Classification of atherothrombotic events in myocardial infarctions survivors with supervised machine learning using data from an electronic health record system

Holger Kunz§, Laura Pasea§, Spiros Denaxas§
§Institute of Health Informatics, University College London, United Kingdom

Abstract. The aim was to build a prediction model for subsequent atherothrombotic events for patients who survived a myocardial infarction. The dataset contained 7,582 patients from a national Electronic Health Record. The prediction is a binary outcome (event and no event) in a period of five years after a myocardial infarction. Different classifiers were tested and XGBoost achieved the best F1-score=0.76. Top features are: imd_score, age_at_entry, egfr_ckdepi_base, height, and SBP_base.

Keywords. classification, atherothrombosis, supervised machine learning, prognosis

1. Introduction

The aim is to use data from Electronic Health Records (EHR) to predict atherothrombotic events as a binary outcome (event and no event) in a period of five years after a myocardial infarction. Data from patients who were still alive one year after their last acute MI from the CALIBER programme [1] were used. CALIBER uses three national structured EHR sources: primary care health records (coded diagnoses, clinical measurements, and prescriptions) from the Clinical Practice Research Datalink (CPRD); coded hospital discharges (Hospital Episode Statistics, HES); and death registrations.

2. Methods

For the supervised classification a dataset of 7,582 patients was used. Follow-up started at year 1 after index acute MI, and patients were censored at the earliest data of the endpoints, death or 5 years of follow-up. The endpoint is defined as a binary outcome whether a patient has an atherothrombotic event in a period of five years. In total 62 clinical features were used. Classifiers such as SVM, XGBoost, Neural Network (MLP), and Naïve Bayes were applied. Five-fold cross-validation was used for hyperparameter tuning on 5307 instances. For the final evaluation 2275 instances were used. The optimization focused on the F1-score as the dataset was imbalanced. The implementation was done in Python.
Results

The best F1-score has XGBoost with F1=0.76 and AUROC=0.83. The best AUROC has the SVM with AUROC=0.84 and F1=0.74. MLP has an AUROC=0.82 and F1=0.75. Naïve Bayes has an AUROC=0.79 and F1=0.69. Testing the XGBoost-model with 2275 instances achieved a confusion matrix of 1244 (TN), 327 (FN), 534 (TP), and 170 (FP). The following top five features were derived with the information weight from XGBoost: 1) imd_score (0.089), 2) age_at_entry (0.0698), 3) egfr_ckdepi_base (0.0578), 4) height (0.0508), and 5) SBP_base (0.048). Figure 1 shows a comparison of different classifiers.

![Figure 1: Comparison of classifiers](image)

3. Discussion and Conclusion

The applied classifiers achieved a reasonable performance. XGBoost has the advantage to provide a feature ranking that can further inform clinicians. Similar discrete single time point models with a fixed time period have been used for the prognosis of cancer [2]. The dataset can also be used for time-to-event analysis [3]. A limitation of this study is the usage of a proportion of the data for testing purposes to achieve a robust and reliable estimate on unseen data. Finally, it may be concluded that the performance of the prediction algorithms and the derived feature ranking could be useful for a potential clinical decision support system (DSS). A further feasibility study could evaluate its usage in clinical practice.

References