High Quality Phenotypic Data and Machine Learning Beat a Generic Risk Score in the Prediction of Mortality in Acute Coronary Syndrome

by Kari Antila (VTT), Niku Oksala (Tampere University Hospital) and Jussi A. Hernesniemi (Tampere University)

We set out to find out if models developed with a hospital's own data beat a current state-of-the art risk predictor for mortality in acute coronary syndrome. Our data of 9,066 patients was collected and integrated from operational clinical electronic health records. Our best classifier, XGBoost, achieved a performance of AUC 0.890 and beat the current generic gold standard, GRACE (AUC 0.822).

The use of electronic health records (EHRs) as a source of "big data" in cardiovascular research is attracting interest and investments. Integrating EHRs from multiple sources can potentially provide huge data sets for analysis. Another potentially very effective approach is to focus more on data quality instead of quantity. We evaluated the applicability of large-scale data integration from multiple electronic sources to produce extensive and high quality cardiovascular (CVD) phenotype data for survival analysis and the possible benefit of using novel machine learning [1]. For this purpose, we integrated clinical data recorded by treating physicians with other EHR data of all consecutive acute coronary syndrome (ACS) patients diagnosed invasively by

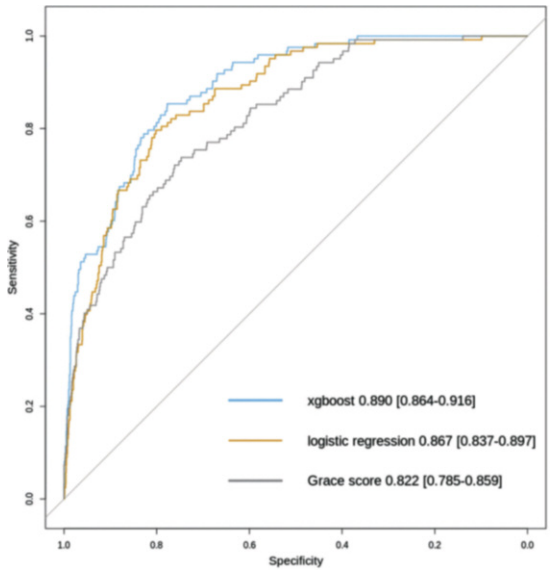

Figure 1: Comparison of model performance by receiving operating characteristic curves for different risk prediction models for six month mortality among patients undergoing coronary angiography in Tays Heart Hospital for acute coronary syndrome during years 2015 and 2016 (n = 1722 with n = 122 fatalities during a six-month follow-up).

coronary angiography over a 10-year period (2007 -2017).

To achieve this, we generated high quality phenotype data for a retrospective analysis of 9,066 consecutive patients (95% of all patients) undergoing coronary angiography for their first episode of ACS in a single tertiary care centre. Our main outcome was six-month mortality. Using regression analysis and machine learning method extreme gradient boosting (XGBoost) [2], multivariable risk prediction models were developed in a separate training set (patients treated in 2007-2014 and 2017, n=7151) and validated and compared to the Global Registry of Acute Coronary Events (GRACE) [3] score in a validation set (patients treated in 2015-2016, n=1771) with the full GRACE score data available.

In the entire study population, overall six-month mortality was 7.3 % (n=660). Many of the same variables were associated highly significantly with six-month mortality in both the regression and XGBoost analyses, indicating good data quality in the training set. Observing the performance of these methods in the validation set revealed that xgboost had the best predictive performance (AUC 0.890) when compared to logistic regression model (AUC 0.871, p=0.012 for difference in AUCs) and compared to the GRACE score (AUC 0.822, p<0.00001 for difference in AUCs) (Figure 1).

These results show that clinical data as recorded by physicians during treatment and conventional EHR data can be combined to produce extensive CVD phenotype data that works effectively in the prediction of mortality after ACS. The use of a machine learning algorithm such as gradient boosting leads to a more accurate prediction of mortality when compared to conventional regression analysis. The use of CVD phenotype data, either by conventional logistic regression or by machine learning, leads to significantly more accurate results when compared to the highly validated GRACE score specifically designed for the prediction of six-month mortality after admission for ACS. In conclusion, the use of both high quality phenotypic data and novel machine learning significantly improves prediction of mortality in ACS over the traditional GRACE score.

This study was part of the MADDEC (Mass Data in Detection and prediction of serious adverse Events in Cardiovascular diseases) project supported by Business Finland research funding (Grant no. 4197/31/2015) as apart of a collaboration between Tays Heart Hospital, University of Tampere, VTT Technical Research Centre Finland Ltd, GE Healthcare Finland Ltd, Fimlab laboratories Ltd, Bittium Ltd and Politechinco di Milano.

References:

[1] J.A. Hernesniemi, S. Mahdiani, J.A.T. Tynkkynen, et al.: " Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome – the MADDEC study", 2019. Annals of Medicine. https://doi.org/10.1080/07853890.2019.1596302

[2] T. Cheng, C. Guestrin: "XGBoost: A Scalable Tree Boosting System", KDD '16, 2016. https://doi.org/10.1145/2939672.2939785

[3] K. Fox, J.M. Gore, K. Eagle, et al. : "Rationale and design of the grace (global registry of acute coronary events) project: A multinational registry of patients hospitalized with acute coronary syndromes", Am Heart J 141:190–199, 2001. https://doi.org/10.1067/mhj.2001.112404

Please contact:

Kari Antila
VTT Technical Research Centre of Finland ltd
+358 40 834 7509