Prediction of 3-Year All-Cause Death in a Percutaneous Coronary Intervention Registry using Machine Learning: A Comparison Between Random Forest and CatBoost Algorithms

Paul-Adrian CĂLBUREAN[a,b], Paul GREBENIŞAN[b], Victor VACARIU[b], Reka-Katalin DRINCAL[b], Oana ŢEPES[b], Iulia GRANCEA[b], Ioana ŞUŞ[b], Cristina SOMKEREKI[b], Valentin SIMON[b], Zoltán DEMJÉN[b], István ADORJÁN[b], Irina PINITILIE[b], Anca Teodora DOLCOŞ[b], Tiberiu OLTEAN[b], László HADADI[b,*], Marius MĂRUŞTERI[a]

[a] University of Medicine, Pharmacy, Sciences and Technology "George Emil Palade" of Târgu Mureş, Gheorghe Marinescu Str., no. 38, Târgu Mureş, Romania
[b] Emergency Institute for Cardiovascular Diseases and Transplantation Târgu Mureş, Târgu Mureş, Romania
E-mails: calbureanpaul@gmail.com; hadadilaci@yahoo.com; marius.marusteri@umfst.ro

* Author to whom correspondence should be addressed

Abstract

Background and Aim: Risk stratification in patients undergoing percutaneous coronary intervention (PCI) procedures is a major objective in clinical practice since it guides appropriate therapy selection. Machine learning (ML) models are complex automated decision systems and numerous algorithms have been developed. Our aim was to compare random forest, a traditional ML algorithm, with gradient boosting with categorical features support (CatBoost), a newer ML algorithm, in predicting 3-year all-cause death in a PCI population. *Materials and Methods:* All patients older than 18 years and treated by PCI in a tertiary care centre between January 2016 – December 2017 have been included after hospital discharge in this registry. Mortality rates at 3 years were documented from the Romanian National Health Insurance System database. A total of 120 clinical variables were used to train the two ML algorithms. Training was performed on 70% of the dataset and testing was performed on the remaining 30% of the dataset. *Results:* A total of 2242 patients were included, of which 336 (14.9%) were deceased at 3-year follow-up. Area under receiver-operator curve for 3-year all-cause mortality prediction for CatBoost was 0.848, while for random forest was 0.802 (DeLong p=0.001). Three most important clinical variables for both ML models were age, left ventricular ejection fraction and serum creatinine. Brier scores for random forest and CatBoost were 0.121 and 0.102 respectively, indicating a good fit of the ML-based models. *Conclusions:* Among aggregated decision trees ML algorithms, CatBoost has superior predictive capacity of adverse clinical events in a PCI population when compared with random forest.

Keywords: Percutaneous coronary intervention; Coronary artery disease; Machine learning (ML); Artificial intelligence (AI)

Funding: This study was funded by the Romanian Academy of Medical Sciences and European Regional Development Fund, MySMIS 107124: Funding Contract 2/Axa 1/31.07.2017/ 107124 SMIS