Predicting the Outcome for Patients in a Heart Transplantation Queue using Deep Learning

Dennis Medved[1], Pierre Nugues[1], and Johan Nilsson[2]

Abstract— **Heart transplantations have made it possible to extend the median survival time to 12 years for patients with end-stage heart diseases. This operation is unfortunately limited by the availability of donor organs and patients have to wait on average about 200 days in a waiting list before being operated. This waiting time varies considerably across the patients. In this paper, we studied the outcome for patients entering a transplantation waiting list using deep learning techniques. We implemented a model in the form of two-layer neural networks and we predicted the outcome as still waiting, transplanted or dead in the waiting list, at three different time points: 180 days, 365 days, and 730 days. As data source, we used the United Network for Organ Sharing (UNOS) registry, where we extracted adult patients (>17 years) from January 2000 to December 2011. We trained our model using the Keras framework, and we report F1 macro scores of respectively 0.674, 0.680, and 0.680 compared to a baseline of 0.271. We also applied a backward elimination procedure, using our neural network, to extract the 10 most significant parameters predicting the patient status for the three different time points.**

I. Introduction

Heart transplantations have made it possible to extend the median survival time to 12 years for patients with end-stage heart diseases. Unfortunately, the need for donated hearts greatly exceeds supply and many candidates die awaiting transplantation. Estimating the probability of dying in the waiting list for a specific time period, could support the decision of surgeons on the priority of a transplantation. In addition, knowing the probability for a patient to be transplanted within a certain time frame would help plan operation resources and inform the patient.

In this study, we have used neural network models to predict the outcome for patients entering a heart transplantation waiting list. We carried out the prediction at three different time points: 180 days, 365 days and 730 days. We categorized the patient status with three possible outcomes: still waiting, transplanted, or dead in the waiting list.

II. Previous Work

A few studies investigated waiting times of allografts. They include heart transplants [10, 4], liver [1], and kidney [3], that all revealed increased waiting times for group O recipients. Other studies proposed models to predict the outcome in heart failures and outlined lists of predictors. [9] is a review of 64 such models, where the possible outcomes were death, hospitalization, and death or hospitalization, depending on the model. The authors could distill a list of 10 consistently used predictors: age, renal function, blood pressure, blood sodium level, etc. Other papers provide models to estimate the survival time after a heart transplantation such as [16, 6], while [5] describe a procedure to extract features predicting the one, five, and ten year survival of patients.

III. Materials and Methods

A. Data Source

UNOS administers the only Organ Procurement and Transplantation Network in the United States of America [7], and is a non-profit organization. The patient data that we used was obtained from the UNOS database. The database contains data from October 1, 1987 and onwards. In the database, there is information that encompass recipient, donor and transplant data. It includes almost 500 variables reflecting different attributes of the patients.

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

B. Study population

We included adult (> 17 years) heart transplantation (HT) patients from January 2000 to December 2011, that either died in the queue, got transplanted, or were still waiting in the queue. We did not include patients, who were removed from the list for other reasons, such as being too sick to be operated. We excluded these patients because they could potentially confuse the model. We assumed that the features predicting the death of a patient would probably be correlated with a removal from the queue.

We used this data set to create three different temporal cohorts, where we recorded the patients' outcome after 180, 365, and 730 days. Table I shows the distribution of outcomes in the different time periods.

The total number of patients included in our data set was of 27,444. We randomly divided the data in train/validation/test in sets of 70%/15%/15% which translates to 19210/4117/4117 patients, respectively.

As features, we included 87 variables describing the patients in the queue that were available at the time of listing such as: age, sex, weight, and blood group.

*This research was supported by Heart Lung Fondation, The Swedish Research Council, and the eSSENCE program.

[1]Department of Computer Science, Lund University, Lund, Sweden {dennis.medved, pierre.nugues}@cs.lth.se

[2]Department of Clinical Sciences Lund, Cardiothoracic Surgery, Lund University and Skåne University Hospital, Lund, Sweden johan.nilsson@med.lu.se

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

TABLE I

THE THREE TEMPORAL COHORTS AND OUTCOME DISTRIBUTION

Days	Dead (%)	Transplanted (%)	Queueing (%)
180	9.7	57.0	33.4
365	11.6	69.1	19.3
730	13.4	77.3	9.3

C. Imputation of Missing Data

As with all large registries, there is missing patient data. No patient has a complete information record and excluding the patients with missing data fields from the cohorts would have reduced the data set to almost nothing. To mitigate this, we chose to impute the missing data, where we applied a probabilistic approach. For each variable, we replaced the missing values with a random value from a discrete uniform distribution of the non-missing values in this variable, following the method in [11].

D. Evaluation

To evaluate the models, we used the F1 score, which is the harmonic mean of precision and recall [8]. These metrics were created for binary classes and to generalize them to more than two classes, we averaged the results using micro and macro averages.

The micro average method consists of summing up the individual true positives, false positives, and false negatives of the system for the different classes and then calculating the average. The macro average takes the average of the precision and recall of the system on the different classes. When the examples are unevenly distributed across the classes, the macro average method is less biased toward the largest class [15].

We also computed a confusion matrix, where each column of the matrix represents the instances of a predicted class, while each row represents the actual class. The diagonal then represents the correctly classified outcomes. Confusion matrices make it easier to visualize the classification errors that a model produces [13].

E. Implementation Details

We used the Keras framework to train the model [2]. It utilizes Python as a programming interface and enables the user to easily create and configure artificial neural networks (ANN) of different architectures. It serves as a high level abstraction, that utilizes Theano as the back-end [14].

We created a network with two hidden layers and 128 nodes in each layer. The hidden layers used the rectified linear unit as activation function and the final output layer used a softmax activation. We selected categorical cross entropy as the loss function and adamax as the optimizer with 30 epochs.

Dropout is a regularization technique for reducing overfitting in neural networks [12]. The idea behind dropout is to randomly drop units, together with their connections, from the neural network during training. The dropout rate controls the probability of a neuron being removed. We chose to use a dropout rate of 0.5.

F. Feature Significance

We wanted to know which features contributed the most to the result of the classification. We utilized backward elimination to find these features.

Backward elimination starts with all the features and removes them one by one from the set. The resulting feature set is then used to produce the classification probabilities. We calculate the F1 macro metric for each of the new feature sets and remove the feature that produced the best score when excluded. We repeat this process until the desired amount of features remain.

IV. RESULTS

We optimized the hyperparameters on the validation set. Using these parameters, Table II shows the precision and recall values we obtained on the test set, while Table III shows the F1 values for 180, 365, and 730 days, respectively. We included a baseline model in the table that always classifies the most frequent class, in this case: the patient was transplanted. The best macro averaged F1 was achieved for 365 days: 0.680. Figure 1 shows the precision-recall curve for this time period.

TABLE II

THE PRECISION AND RECALL VALUES FOR 180, 365, AND 730 DAYS OBTAINED ON THE TEST SET

Days	Class	Precision	Recall	F1
180	Dead	0.680	0.644	0.664
	Transplanted	0.764	0.887	0.820
	Queuing	0.654	0.485	0.557
365	Dead	0.782	0.684	0.705
	Transplanted	0.842	0.967	0.900
	Queuing	0.605	0.314	0.413
730	Dead	0.770	0.747	0.759
	Transplanted	0.918	0.992	0.954
	Queuing	0.606	0.226	0.329
Baseline 180	Dead	0.000	0.000	0.000
	Transplanted	0.567	1.000	0.724
	Queuing	0.000	0.000	0.000
Baseline 365	Dead	0.000	0.000	0.000
	Transplanted	0.77	1.000	0.869
	Queuing	0.000	0.000	0.000
Baseline 730	Dead	0.000	0.000	0.000
	Transplanted	0.685	1.000	0.813
	Queuing	0.000	0.000	0.000

Using the neural network and backward elimination, we extracted the ten most important features, shown in Table IV. The features are ranked within the sets, according to their removal order. We evaluated these feature sets and Table V shows the results. Figure 2 shows the confusion matrix for the 365 days time period that reveals that the most misclassified outcome is queueing as transplanted.

We wanted to look at the distributions of outcomes depending on the patient having blood group O or not, mostly because previous studies had shown that it was a predictor. In addition, it was also implicitly included in the ten most predictive features for 365 days. Table VI shows that there is a 17% absolute difference between the number of transplanted.

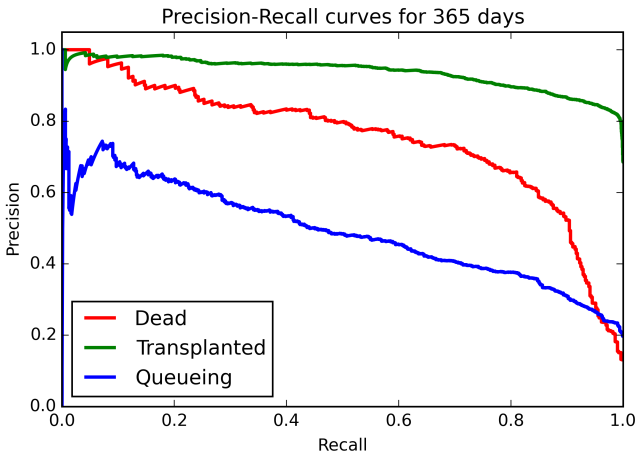

Fig. 1. Precision-recall curves for 365 days

TABLE III
THE F1 VALUES FOR 180, 365, AND 730 DAYS OBTAINED ON THE TEST SET

Days	F1 (micro)	F1 (macro)
180	0.750	0.675
365	0.760	0.680
730	0.888	0.680
Baseline 180	0.567	0.241
Baseline 365	0.685	0.271
Baseline 730	0.769	0.290

V. DISCUSSION

The distribution of patient outcomes within the cohorts is quite imbalanced, where transplanted is the outcome for 57-77% of the patients, during the chosen time periods. We tried a simple baseline, where we classified all the patient outcomes as the most frequent, see Table III for the results. It produced quite good micro averaged values, mostly because these metrics are biased towards the largest class, but comparatively bad macro values.

The largest misclassification error in Figure 2 corresponds to queueing as transplanted. This is probably because it is hard to differentiate between the patients that were transplanted at a certain time point versus those that are still waiting in the queue, based on the available features.

We carried out a backward elimination using our neural network and the ten most contributing features is shown in Table IV. This results in a decrease of only about 2% (absolute difference) from the F1 macro score with all the features, see Table V. This means that most of the predictive power from the ANN comes from a few features. Neural networks do a kind of feature selection naturally as part of the model, weighing up more predictive features and weighing down the less predictive. Because of this, feature search for neural networks is usually not needed. But considering it is hard to interpret the matrices produced by the ANN model directly, we carried out a backward elimination to approximate the features importance.

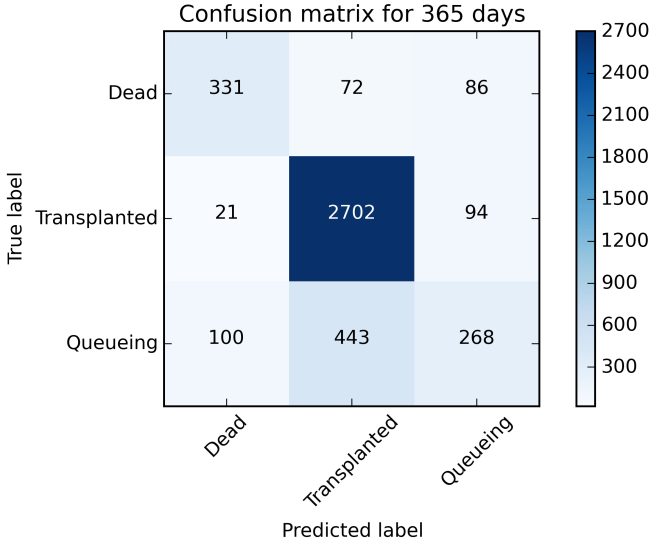

Fig. 2. Confusion matrix for 365 days time period.

The features shared by all of the three sets are: urgency status 2, weight, height and body mass index (BMI). BMI can be considered a feature transformation of weight and height as BMI = weight $\times$ height^2, but it provided extra predictive information over the constituent variables. A sufficiently complex neural network could probably approximate this transformation and therefore BMI would probably not be needed.

Table VI shows some discrepancy between the number of transplanted patients depending on having blood group O. This can probably be explained by the fact that only patients that are blood-group compatible with the donor are transplanted. Even though type O is quite common, patients of this group can only receive from donors from the same blood group and can give to all other types.

A. Future Work

We did not have time to fully optimize the hyperparameters of the neural network and there are some variables that are available that we did not include, both which could produce better results.

We also plan to build a more advanced model based on networks similar to those we described in this paper to be able to estimate the probability the patient would die or would be transplanted depending on the time s/he spent in the waiting list.

ACKNOWLEDGMENT

This work is based on OPTN data as of October 1, 2013 and was supported in part by the Health Resources and Services Administration contract 234-2005-370011C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services. This research was supported by Heart Lung Foundation, The Swedish Research Council, and the eSSENCE program.

TABLE IV

THE TEN MOST CONTRIBUTING FEATURES FOR EACH TIME PERIOD USING BACKWARD ELIMINATION, IN ORDER OF IMPORTANCE.

Rank	180 days	365 days	730 days
1	Urgency status 2	BMI	BMI
2	Weight	Weight	Weight
3	BMI	Height	Height
4	Height	Urgency status 2	Urgency status 2
5	Inotropes	Creatine clearance	Creatinine
6	Blood group: AB	Inotropes	Functional status
7	Life support	Blood group: A	Pulmonary Vascular Resistance
8	Blood group: B	Life support	Educational level: none
9	Inotropic support	Blood group: AB	Ventricular assist type: LVAD + RVAD
10	Ethnicity: black	Blood group: B	Educational level: grade school

TABLE V

EVALUATION ON THE TEST WITH THE 10 BEST FEATURES FOUND FOR EACH TIME PERIOD.

Days	F1 (micro)	F1 (macro)
180	0.710	0.657
365	0.714	0.655
730	0.889	0.660

TABLE VI

THE DISTRIBUTION OF OUTCOMES DEPENDING ON BLOOD GROUP FOR 365 DAYS

Blood group	Dead (%)	Transplanted (%)	Queueing (%)
O	14.2	59.3	26.5
not O	9.7	76.4	13.8

REFERENCES

[1] Michele Barone et al. "ABO blood group-related waiting list disparities in liver transplant candidates: effect of the MELD adoption." In: *Transplantation* 85.6 (2008), pp. 844–849.

[2] François Chollet. *Keras*. https://github.com/fchollet/keras. 2015.

[3] Petra Glander et al. "The 'blood group O problem' in kidney transplantation—time to change?" In: *Nephrology Dialysis Transplantation* 25.6 (2010), p. 1998.

[4] J.C. Hussey, J. Parameshwar, and N.R. Banner. "Influence of Blood Group on Mortality and Waiting Time Before Heart Transplantation in the United Kingdom: Implications for Equity of Access". In: *The Journal of Heart and Lung Transplantation* 26.1 (2007), pp. 30–33. ISSN: 1053-2498.

[5] D. Medved, P. Nugues, and J. Nilsson. "Selection of an optimal feature set to predict heart transplantation outcomes". In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Aug. 2016, pp. 3290–3293.

[6] Johan Nilsson et al. "The International Heart Transplant Survival Algorithm (IHTSA): A New Model to Improve Organ Sharing and Survival". In: *PLoS ONE* 10.3 (2015), e0118644.

[7] United Network for Organ Sharing. *Organ Procurement and Transplantation Network Data*. 2015. URL: http://optn.transplant.hrsa.gov/converge/data/default.asp (visited on 11/19/2015).

[8] DM Powers. "Evaluation: From Precision, Recall and F Factor to ROC, Informedness, Markedness & Correaltion". In: *School of Informatics and Engineering, Flinders University of South Australia Adelaide* (2007).

[9] Kazem Rahimi et al. "Risk Prediction in Patients With Heart Failure". In: *JACC: Heart Failure* 2.5 (2014), pp. 440–446. ISSN: 2213-1779.

[10] Helena Rexius, Folke Nilsson, and Anders Jeppsson. "On the Allocation of Cardiac Allografts from Blood Group-O Donors". In: *Scandinavian Cardiovascular Journal* 36.6 (2002), pp. 342–344.

[11] Michael Schemper and Georg Heinze. "Probability imputation revisited for prognostic factor studies". In: *Statistics in medicine* 16.1 (1997), pp. 73–80.

[12] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[13] Adi L Tarca et al. "Machine learning and its applications to biology". In: *PLOS Computational Biology* 3.6 (2007), e116.

[14] Theano Development Team. "Theano: A Python framework for fast computation of mathematical expressions". In: *arXiv e-prints* abs/1605.02688 (May 2016). URL: http://arxiv.org/abs/1605.02688.

[15] Vincent Van Asch. *Macro-and micro-averaged evaluation measures*. Tech. rep. University of Antwerp, 2013.

[16] Eric S. Weiss et al. "Creation of a Quantitative Recipient Risk Index for Mortality Prediction After Cardiac Transplantation (IMPACT)". In: *The Annals of Thoracic Surgery* 92.3 (2011), pp. 914–922.