Computer Science • 19(2) 2018 https://doi.org/10.7494/csci.2018.19.2.2746

Wojciech Jarmulski
Alicja Wieczorkowska
Mariusz Trzaska
Michał Ciszek
Leszek Paczek

MACHINE LEARNING MODELS FOR PREDICTING PATIENTS SURVIVAL AFTER LIVER TRANSPLANTATION

Abstract *In our work, we have built models predicting whether a patient will lose an organ after a liver transplant within a specified time horizon. We have used the observations of bilirubin and creatinine in the whole first year after the transplantation to derive predictors, capturing not only their static value but also their variability. Our models indeed have a predictive power that proves the value of incorporating variability of biochemical measurements, and it is the first contribution of our paper. As the second contribution we have identified that full-complexity models such as random forests and gradient boosting lack sufficient interpretability despite having the best predictive power, which is important in medicine. We have found that generalized additive models (GAM) provide the desired interpretability, and their predictive power is closer to the predictions of full-complexity models than to the predictions of simple linear models.*

Keywords machine learning, models interpretability, survival prediction, generalized additive models, liver transplant

Citation Computer Science 19(2) 2018: 223–239

1. Introduction

The liver is a complex organ with multiple functions in the human body; for this reason, the quantitative measurement of its state is not straightforward. In contrast to the kidneys (whose function can be assessed indirectly by one parameter – creatinine serum concentration), assessment of the functioning of the liver is more complex and requires the measurement of different parameters [22]:

- excretory function – measured by bilirubin,
- secretory function – measured by albumin and INR (international normalized ratio of prothrombin times),
- detoxification – measured by ammonium.

The various liver functions make determining its state difficult. Additionally, liver damage measured by transaminase can also be accounted for.

The Model for End-Stage Liver Disease (MELD) is the most popular indicator for determining liver condition. It uses bilirubin, creatinine, and INR as presented in the following formula:

$$MELD = 3.78 \cdot ln(bilirubin) + 9.57 \cdot ln(creatinine) + 11.2 \cdot ln(INR) + 6.43$$

In this equation, bilirubin is related to the serum concentration of bilirubin (measured in mg/dL), and creatinine is related to the serum concentration of creatinine (measured in mg/dL). MELD was originally devised as a tool for determining the need for liver transplantation. Today, it has more general application in liver diseases, including assessment of a patient's health after liver transplantation. MELD has some predictive power regarding a patients survival after liver transplantation; however, it is often noted that a more precise and fit-for-purpose indicator or model is needed [6, 17, 26].

We propose to observe the changes of biochemical measurements over time (rather than just base the prediction on a single static measurement) and apply various machine-learning techniques to build models for our problem. The models cover logistic regression, generalized additive models, random forests, and gradient-boosting models. We compared them not only from the perspective of the prediction results but also their interpretability and applicability in medicine. By the interpretability of the models, we mean that users can understand the contribution of individual predictors in the model; i.e., we want models that can quantify the impact of each predictor. We aim to choose the best model for the defined problem by taking into consideration these criteria.

The contribution of this paper is the application and comparison of various machine-learning techniques used to build models on the observed changes of biochemical measurements over time. Our other contribution is verifying the usefulness of biochemical measurements during the period of the first year after liver transplantation on the prediction of a patient's survival within a selected time horizon. The chosen biochemical measurements are bilirubin and creatinine (due to their presence in MELD). INR was skipped due to insufficient measurements in the available data.

Unlike MELD, where the parameters represent a single measurement at a specific point in time, inputs to our models capture the variability of bilirubin and creatinine during the period of a whole year.

Machine-learning applications in medicine are part of a dynamically growing field of research [13, 20]. Random forests and their variations [15, 21] and gradient boosting models [24] are both popular methods in medial classification and survival problems. Generalized additive models have also already been applied in medical problems – in [25], they were used to determine heart transplant survival, and in [3], GAMs were chosen for predicting pneumonia risk and also for their interpretability. There have been many analyses on the prediction of survival after liver transplantation [10, 18, 23], but to the best of our knowledge, none of them have focused on biochemical measurement variability over a period of time after liver transplantation.

The rest of the paper is structured as follows. Section 2 describes the medical dataset as well as the performed data cleaning and preparation operations. Section 3 defines the problem and presents the performed analysis and applied machine-learning methods. In Section 4, we present the results and discuss their interpretation. We conclude in Section 5.

2. Dataset

In this section, we present the analyzed medical dataset, performed data cleaning, and subsequent data preparation for our analysis.

2.1. Dataset description

The dataset for this particular analysis was provided by the Department of Immunology, Transplantology, and Internal Diseases at the Medical University of Warsaw. The data consists of observations (including visits to doctors) of patients who underwent liver transplantation. The observations range from 1994 through the end of 2015. Since 2009, most of the information (e.g., biochemical measurements) is in digital form by default; hence, the data is growing continuously with the information from current visits being added. However, the data prior to 2009 had to be input manually; thus, it is less represented in the data set. Additionally, it is subject to human errors created during input.

The raw input data contains information on 1095 patients with a total of 48,772 observations. Each observation is assigned to a point in time and may contain information about the values of biochemical measurements (in particular, bilirubin and creatinine). The first and most visible characteristic about the available time series is their sparseness and uneven distribution over time. Additionally, the distribution varies between particular biochemical measurements for a given patient; e.g., one sample patient may have only three measurements of bilirubin and more than ten measurements of creatinine.

Moreover, there is noise in the data, which can be divided into the following two groups:

- human errors being the result of manual data input,
- measurement errors from biochemical laboratories; e.g., in the bilirubin data, peaks appear that are measurements not from the patient's body but from a gall container.

The result of the noise in the measurements has been illustrated for a sample patient in Figure 1.

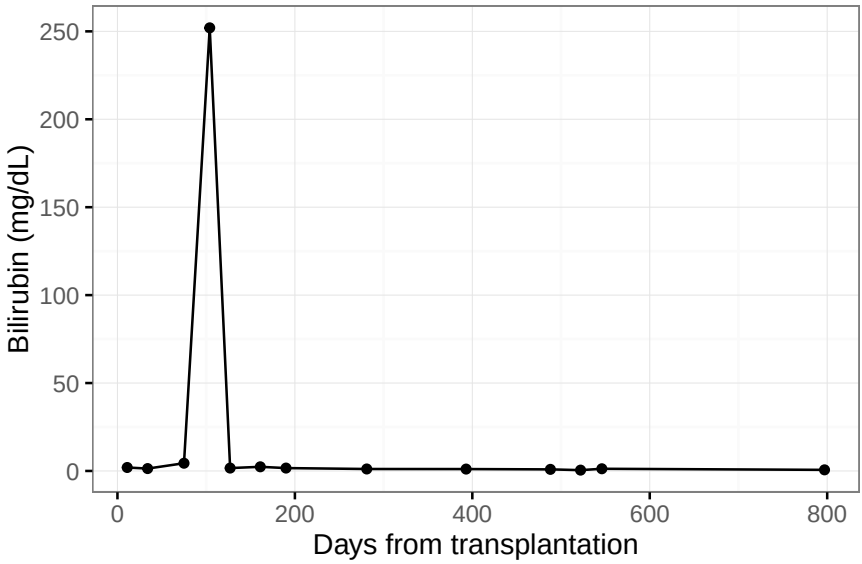


Figure 1. Bilirubine measurements for a sample patient with erroneous measurement

2.2. Data cleaning

The following data-cleaning steps have been performed in preparation for the analysis:

1. Removal of observations that do not contain any measurements – neither for bilirubin nor creatinine.
2. Removal of observations that did not occur within the first year after transplantation – the analysis is focused on measurements during that period only. In this research, we focus on medium- and long-term survival (since it is relatively uncommon in the literature and the majority of the existing research covers short-term survival) [6, 10, 17, 18, 26].
3. Removal of extreme values of bilirubin and creatinine; i.e., erroneous measurements (see Dataset description subsection). The thresholds for the extreme values were chosen based on the domain knowledge.

4. Exclusion of patients whose last observation ended less than one year after transplantation.
5. Exclusion of patients for whom there is no information if they lost their transplant within the analyzed time horizon; i.e., lost to follow-up. This research is focused on the application of classification methods on the survival problem; thus, patients with unknown outcomes had to be excluded from the analysis.
6. Exclusion of patients who had fewer than three observations of both bilirubin and creatinine in the first year after liver transplantation. The analysis is focused on measurement behavior, and the derivation of some of the parameters (see Data Preparation subsection) would not be possible with fewer than three measurements.

2.2.1. Bias

Steps 1 and 2 are standard data-cleaning operations that remove observations not needed for modeling the problem and, thus, do not introduce any bias to the results. Extreme values in the data are assumed to be distributed randomly, as there is no indication of any systematic errors at the source. Therefore, Step 3 likely does not bring any bias. The exclusion of patients in Step 4 is in line with the experiment design; although it might bring bias to the overall class of survival prediction, our focus is on the subset of patients who survived at least one year. Patients are lost to follow-up in our data almost exclusively due to changing their medical units, which can be caused by random events such as changing one's place of residence. Therefore, we assume that no bias was introduced in Step 5. Finally, the data contains patients who have very few observations. This is related to the missing information for patients who were operated on during the early years of the data period. Exclusion of these patients in Step 6 does not bring any significant bias, as it can be assumed that the characteristics of the problem did not change throughout the years within the analyzed period.

2.3. Data preparation

Measurements of bilirubin and creatinine during the first year after a patient's liver transplantation have been captured in the following parameters (Figure 2 presents sample bilirubin and creatinine inputs to derive them):

- minimum, maximum, last value,
- mean, standard deviation, variance,
- skewness and kurtosis,
- amplitude (difference between max and min),
- change (difference between last and first value) expressed as nominal (absolute) value and as percentage,
- percentage of observations above and below norm,
- alpha – slope of regression line,
- average volatility and average daily volatility.

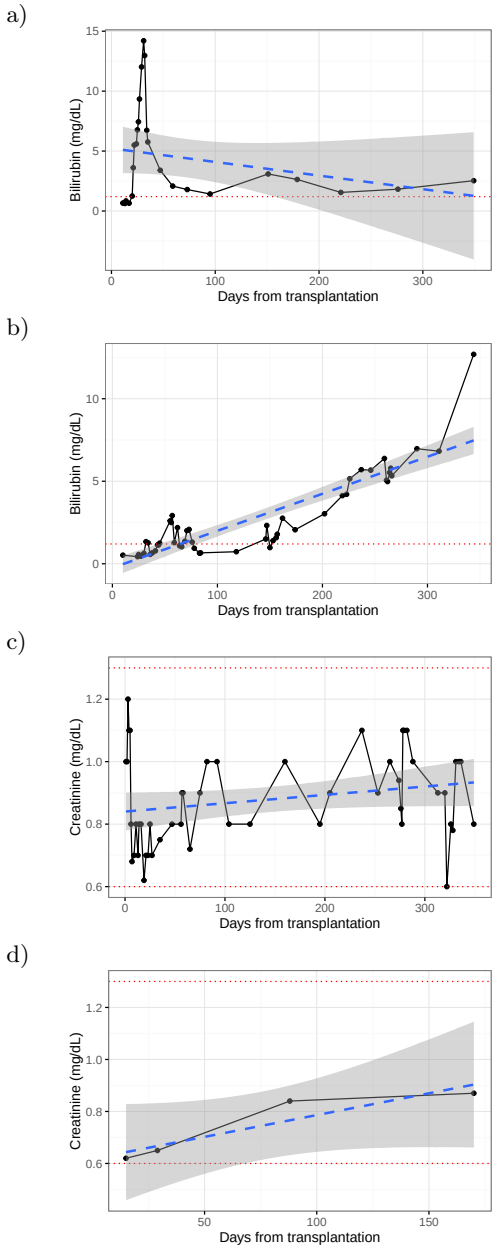
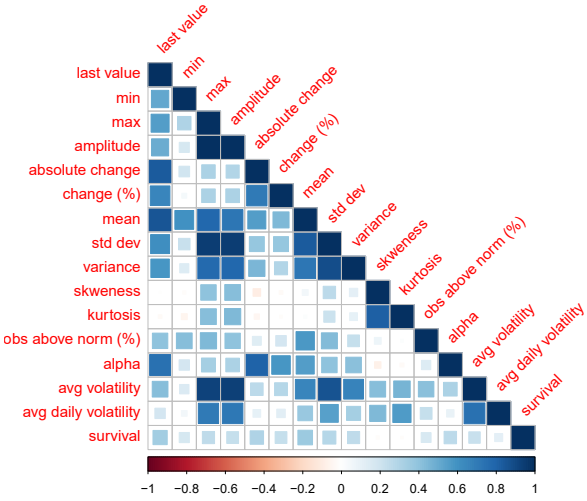


Figure 2. Bilirubin (a, b) and creatinine (c, d) observations from sample patients. Black solid lines represent measurement values in first year after liver transplantation. Blue dashed lines represent regression line, and gray shades – standard error. Vertical dotted lines represent maximum norm values for bilirubin and creatinine and minimum norm value for creatinine (there is no minimum norm for bilirubin)

In order to verify the linear dependencies between the derived parameters, a correlation matrix has been calculated where each value is Pearson's correlation coefficient between a pair of variables. Figure 3 is a visualization of these matrices separately for the bilirubin and creatinine parameters for the three-year survival time horizon.

a)



b)

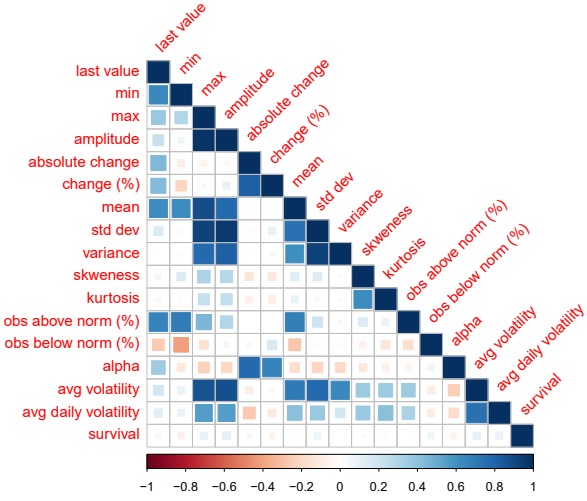


Figure 3. Correlations between derived parameters for bilirubin (a) and creatinine (b) in three-year survival horizon

As can be seen, the dependent variable for a patient's survival is not linearly correlated with any of the parameters (the last row in the matrices). Additionally, before each analysis, those parameters correlated more than 70% are removed using

an automated procedure that removes the most correlated parameters until the desired level is reached. This operation is needed, as some methods (e.g., logistic regression or generalized additive models) do not provide reliable interpretations of the results for data with correlated variables [14].

3. Analysis

Predicting whether a patient will lose an organ[1] during a specified time horizon can be seen as a classification problem. Patients who lost their organs before the specified time period are assigned to Class "1"; otherwise, they belong to Class "0". The analyzed survival time horizon values have been set to three, five, seven, and ten years. These values are most often used in the medical literature [6, 17, 26].

The analysis was performed in the R programming language in Version 3.2.2. All of the techniques and methods described in this section were based on the dedicated packages for that language available via the R package repository – CRAN[2].

3.1. Class imbalance

Table 1 presents the total number of patients and the number of patients assigned to Class "1" for different survival time horizons.

Table 1

Observations and their class distribution for different time horizons. Class "1" represents patients who lost their organ within specified time horizon.

Survival time horizon	Number of observations	Number of Class "1" observations	Percentage of Class "1" observations
Three years	655	49	7.5%
Five years	467	72	15.4%
Seven years	348	86	24.7%
Ten years	204	93	45.6%

It may be observed that, with longer time horizons, the number of patients decreases due to a lack of information at a given point in time about some patient's survival in the longer period. For example, a patient who was five and a half years after transplantation at the end of the analyzed period (the end of 2015) would be included in the three- and five-year survival horizon analyses yet excluded from the seven- and ten-year analyses, respectively. Moreover, with the shorter survival time horizon, there is smaller percentage of Class "1" patients in their total number and larger imbalance between the number of patients in the two classes. This is related to the fact that, during longer periods, more patients would die due to other reasons than losing their livers.

[1]The loss of a liver results in a patient's death in the majority of cases; however, in some cases, it leads to the retransplantation of an organ.

[2]https://cran.r-project.org/

Class imbalance may hamper the performance of some classifiers [14]. There are various approaches to dealing with this issue – simple undersampling and oversampling techniques [11] and more-complex techniques involving combined over- and under-sampling with synthetic sample generation like SMOTE (Synthetic Minority Over-sampling Technique) [4] and ROSE (Random OverSampling Examples) [19]. For this research, ROSE was chosen as a method of choice (R package *ROSE*) due to its simplicity of use and best results achieved. Interestingly, among all of the classifiers tested, the method dealing with class imbalance improved the predictive power of only logistic regression.

3.2. Methods

This subsection describes the methods we have chosen to apply to our dataset to build the classification models for the specified survival problem. We have started with the simplest but most interpretable logistic regression, which assumes linearity and no interactions between the predictors. The generalized additive models include non-linear predictor dependencies while maintaining high interpretability. Finally, we review two types of full-complexity models: random forests and gradient boosting, which model both the non-linearity and interactions between the predictors but are the least interpretable. This comparison is presented in Table 2.

Table 2

Comparison of interpretability and accuracy of different complexity models [16] (+++ means highest level, + lowest)

Model	Form	Interpretability	Accuracy
Generalized linear model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized additive model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full-complexity model	$y = f(x_1, ..., x_n)$	+	+++

Logistic regression is a special case of a generalized linear model (GLM) [11]. GLMs have the following form:

$$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \tag{1}$$

where y is a dependent variable, x_i represent the predictors, and β_i are the coefficients. Function $g(y)$ is the link function; its form depends on the model application. If the link function is identify ($g(y) = y$), then equation 1 describes the linear regression model. For the binary classification model, the link function is logit function $g(y) = log(\frac{y}{1-y})$, and equation 1 becomes the logistic regression that we use throughout the analysis. Logistic regression is the simplest of the methods used in our analysis as well as the most interpretable. It assumes the linear impact of single predictors on the link function of the dependent variable, which is measured by the values of coefficients β_i.

Generalized additive models (GAM) [12] are the extension for generalized linear models and have the following form:

$$g(y) = f_1(x_1) + ... + f_n(x_n) \tag{2}$$

where f_i are called shape functions and are assumed to be smooth and non-linear. This allows a single predictor to have a non-linear impact on the dependent variable. Shape functions are usually based on splines and can take different variations: cubic regression splines, B-splines, P-splines, thin-plate regression splines, and others [27]. In our analysis, we used cubic regression splines due to their universality and good performance. Cubic regression splines are the shape function for each predictor x in the following form: $s(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3$, where α_i are the parameters calculated during the spline-fitting process. We used R package *mgcv* with the default settings.

Random forests belong to the full-complexity models and have the following form:

$$y = f(x_1, ..., x_n).$$

Full complexity models are usually more accurate than simpler models because they model both non-linearity and interaction; however, they are so complex that it is very hard to interpret them [16]. Random forests [2] are ensembles of decision trees. The method builds a large collection of decorrelated decision trees and then averages their outputs in regression problems or takes the majority vote in classification problems. This allows us to reduce the variance of an estimated prediction of the underlying decision trees. We chose random forests (R package *randomForest*) as a representative full-complexity model because they often demonstrate better performance than other standard classifiers, are easy to tune, and are robust to overfitting; for these reasons, they are often recommended as a universal machine-learning method [7]. We used the default number of variables randomly sampled as candidates at each tree split equal to the square root of the number of predictors and built the classifiers with 1000 trees.

Gradient boosting is another technique for full-complexity models and, similar to random forests, creates ensembles of weak learners typically decision trees. Unlike random forests (which use the bagging technique to create ensembles), gradient-boosting models are based on boosting method [8, 9] and often lead to better accuracy than with random forests. Their disadvantage is that they are more difficult to train than random forests, as the model itself has more parameters that must be tuned; namely, shrinkage (learning rate), tree size, the number of trees, and the subsampling rate (the subset of predictors used when building a single tree). Nevertheless, a gradient-boosting model is one of the most powerful machine-learning techniques; its variations are used in the winning solutions of many data science competitions [5]. For the analysis, we used R package *gbm* with a default shrinkage of 0.001 and 1000 trees in the final classifier.

4. Results and discussion

To measure how well a particular model performed at predicting whether a patient lost an organ during the survival time horizon, we evaluated the ability of classifiers to properly assign test patients to either Class "1" or "0". For each classifier, we took the probabilities obtained by the tested patients and measured the area under the receiver operating characteristic curve (i.e., AUC-ROC). The ROC curve graphically displays the trade-off between sensitivity and specificity (Fig. 4) and is useful in assigning the best cut-offs for clinical use [1]. AUC represents the overall accuracy of a classifier and provides a useful parameter for comparing performance between models. AUC is computed by integrating the ROC curve and is lower bounded by 0.5. AUC can be interpreted as the probability of a model properly assigning a patient to Class "1" when choosing from one randomly sampled patient in Class "1"' and one from Class "0".

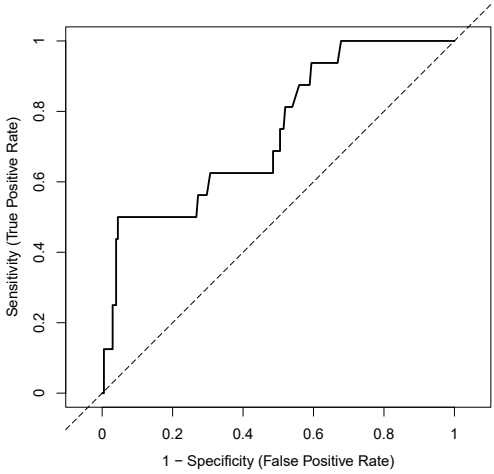


Figure 4. ROC curve (black line) with AUC of 0.7421 for random forest model for three-year survival time horizon. Dashed line represents baseline ROC curve

The test results have been obtained using 3-fold cross validation repeated 20 times for each method and each time horizon. This allowed us to obtain reliable AUC values represented by their mean and standard deviation. This also introduced correction for overfitting. The results for all of the machine-learning methods are reported in Table 3 and compared side by side in Figure 5.

In line with expectations, the best results on average were achieved by the full-complexity models, while logistic regression gave the weakest results. Interestingly, the GAM results are closer on average to the full-complexity models than to the linear models. On average, there was no significant difference (t-test with $p < 0.05$) between the random forests and the gradient boosting models; however, on most survival time horizons, random forests gave slightly better results.

Table 3

AUC for survival prediction across time horizons for tested machine-learning methods and MELD

Survival time horizon	Logistic regression	GAM	Random forest	Gradient boosting	MELD
Three years	0.6175368	0.6643713	0.6877293	0.6697647	0.578869
Five years	0.6135029	0.6744619	0.6155737	0.6455466	0.5796932
Seven years	0.6649711	0.6682012	0.7213799	0.7078518	0.5911405
Ten years	0.6659613	0.6818575	0.7335804	0.7270159	0.5769097
Average	0.6404930	0.6722230	0.6895658	0.6875448	0.5816531

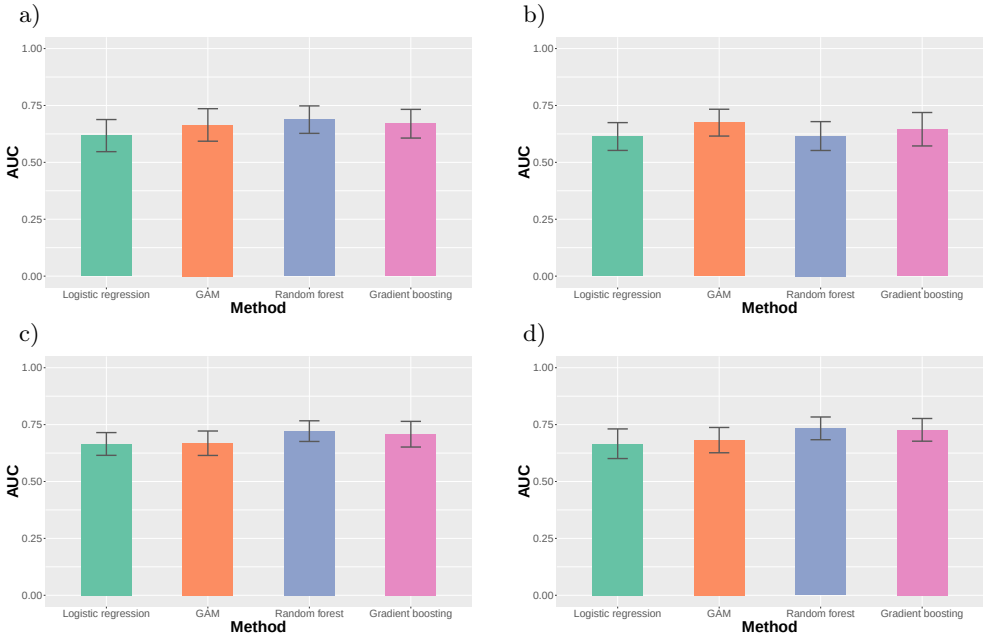


Figure 5. Classifier AUC results with standard deviation (error bars on top of each bar) for various methods across different survival horizons: 3-year (a); 5-year (b); 7-year (c); 10-year (d). On each subfigure, bars represent machine-learning methods (starting from left): logistic regression (green); GAM (orange); random forest (blue); and gradient boosting (purple)

Although our research is focused on a comparison of the modeling methods, we have also provided a comparison of our models with MELD (Tab. 3). MELD is significantly worse than all of the models irrespective of the method used. This validates our approach to incorporate variability into our models. In the GAM models, it is also possible to examine the shape functions and explicitly examine the direct impact of each predictor on the outcome probability. Figure 6 presents significant (i.e., non-zero) shape functions for the final GAM model, predicting the probability of a patient losing an organ during the five-year time horizon.

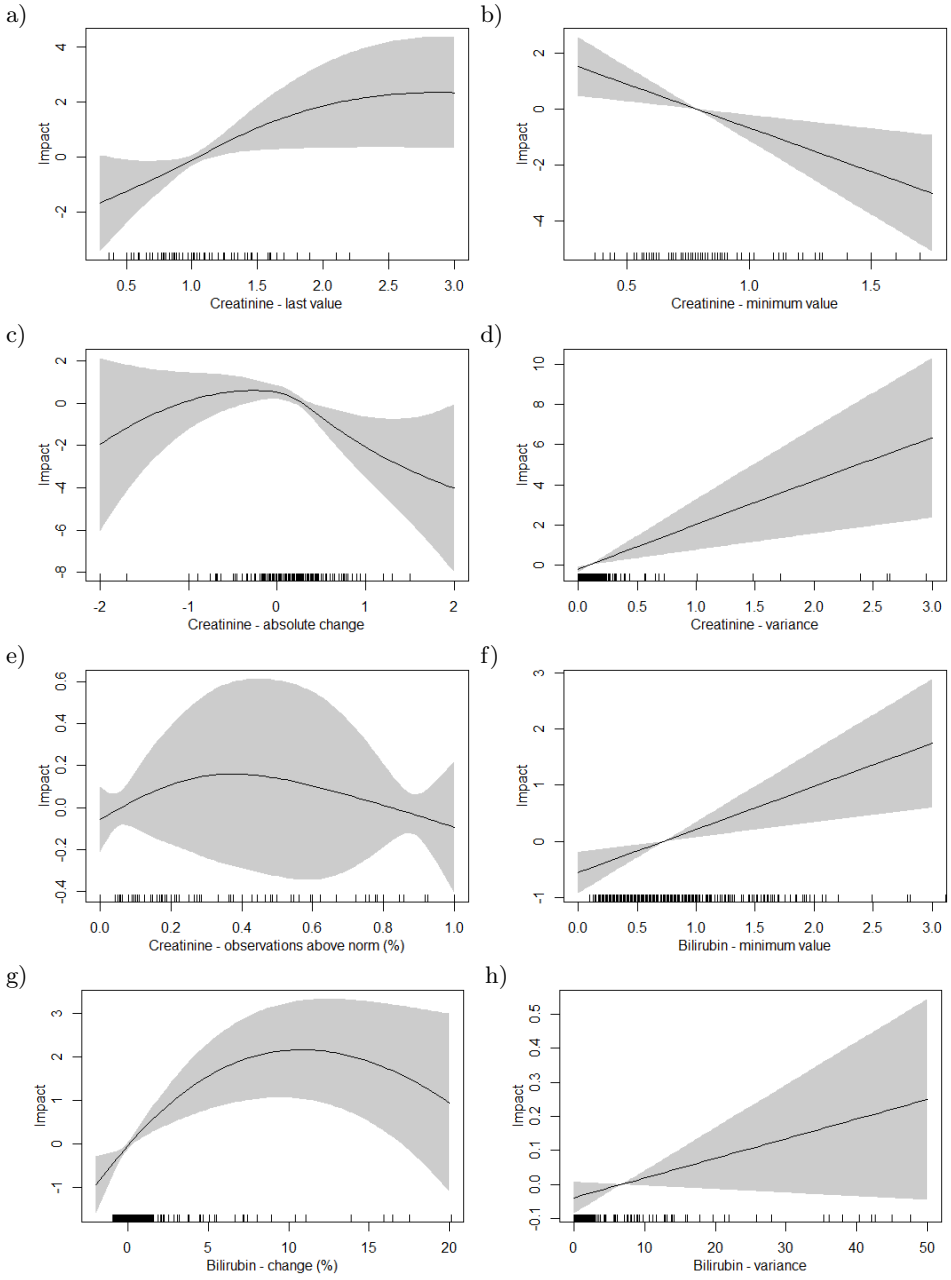


Figure 6. Visualization of impact on output classifier probability of selected predictors. Gray shades represent confidence bands, and ticks on horizontal axis represent observations. Impact above 0 means that probability of patient being assigned to Class "1" is increased, while impact below 0 decreases this probability. Impact is equivalent to odds ratio in logistic regression [11]

While some shape functions (e.g., Fig. 6b and 6h) in most of their domains could be substituted with linear functions without significant losses, other predictions (e.g., Fig. 6c and 6e) have a clearly non-linear impact on the output, and logistic regression would not be able to capture such complex dependencies. For instance, when the last value of creatinine (Fig. 6a) is below 1.0, it decreases the probability of a patient losing an organ. For values above 1.0, the risk of losing an organ increases proportionally with the value. Finally, for very high values of creatinine (above 25), the risk stops increasing remaining at the same (although high) level.

From a medical perspective, we have proven that, using the observations of bilirubin and creatinine during the first year, it is possible to model whether a patient will lose an organ during a specified time horizon, with AUC reaching a value of 0.73 in the best-case scenario. This result could be further improved by adding observations from other biochemical measurements; however, it proves that predictors derived from variability measurements of bilirubin and creatinine already have predictive power in our problem. Interestingly, there is a better predictability for longer survival time horizons than for shorter periods (AUC within a range of 0.72–0.73 for seven- and ten-year time horizons vs. AUC of 0.67–0.68 for three- and five-year time horizons). This could be explained by the fact that serious consequences occur during shorter time horizons after liver transplantation; thus, it is easier to predict whether patient would lose an organ within a longer period.

Finally, we have not only provided results of the models' predictive power but also the impact of each predictor in the GAM models in an easily interpretable graphical form. Based on this, medical doctors can understand how predictions are made and, in some cases, discard particular parameters if they have been wrongly set up. In medical applications, it might be worth sacrificing some predictive power for gaining the possibility of interpreting a parameter's impact on the output. This is provided by the GAMs.

Our study has a few limitations. First, we concentrated only on patients who survived at least one year after liver transplantation. Second, to derive the variability predictors for the first year after liver transplantation, we had to exclude patients who had very few observations during this period (fewer than three), which limits the applicability of our models. Finally, we have proven the predictive power of the variability predictors but did not compare them against models with other predictor selections.

Our aim for future research is to improve the GAM results while maintaining their interpretability, which would make them even more competitive as a machine-learning method. Having proven that variability predictors can build successful predictive models, we plan to compare the standard approach where only static measurements are used for modeling against models using additional variability predictors.

5. Conclusion

In this work, we have built models predicting whether a patient will lose an organ after liver transplantation within a specified time horizon. Inspired by MELD (the

most popular indicator for determining liver condition), we have used the observations of bilirubin and creatinine. Unlike MELD, our predictors are derived from the observations during the whole first year after transplantation, trying to capture not only the static value but also variability. These predictors indeed have a predictive power, which proves the value of incorporating biochemical measurement variability in models (which is the first contribution of our paper). We hypothesize that the accuracy could be increased by adding other biochemical measurements.

As a result of the analysis, we have found that full-complexity models such as random forests and gradient boosting lack sufficient interpretability despite having the best predictive power (which is important in medicine). Our contribution is the finding that generalized additive models provide the desired interpretability, and their results in prediction are relatively closer to full-complexity models than simple linear models are. This property of GAMs makes them well-suited models in medical applications where, apart from predictive power, interpretability is also important.

Acknowledgements

This work was partially supported by the Research Center of PJAIT, supported by the Ministry of Science and Higher Education in Poland.

The authors would like to show gratitude to the Department of Immunology, Transplantology, and Internal Diseases at the Medical University of Warsaw for providing the dataset.

References

[1] Boyd J.: Statistical analysis and presentation of data. In: *Evidence-Based Laboratory Medicine*, pp. 113–140, AACC Press Washington, DC, 2007.

[2] Breiman L.: Random forests, *Machine Learning*, vol. 45(1), pp. 5–32, 2001.

[3] Caruana R., Lou Y., Gehrke J., Koch P., Sturm M., Elhadad N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. ACM, 2015.

[4] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.: SMOTE: synthetic minority over-sampling technique, *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[5] Chen T., Guestrin C.: XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016.

[6] Cholongitas E., Marelli L., Shusang V., Senzolo M., Rolles K., Patch D., Burroughs A.K.: A systematic review of the performance of the model for end-stage liver disease (MELD) in the setting of liver transplantation, *Liver Transplantation*, vol. 12(7), pp. 1049–1061, 2006.

[7] Fernández-Delgado M., Cernadas E., Barro S., Amorim D.: Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, vol. 15(1), pp. 3133–3181, 2014.

[8] Friedman J.H.: Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, vol. 29(5), pp. 1189–1232, 2001.

[9] Friedman J.H.: Stochastic gradient boosting, *Computational Statistics & Data Analysis*, vol. 38(4), pp. 367–378, 2002.

[10] Habib S., Berk B., Chang C.C.H., Demetris A.J., Fontes P., Dvorchik I., Eghtesad B., Marcos A., Shakil A.O.: MELD and prediction of post-liver transplantation survival, *Liver Transplantation*, vol. 12(3), pp. 440–447, 2006.

[11] Hastie T.J., Tibshirani R.J., Friedman J.: *The elements of statistical learning: data mining, inference, and prediction. Second Edition*, Springer, 2009.

[12] Hastie T.J., Tibshirani R.J.: *Generalized additive models*, CRC Press, 1990.

[13] Herland M., Khoshgoftaar T.M., Wald R.: A review of data mining using big data in health informatics, *Journal of Big Data*, vol. 1(1), pp. 1–35, 2014.

[14] Kuhn M., Johnson K.: *Applied predictive modeling*, Springer, 2013.

[15] Liu K.H., Huang D.S.: Cancer classification using Rotation Forest, *Computers in Biology and Medicine*, vol. 38(5), pp. 601–610, 2008.

[16] Lou Y., Caruana R., Gehrke J.: Intelligible models for classification and regression. In: *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 150–158. ACM, 2012.

[17] Luca A., Angermayr B., Bertolini G., Koenig F., Vizzini G., Ploner M., Peck-Radosavljevic M., Gridelli B., Bosch J.: An integrated MELD model including serum sodium and age improves the prediction of early mortality in patients with cirrhosis, *Liver Transplantation*, vol. 13(8), pp. 1174–1180, 2007.

[18] Mazzaferro V., Llovet J.M., Miceli R., Bhoori S., Schiavo M., Mariani L., Camerini T., Roayaie S., Schwartz M.E., Grazi G.L., Adam R., Neuhaus P., Salizzoni M., Bruix J., Forner A., De Carlis L., Cillo U., Burroughs A.K., Troisi R., Rossi M., Gerunda G.E., Lerut J., Belghiti J., Boin I., Gugenheim J., Rochling F., Van Hoek B., Majno P., Metroticket Investigator Study Group: Predicting survival after liver transplantation in patients with hepatocellular carcinoma beyond the Milan criteria: a retrospective, exploratory analysis, *The Lancet Oncology*, vol. 10(1), pp. 35–43, 2009.

[19] Menardi G., Torelli N.: Training and assessing classification rules with imbalanced data, *Data Mining and Knowledge Discovery*, vol. 28(1), pp. 92–122, 2014.

[20] Miotto R., Li L., Kidd B.A., Dudley J.T.: Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, *Scientific Reports*, vol. 6, p. 26094, 2016. `http://dx.doi.org/10.1038/srep26094`.

[21] Ozcift A., Gulten A.: Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms, *Computer Methods and Programs in Biomedicine*, vol. 104(3), pp. 443–451, 2011.

[22] Pratt D.S., Kaplan M.M.: Evaluation of liver function. In: *Harrison's Principles of Internal Medicine*, 17th ed., pp. 1923–1926. McGraw-Hill Medical Publishing Division, New York, 2008.

[23] Roberts M.S., Angus D.C., Bryce C.L., Valenta Z., Weissfeld L.: Survival after liver transplantation in the United States: a disease-specific analysis of the UNOS database, *Liver Transplantation*, vol. 10(7), pp. 886–897, 2004.

[24] Thongkam J., Xu G., Zhang Y., Huang F.: Breast cancer survivability via AdaBoost algorithms. In: *Proceedings of the second Australasian workshop on Health data and knowledge management*, vol. 80, pp. 55–64, 2008.

[25] Tsujitani M., Tanaka Y.: Analysis of heart transplant survival data using generalized additive models, *Computational and Mathematical Methods in Medicine*, 2013.

[26] Watt K., Menke T., Lyden E., McCashland T.M.: Mortality while awaiting liver retransplantation: predictability of MELD scores, *Transplantation Proceedings*, vol. 37, pp. 2172–2173, 2005.

[27] Wood S.: *Generalized additive models: an introduction with R*, CRC Press, 2006.

Affiliations

Wojciech Jarmulski
Polish-Japanese Academy of Information Technology, wojciech.jarmulski@pja.edu.pl

Alicja Wieczorkowska
Polish-Japanese Academy of Information Technology, alicja@poljap.edu.pl

Mariusz Trzaska
Polish-Japanese Academy of Information Technology, mtrzaska@pjwstk.edu.pl

Michał Ciszek
Medical University of Warsaw, Department of Immunology, Transplantology and Internal Diseases, michal.ciszek@wum.edu.pl

Leszek Paczek
Medical University of Warsaw, Department of Immunology, Transplantology and Internal Diseases, leszek.paczek@wum.edu.pl

Received: 05.01.2018
Revised: 17.02.2018
Accepted: 17.02.2018