Prediction of Type 2 Diabetes Mellitus with an Enhanced k-Fold Cross-Validation Method

Vlad-Florin CHELARU

Faculty of Medicine, Iuliu Hațieganu University of Medicine and Pharmacy Cluj-Napoca, Victor Babeş Str., no. 8, 400012 Cluj-Napoca, Romania
E-mail: vladflorinchelaru@gmail.com

* Author to whom correspondence should be addressed

Abstract

Background and Aim: Multiple models available for predicting the development of Type II Diabetes Mellitus (T2DM) contain at least seven predictors (Cambridge Score – 7; QDScore – 10; QDiabetes – at least 16), some of which require invasive techniques. Our aim was to identify, if any, a predictive model based on observations or non-invasive measurements with minimum number of predictors for T2DM. *Materials and Methods:* We used the NAGALA study, a prospective follow-up study conducted on Japanese population with observations and measurements (e.g., blood pressure or ultrasound non-alcoholic fatty liver disease - NAFLD) as input data and logistic regression analysis as method. We divided the database in 12 balanced folds, and we then recombined the folds in a 9:3 ratio to create 220 pairs of training-testing sets. We employed a layer-based approach, adding each predictor to the model akin to a forward stepwise algorithm. In each layer, we maximized the standardized accuracy (defined as the mean of sensitivity and specificity). *Results:* We analyzed 15464 patients, of which 373 (2.41%) developed T2DM. The median follow-up time was 5.39 years. The final model included the following predictors, in order: NAFLD, alcohol consumption, age, and obesity (body mass index over 25kg/m^2). In testing, the model obtained a mean standardized accuracy of 73.9% (on testing sets), and after training the reported model on the entire dataset, the final model had a sensitivity of 77.5% [95%CI 73.2% to 81.7%] (where CI = confidence interval) and a specificity of 71.7% [95%CI 71% to 72.4%] (on the entire dataset). *Conclusion*: The final model has limited clinical usability. Future research should compare the reliability of this method against standard methods, as well as its robustness when predictors are correlated.

Keywords: Diabetes Mellitus type 2 (T2DM); Non-Alcoholic Fatty Liver Disease (NAFLD); Statistics; Logistic models; Reproducibility of results