

Predicting the risk of all-cause mortality in patients with type 2 diabetes mellitus

Vershinina O.^{1*}, Kondakova E.¹, Sabbatinelli J.^{2,3}, Colombaretti D.², Bonfigli A.R.⁴, Franceschi C.¹, Ivanchenko M.¹, Olivieri F.^{2,5}

¹ Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia

² Department of Clinical and Molecular Sciences, Università Politecnica delle Marche, Ancona, Italy

³ Clinic of Laboratory and Precision Medicine, IRCCS INRCA, Ancona, Italy

⁴ Scientific Direction, IRCCS INRCA, Ancona, Italy

⁵ Advanced Technology Center for Aging Research, IRCCS INRCA, Ancona, Italy

* olya.vershinina@itmm.unn.ru

Key words: diabetes; mortality risk; machine learning; artificial intelligence; explainability

Motivation and Aim: Diabetes mellitus is a non-infectious endocrine disease characterized by high blood sugar levels. According to the International Diabetes Federation, its prevalence is steadily increasing: in 2021, there were approximately 536.6 million cases of diabetes diagnosed among people aged 20–79 years, and by 2045 this number is estimated to increase to 783.2 million people [1]. Diabetes leads to serious complications and increased mortality – diabetes was responsible for 6.7 million deaths worldwide in 2021. To improve the life quality and expectancy of patients with diabetes, it is necessary to develop personalized treatment strategies. Information about individual mortality risk predictions may be of great importance in choosing appropriate medications and lifestyle recommendations. Based on machine learning algorithms and explainable artificial intelligence, we develop an explainable model for predicting the mortality risk of patients with diabetes and identify important predictors that influence life expectancy.

Methods and Algorithms: We consider a dataset of 568 patients with type 2 diabetes mellitus (T2DM) who were followed at the Italian National Research Center on Aging (Ancona, Italy) for 16.8 years [2]. Among the patients, there were 308 men and 260 women aged from 40 to 87 years, the average age of patients was 65.74 ± 8.15 years. The dataset includes more than 100 different demographic, laboratory and clinical characteristics (features) that are considered risk factors for mortality. The data was pre-processed and divided into training and test sets. In constructing the predictive model, we considered the two most widely used survival models – Cox proportional hazards regression [3] and random survival forests [4]. In order to obtain an optimal and small subset of the most significant features, different feature selection strategies were used with each survival model: feature selection based on univariate Cox regression, forward selection method, backward selection method, selection based on feature importance. The quality of mortality risk prediction models was assessed using two metrics: Harrell's C-index and area under the receiver operating characteristic curve (AUC). Both quality metrics can take values from 0 (poor prediction) to 1 (perfect prediction). In order to increase user confidence in the final prediction model, an explainable artificial intelligence approach was used – SHapley Additive exPlanations (SHAP) [5], which allows interpreting model predictions both globally at the level of the entire sample and locally at the level of specific patients.

Results: As a result of experiments with different feature selection strategies and different survival algorithms, more than ten models were obtained that predict the mortality risks of patients with T2DM. It is interesting to note that all the resulting feature subsets included the age of the patients as one of the most important characteristics affecting life expectancy. More accurate models were obtained using random survival forests, for which the quality metric of the 15-year prediction on the test set varied within the AUC range from 0.77 to 0.82. The model with the highest AUC value was selected as the final model. The risk of mortality predicted by the model was significantly associated with survival, which was also confirmed by constructing Kaplan–Meier survival curves and using the log-rank test. Using the SHAP method, graphs of the global importance of features were constructed, as well as graphs of local explainability, which reflect which features increase the risk assessment of specific patients, and which, on the contrary, decrease.

Conclusion: Thus, an explainable machine learning model was developed to predict the mortality risks of patients with type 2 diabetes, which has high quality scores on both the training and test sets. The advantage of the model is the ability to assess the contribution of features to predictions and explain the predictions made.

References

1. Sun H. et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract.* 2022;183:109119. doi 10.1016/j.diabres.2021.109119
2. Bonfigli A.R. et al. Leukocyte telomere length and mortality risk in patients with type 2 diabetes. *Oncotarget.* 2016;7(32):50835-50844. doi 10.18632/oncotarget.1061
3. Cox D.R. *Analysis of Survival Data.* Chapman and Hall/CRC, 1984
4. Ishwaran H., Kogalur U.B., Blackstone E.H., Lauer M.S. Random survival forests. *Ann Appl Stat.* 2008;2(3):841-860. doi 10.1214/08-AOAS169
5. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 2017;4765-4774