.001 ¹
Congenital		3.1% ¹⁴⁵ / ₄₇₅₀	2.8% ¹⁷¹ / ₆₁₂₀	2.4% ⁷³⁹ / ₃₁₃₁₅	
Graft failure		2.7% ¹²⁶ / ₄₇₅₀	2.8%	3.1% ⁹⁷⁵ / ₃₁₃₁₅	

			171 / 6120		
ICM		30.7% ^{1460 / 4750}	34.2% 2094 / 6120	43.8% 13711 / 31315	
NICM		60.1% ^{2856 / 4750}	57.0% 3488 / 6120	46.7% 14616 / 31315	
Other		2.3% ^{111 / 4750}	1.9% 118 / 6120	2.0% ^{617 / 31315}	
Valve		1.1% ^{52 / 4750}	1.3% 78 / 6120	2.1% ^{657 / 31315}	
Diabetes mellitus [*]	41700	29% ^{1382 / 4748}	29% 1780 / 6114	24% ^{7313 / 30838}	$\chi^2_2=124$, P<0.001 ¹
Infection within 2 weeks ^{*,#,†}	42181	9.1% ^{430 / 4750}	10.8% 662 / 6120	10.7% ^{3343 / 31311}	$\chi^2_2=12$, P=0.002 ¹
Dialysis [#]	41359	4.7% ^{222 / 4748}	4.2% 256 / 6117	3.3% ^{1020 / 30494}	$\chi^2_2=27$, P<0.001 ¹
Previous blood transfusion [*]	39870	21% ^{1009 / 4732}	27% 1598 / 5993	22% ^{6387 / 29145}	$\chi^2_2=69$, P<0.001 ¹
Previous organ transpl. [*]	42185	3.0% ^{142 / 4750}	3.0% 182 / 6120	3.4% ^{1078 / 31315}	$\chi^2_2=5.4$, P=0.069 ¹
Prior cardiac surgery (non-OHT) [*]	29165	50% ^{2380 / 4750}	56% 3416 / 6120	48% ^{8727 / 18295}	$\chi^2_2=121$, P<0.001 ¹
Medical condition at OHT [*]	42185				$\chi^2_4=78$, P<0.001 ¹
Home		53% ^{2516 / 4750}	54% 3296 / 6120	49% ^{15305 / 31315}	
Hospital		16% ^{759 / 4750}	17% 1050 / 6120	18% ^{5478 / 31315}	
ICU		31% ^{1475 / 4750}	29% 1774 / 6120	34% ^{10532 / 31315}	
Ventilator at OHT ^{*,#,†}	42185	0.9% ^{42 / 4750}	0.9% 58 / 6120	2.6% ^{807 / 31315}	$\chi^2_2=105$, P<0.001 ¹

ECMO*	42185	1.3% ^{64/4750}	0.8% ^{50/6120}	0.5% ^{150/31315}	$\chi^2_2=54$, P<0.001 ¹
IABP* ^{*,#}	42185	8.7% ^{414/4750}	7.0% ^{428/6120}	5.3% ^{1667/31315}	$\chi^2_2=99$, P<0.001 ¹
VAD at OHT* ^{*,#}	36831				$\chi^2_{10}=3317$, P<0.001 ¹
LVAD		45.6% ^{2168/4750}	45.5% ^{2787/6120}	19.4% ^{5032/25961}	
LVAD + RVAD		1.5% ^{71/4750}	2.0% ^{121/6120}	2.4% ^{625/25961}	
No device		51.9% ^{2463/4750}	50.7% ^{3102/6120}	68.5% ^{17772/25961}	
RVAD		0.2% ^{11/4750}	0.2% ^{10/6120}	0.2% ^{54/25961}	
TAH		0.8% ^{37/4750}	1.6% ^{100/6120}	0.8% ^{196/25961}	
Unknown device		0.0% ^{0/4750}	0.0% ^{0/6120}	8.8% ^{2282/25961}	
PVR (wood units) *	34075	1.4 2.0 2.8 2.3 ± 1.4	1.4 2.0 3.0 2.4 ± 1.5	1.4 2.1 3.1 2.5 ± 1.7	F_2 34072=23, P<0.001 ²
SPP (mmHg) *	39300	29 37 47 39 ± 13	30 38 49 40 ± 14	32 41 52 42 ± 14	F_2 39297=197, P<0.001 ²
Creatinine ($\mu\text{mol/l}$) ^{*,#}	41367	85 106 133 120 ± 74	85 106 133 119 ± 69	88 106 133 120 ± 68	F_2 41364=4.8, P=0.008 ²
Serum bilirubin ($\mu\text{mol/l}$) ^{*,#}	40640	6.8 12.0 17.1 16.6 ± 31.8	8.6 12.0 18.8 17.0 ± 33.1	8.6 13.7 22.2 20.9 ± 41.6	F_2 40637=313, P<0.001 ²
PRA > 10%*	36978	24% ^{916/3773}	22% ^{1149/5287}	15% ^{4188/27918}	$\chi^2_2=306$, P<0.001 ¹
HLA-DR mismatch*	27383				

$\chi^2_4=18,$
 $P=0.001^1$

0	4.6% ^{156/3410}	5.2% ^{226/4312}	4.2% ^{830/19661}
1	37.6% ^{1283/3410}	38.1% ^{1642/4312}	40.1% ^{7883/19661}
2	57.8% ^{1971/3410}	56.7% ^{2444/4312}	55.7% ^{10948/19661}
ABO blood type*	42185		$\chi^2_6=36,$ $P<0.001^1$
A	40.1% ^{1903/4750}	38.9% ^{2382/6120}	42.2% ^{13222/31315}
AB	5.2% ^{248/4750}	6.0% ^{369/6120}	5.3% ^{1661/31315}
B	15.4% ^{730/4750}	15.4% ^{943/6120}	13.9% ^{4352/31315}
O	39.3% ^{1869/4750}	39.6% ^{2426/6120}	38.6% ^{12080/31315}

a b c represents the lower quartile a, the median b, and the upper quartile c for continuous variables. $x \pm s$ represents $X \pm 1$ SD. N is the number of non-missing values. Variables included in * IHTSA, # IMPACT, and † PRN model. SPP, Systolic pulmonary pressure; PRA, panel reactive antibody level, HLA, human leucocyte antigen; OHT, orthotopic heart transplantation; VAD, ventricular assist device; LVAD, left ventricular assist device; RVAD, Right ventricular assist device; TAH, total artificial heart; ECMO, extracorporeal membrane oxygenation. ICU, Intensive care unit; NICM, non-ischemic cardiomyopathy; ICM, ischemic cardiomyopathy; PVR, pulmonary vascular resistance; IABP, intra aortic balloon pump
 Tests used: 1Pearson test; 2Kruskal-Wallis test.

Table 1b. Donor baseline characteristics for the study cohorts.

	N	Ext. validation 2017-2018 N=4,750	Test 2014- 2016 N=6,120	Train/val. 1987-2013 N=31,315	Test Statistic
Donor age (yrs) ^{*,†}	42185	24 31 40 33 ± 11	23 31 41 32 ± 11	21 30 42 32 ± 12	$F_{2,42182}=18,$ $P<0.001^2$
Donor female gender [*]	42185	30% 1443/ 4750	30% 1828/ 6120	29% 9214/ 31315	$\chi^2_2=2.1,$ $P=0.36^1$
Donor height (cm) [*]	42144	168.0 175.0 180.3 174.0 ± 9.6	167.6 175.0 180.3 173.8 ± 9.6	167.6 175.3 182.9 174.6 ± 9.5	$F_{2,42141}=19,$ $P<0.001^2$
Donor weight (kg) [*]	42179	70 81 95 84 ± 19	70 80 93 83 ± 20	68 79 90 81 ± 18	$F_{2,42176}=69,$ $P<0.001^2$
Donor ABO blood type [*]	42184				$\chi^2_6=7.7,$ $P=0.26^1$
A		36.7% 1745/ 4750	35.3% 2163/ 6120	36.4% 11413/ 31314	
AB		2.1% 99/ 4750	2.5% 154/ 6120	2.3% 706/ 31314	
B		10.9% 520/ 4750	11.3% 690/ 6120	10.5% 3295/ 31314	
O		50.2% 2386/ 4750	50.9% 3113/ 6120	50.8% 15900/ 31314	
Donor cause of death [*]	42149				$\chi^2_8=2669,$ $P<0.001^1$
Anoxia		38.2% 1813/ 4748	29.8% 1823/ 6118	12.3% 3850/ 31283	
Cerebrovascular/Stroke		14.8% 703/ 4748	19.1% 1166/ 6118	25.6% 8000/ 31283	
CNS tumor		0.5% 23/ 4748	0.3% 21/ 6118	1.0% 313/ 31283	

Head trauma		44.6% ^{2119/4748}	48.6% ^{2971/6118}	59.3% ^{18544/31283}	
Other		1.9% ^{90/4748}	2.2% ^{137/6118}	1.8% ^{576/31283}	
Ischemic time (min) ^{*,†}	40394	138 184 223 183 ± 64	143 186 227 187 ± 62	146 188 229 190 ± 63	$F_{2,40391}=23$, P<0.001 ²
Recipient-donor height ratio [*]	42037	0.963 1.000 1.036 1.000 ± 0.056	0.967 1.000 1.040 1.002 ± 0.057	0.958 1.000 1.036 0.998 ± 0.061	$F_{2,42034}=16$, P<0.001 ²
Recipient-donor weight ratio [*]	42086	0.87 1.03 1.18 1.03 ± 0.23	0.88 1.02 1.17 1.03 ± 0.22	0.88 1.02 1.16 1.02 ± 0.21	$F_{2,42083}=2.6$, P=0.071 ²

a b c represents the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables. $x \pm s$ represents $X \pm 1$ SD. *N* is the number of non-missing values. Variables included in ^{*}IHTSA, [#]IMPACT, and [†]PRN model. CNS, Central nervous system tumours. Tests used: ¹Pearson test; ²Kruskal-Wallis test.

Table 2. AUROC - External validation cohort

	N	ROC area	[95% conf.]	P
PRN-Lasso	4750	0.628	0.602-0.654	
EBM ¹³	4750	0.634	0.607-0.660	0.173
IHTSA ¹⁰	4750	0.635	0.609-0.662	0.488
IHTSA recalibrated ²⁴	4750	0.643	0.617-0.669	0.197
IMPACT ¹¹	4750	0.602	0.575-0.628	0.094

PRN, Partial Response Network; EBM, Explainable Boosting Machines; IHTSA, International Heart Transplant Survival Algorithm; IMPACT, Index for Mortality Prediction After Cardiac Transplantation

Figures

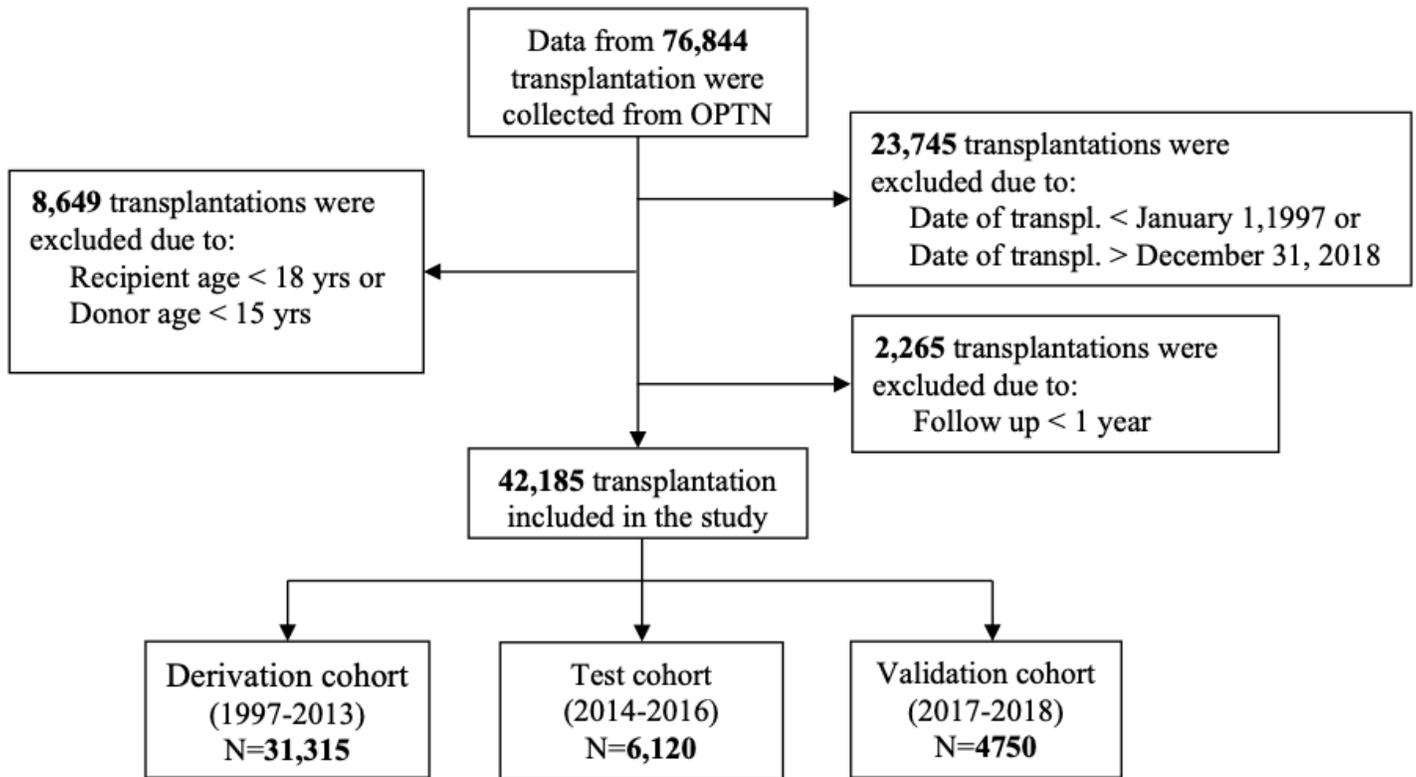


Figure 1

Flow chart.

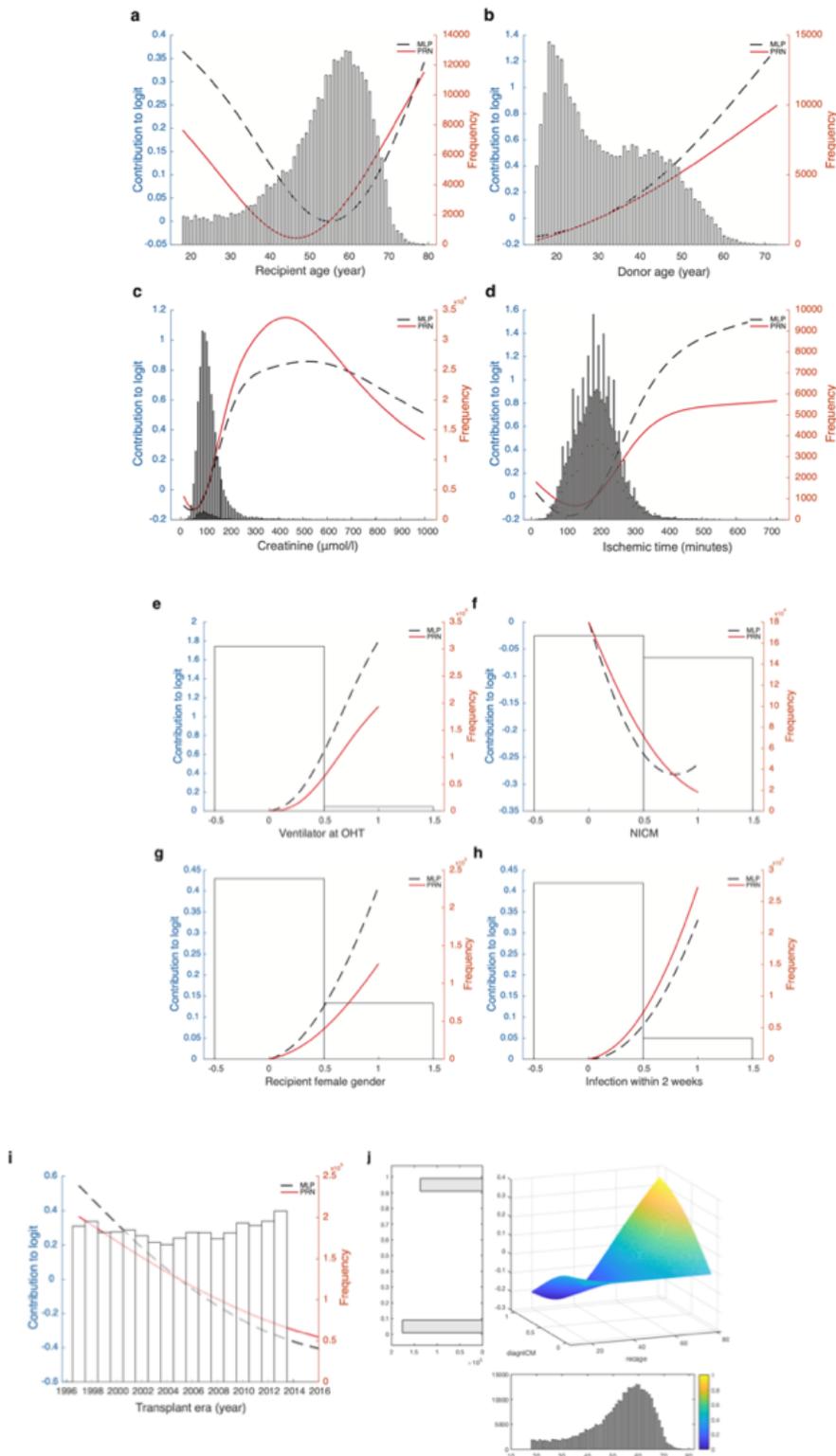


Figure 2

Partial response representing the additive contribution to the logarithm of the odds of death in year one.

The partial response representing the additive contribution of: panel a) recipient age (≥ 18 yrs), b) donor age (≥ 15 yrs), c) creatinine, d) ischemic time, e) ventilator at OHT, f) diagnosis: non-ischemic cardiomyopathy, g) recipient female gender, h) infection within 2 weeks, and i) transplantation era to the

logarithm of the odds of death in year one i.e. the logit. The black line shows the initial estimate of the partial response directly from the ANOVA decomposition and the red curve is the same function recalibrated in the second step of back error propagation. Overlaid in the figure is the histogram of the input variable. Panel j shows the partial response from the bivariate effect between recipient age and diagnosis of ICM. MLP denotes the partial responses obtained from the original neural network; PRN denotes the partial responses obtained after subsequent re-training of the SENN in fig. 6.

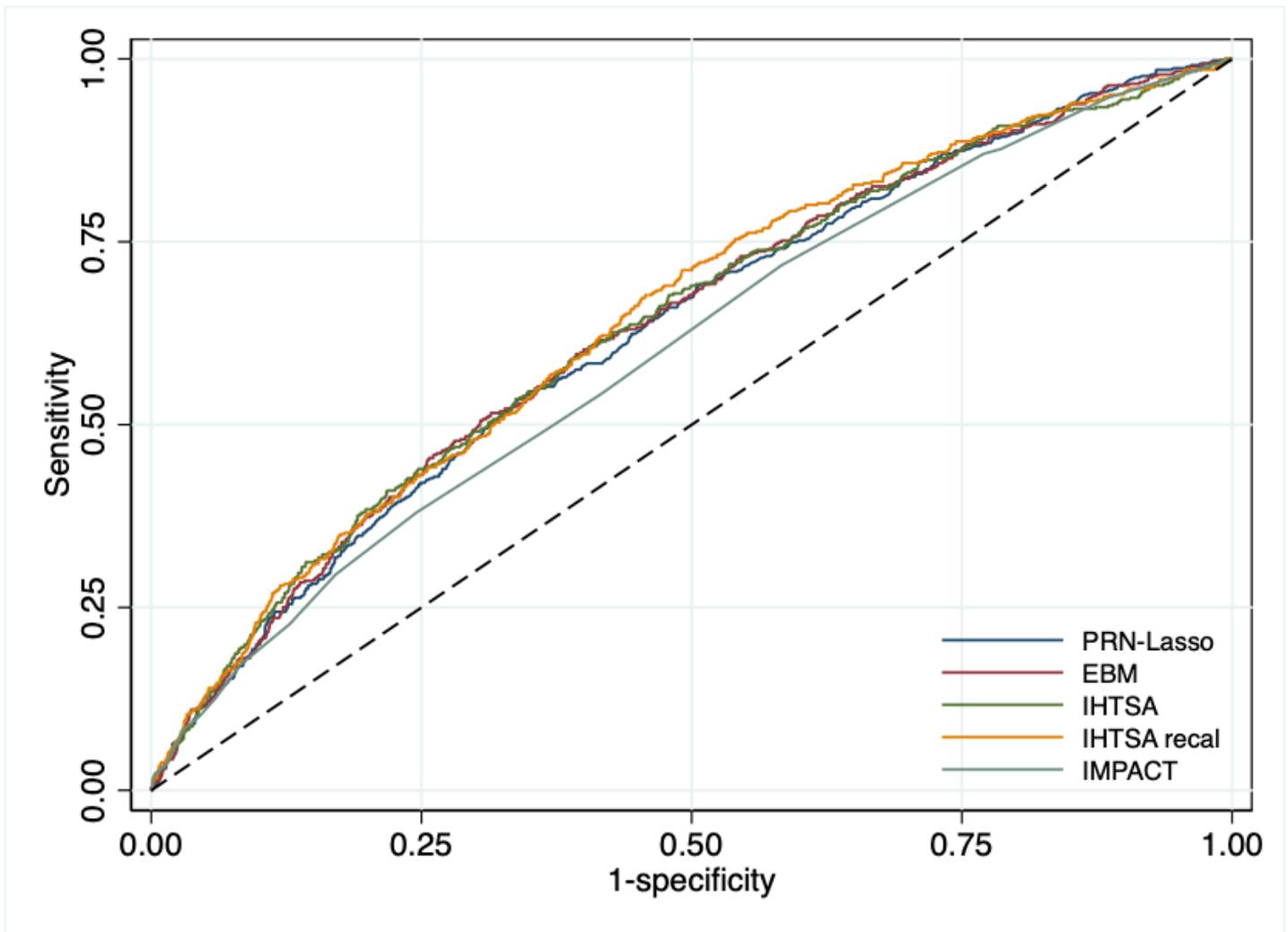


Figure 3

ROC curves of the different models

The ROC curves show the sensitivity and 1-specificity of prediction of one-year mortality for the PRN with 10 responses (navy blue solid line), EBM (maroon solid line), IHTSA (forest green solid line), IHTSA recalibrated (dark orange solid line) and IMPACT (teal green solid line) risk algorithm. The black dash line

represents the absence of discrimination. Heart transplanted patients in the blinded validation cohort 2017-2018 (n=4570) are included in the analysis.

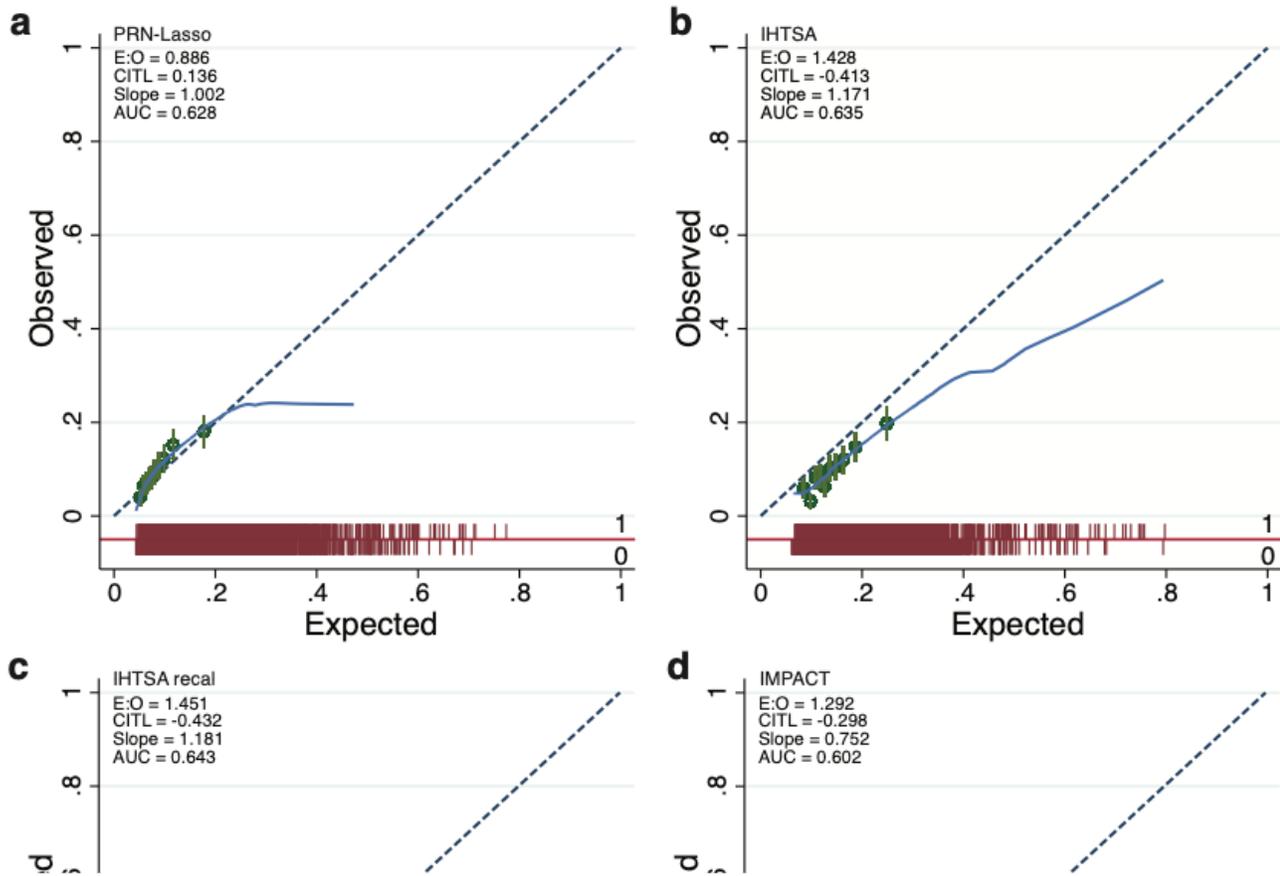


Figure 4

Calibration plot of PRN

The calibration plot, panel a-d, shows observed against expected probabilities for assessment of the four prediction models with the overall external validation cohort according to the TRIPOD guidelines. The red spike plot shows the distribution of events and non-events. Green circles the expected probabilities in

groups with 95% CI. The green line demonstrates the lowest smoother. The model performance statistics are summarised with EO: exp/obs ratio; CITL: calibration-in-the-large; and AUC: area under the curve.

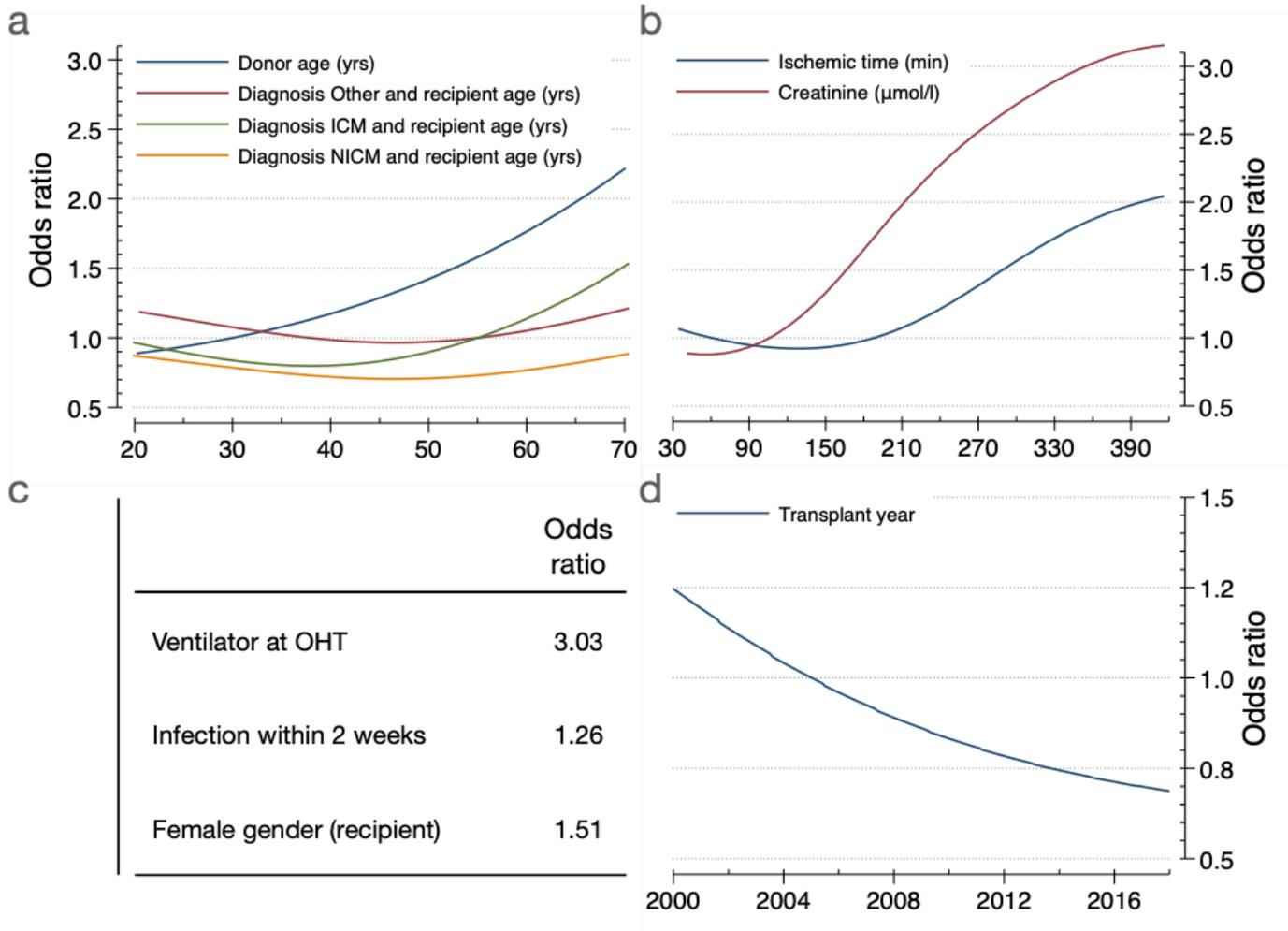


Figure 5

Nomogram predicting one-year mortality after heart transplantation

The partial response representing the additive contribution: panel a) donor age (red solid line), recipient age and diagnosis Other (red solid line), recipient age and diagnosis ICM (green solid line), recipient age and diagnosis NICM (orange solid line); b) creatinine (red solid line), and ischemic time (blue solid line), c) ventilator at OHT, recipient female gender, infection within 2 weeks; d) transplantation year to the odds of death in year one. To calculate the logarithm of the total odds for an individual patient, the logarithm of the odds ratios (contributing logits) are summed for all the contributing factors adding also an overall intercept of -2.3354.

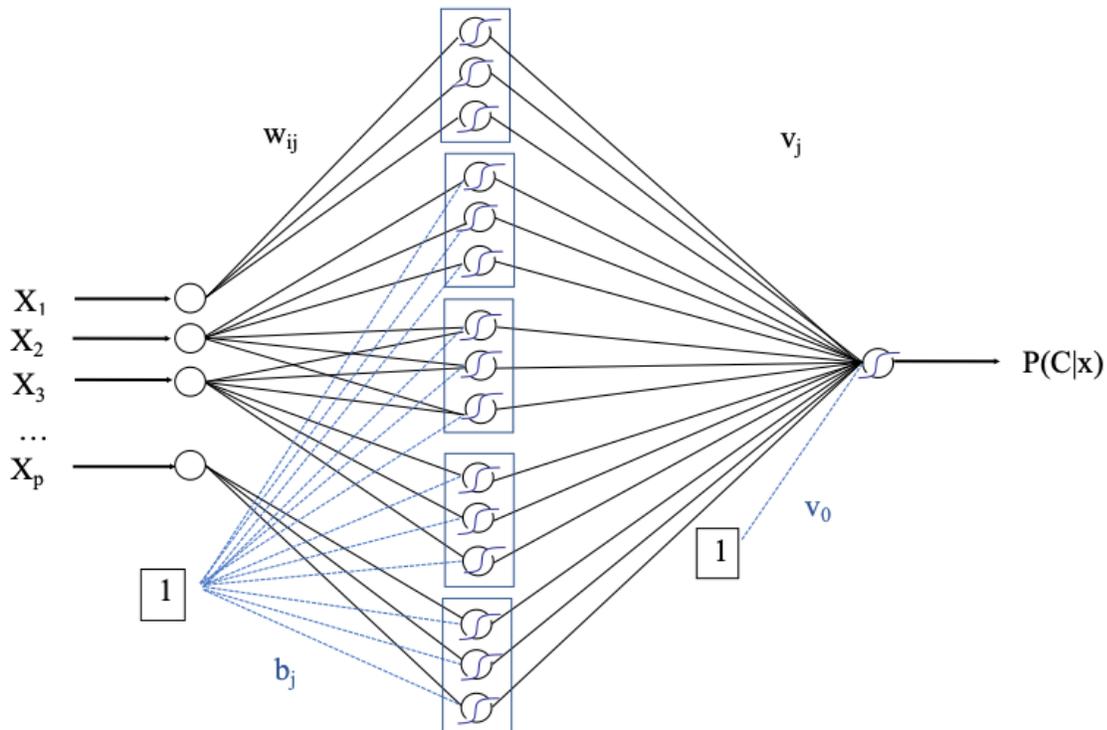
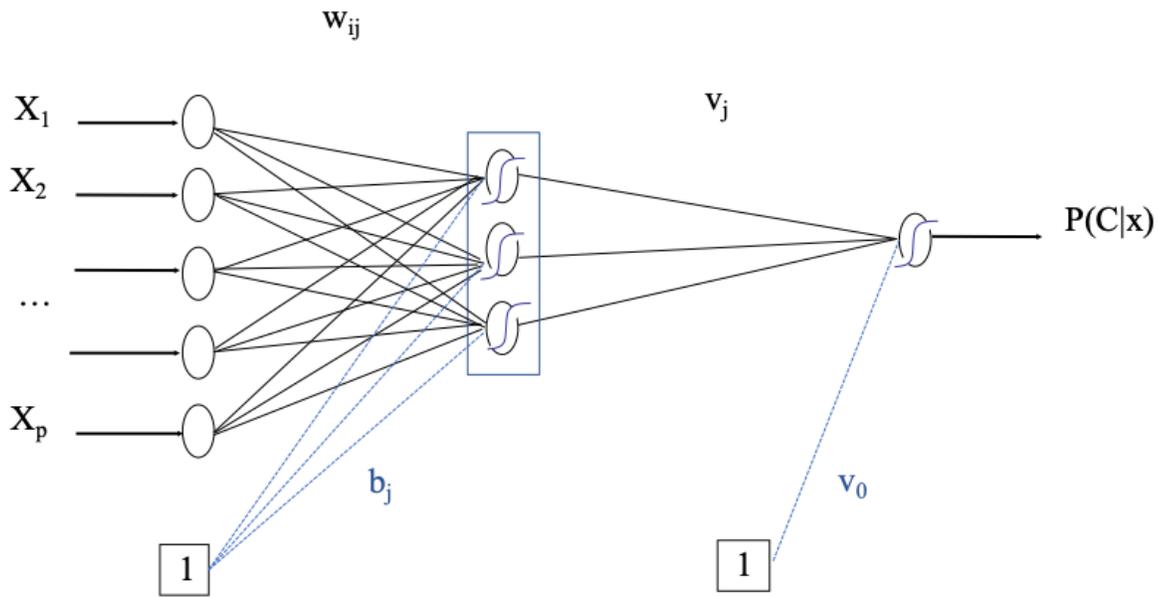


Figure 6

Structure of the original fully connected MLP (top) and the Partial Response Network (PRN) (bottom) which has the structure of a Self-Explaining Neural Network

The structure of the Partial Response Network (PRN) is comprised of modular replicas of the relevant weights from the original MLP for each univariate or bivariate response retained by the Lasso, adjusted

by eqs. (6-11) to initialise the PRN with the exact functional response of the Lasso. Further training smooths out the responses and improves predictive performance.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementInformation.pdf](#)