Deep and machine learning models to improve risk prediction of cardiovascular disease using data extraction from electronic health records

I Korsakov, A Gusev, T Kuznetsova, D Gavrilov, R Novitskiy
European Heart Journal, Volume 40, Issue Supplement_1, October 2019, ehz748.0670,
https://doi.org/10.1093/eurheartj/ehz748.0670
Published: 21 October 2019

Abstract

Advances in precision medicine will require an increasingly individualized prognostic evaluation of patients in order to provide the patient with appropriate therapy. The traditional statistical methods of predictive modeling, such as SCORE, PROCAM, and Framingham, according to the European guidelines for the prevention of cardiovascular disease, not adapted for all patients and require significant human involvement in the selection of predictive variables, transformation and imputation of variables. In ROC-analysis for prediction of significant cardiovascular disease (CVD), the areas under the curve for Framingham: 0.62–0.72, for SCORE: 0.66–0.73 and for PROCAM: 0.60–0.69. To improve it, we apply for approaches to predict a CVD event rely on conventional risk factors by machine learning and deep learning models to 10-year CVD event prediction by using longitudinal electronic health record (EHR).

Methods

For machine learning, we applied logistic regression (LR) and recurrent neural networks with long short-term memory (LSTM) units as a deep learning algorithm. We extract from longitudinal EHR the following features: demographic, vital signs, diagnoses (ICD-10-cm: I21-I22.9: I61-I63.9) and medication. The problem in this step, that near 80 percent of clinical information in EHR is “unstructured” and contains errors and typos. Missing data are important for the correct training process using by deep learning & machine learning algorithm. The study cohort included patients between the ages of 21 to 75 with a dynamic observation window. In total, we got 31517 individuals in the dataset, but only 3652 individuals have all features or missing features values can be easy to impute. Among these 3652 individuals, 29.4% has a CVD, mean age 49.4 years, 68,2% female.

Evaluation

We randomly divided the dataset into a training and a test set with an 80/20 split. The LR was implemented with Python Scikit-Learn and the LSTM model was implemented with Keras using Tensorflow as the backend.

Results

We applied machine learning and deep learning models using the same features as traditional risk scale and longitudinal EHR features for CVD prediction, respectively. Machine learning model (LR) achieved an AUROC of 0.74–0.76 and deep learning (LSTM) 0.75–0.76. By using features from EHR logistic regression and deep learning models improved the AUROC to 0.78–0.79.

Conclusion

The machine learning models outperformed a traditional clinically-used predictive model for CVD risk prediction (i.e. SCORE, PROCAM, and Framingham equations). This approach was used to create a clinical decision support system (CDSS). It uses both traditional risk scales and models based on neural networks. Especially important is the fact that the system can calculate the risks of cardiovascular disease automatically

and recalculate immediately after adding new information to the EHR. The results are delivered to the user's personal account.

Digital Health: Big Data Analysis

Topic: cardiovascular diseases, framingham heart study, clinical decision support systems, demography, memory, short-term, roc curve, guidelines, patient prognosis, vital signs, electronic medical records, predictor variable, cardiovascular disease prevention, weight measurement scales, precision medicine, imputation, missing data, datasets, machine learning, mobile health, area under the roc curve, deep learning, big data
Issue Section: Sunday 1 September 2019