Electronic health records (EHRs) of patient information, collected during each clinical encounter, have been increasingly available in the last decade. This provides a basis for boosting the quality of healthcare analytics and data-based clinical decision making. Structured patient information such as demographics (e.g. age and gender), diagnoses, procedures and prescriptions can be utilized to support the prediction of negative outcomes such as stroke, bleeding and mortality in complex patients.

This study focuses on predicting health risks in patients with atrial fibrillation (AF) treated with Warfarin from a primary care database (CPRD—the UK Clinical Practice Research Datalink). In particular, we use the CPRD 2017 corpus of 26K patients, with a 80%/10%/10% train/development/test patient split. We formulate our task as a binary classification problem. To determine a label associated with each patient, we base the analysis on the start date of the AF diagnosis, i.e., $af_{start}$. In a window of 4 years after $af_{start}$, the patient is assigned a positive label if there is evidence of one of death, ischemic stroke/systemic embolism, or a bleeding event; otherwise, the patient receives a negative label.1

Each patient’s history is given as a temporal sequence of clinical consultation events taken by the patient before $af_{start}$. Here, each clinical event is represented by a list of 5-character medical readcodes. In addition to the patient’s history, the patients’ gender and age group is available.2

We compare the following machine learning methods: a neural network-based model and conventional Support Vector Machine (SVM) and Naïve Bayes (NB) models. Our neural network model is based on the long short-term memory (LSTM) architecture to model each patient’s history. In particular, gender, age group and readcodes are represented by vector embeddings. Each event in the patient’s history is represented as the average vector of all the readcode embeddings associated with the event. We apply a LSTM to the sequence of clinical event vectors, then use the “last” LSTM hidden state vector to represent the whole patient’s event history. This vector is concatenated with the corresponding gender and age group embeddings to produce a final vector representation of the patient information. The final vector is fed into a multilayer perceptron with softmax output for classification.

The SVM and NB models also use gender, age group and readcode features. We model each patient’s event history as a discrete sequence of the 5-character readcodes without considering temporal information. We calculate the value of each feature by using the TFIDF weighting scheme. We find that our LSTM-based model produces the highest F1 score w.r.t. the positive label at 49.3% which is 7.6% and 1.4% absolute higher than scores of SVM (41.7%) and NB (47.9%), respectively. This is reasonable because LSTM can capture useful temporal information in the patient’s event history.

We also find that readcode hierarchy information is useful to improve the model performance. We create another list of codes of the first 3 characters of readcodes for each history event. When additionally incorporating this list into our models, the F1 scores produced by LSTM-based, SVM and NB models for the positive label are improved to 50.7%, 43.0%, and 49.1%, respectively.

Our models with performances of around 50% are not yet suitable for direct clinical application. Future work will focus on improving these models with better learning and data representation strategies.

---

1In each of training, development and test sets, the percentage of positive labels is about 30%.

2In fact, the CPRD contains patients’ ages only. To reduce sparsity among patients’ ages, we group ages into bins of 10.

3For LSTM-based model, the event vector is created by concatenating the average vector of readcode embeddings and the average vector of 3-character code embeddings. For SVM and NB, each patient’s event history is considered as a document of the 5-character readcodes and 3-character codes.