Machine Learning Modeling to Predict the Patient Outcomes with Venous Thromboembolism

Kexin Zhang[1], Jung In Park, PhD, RN[2]
[1]Department of Statistics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, CA; [2]Sue & Bill Gross School of Nursing, University of California, Irvine, Irvine, CA

Introduction

Venous thromboembolism (VTE) is the third most common vascular diagnosis, and it is associated with a high risk of mortality and a significant cost burden. Nearly half of VTEs are related to hospitalization and surgery, and most of them do not occur until after discharge. While these incidents are preventable, existing studies have found that fewer than half of hospitalized patients receive appropriate prevention, which can lead to readmission with acute VTE. Therefore, it is essential to monitor the patient at risk for VTE after discharge to prevent readmission with VTE. Machine Learning (ML) modeling has been suggested as an effective way to accurately predict patient outcomes. This project aims to develop and validate ML models to predict the 30-day readmission with acute VTE using MIMIC-III (Medical Information Mart for Intensive Care) database.

Methods

We used the MIMIC-III (version 1.4) database, which comprises large de-identified electronic health records from a single medical center. We included the adult patients (18 or older) who were re-admitted within 30 days after discharge and whose records existed between 2001 and 2012. The following features were included for the modeling: age, ethnicity, insurance, gender, number of admissions, length of stay, readmission days (the difference between last discharge time and next admission time), diagnosis (ICD-9), procedures, and lab results. Since each hospitalization might have led to different health outcomes, we used each unique admission as the unit of analysis. Three Machine Learning models were developed: Balanced Random Forest (BRF), Logistic Regression (LR), and XGBoost. We applied Synthetic Minority Oversampling Technique (SMOTE) to counter the highly imbalanced classes by increasing the number of minority samples and used the new training set for Logistic Regression. Three factors were used to evaluate the performance of three models: sensitivity, specificity, and AUC. We adjusted the values in parameters using different regularization techniques, such as limiting the max number of tree depths, to balance the differences between sensitivity and specificity to achieve the best performance.

Results

Table 1. Prediction Model Evaluation

	AUC	Sensitivity	Specificity
Balanced Random Forest	0.87	0.92	0.81
Logistic Regression	0.68	0.42	0.95
XGBoost	0.90	0.92	0.87

Two classes were identified among the 4,630 unique readmissions: readmission with VTE (n_1=186) and without VTE (n_2=4,444). We divided the dataset into two sets: 75% of the data for the training set and 25% for the test set. Based on the evaluation results, XGBoost model showed the highest performance (AUC: 0.90, Sensitivity: 0.92, Specificity: 0.87), followed by BRF model (AUC: 0.87, Sensitivity: 0.92, Specificity: 0.81), and LR model (AUC: 0.68, sensitivity: 0.42, specificity: 0.95). Although the LR model showed the highest specificity, other measures were lower than the other two models. We utilized XGBoost models to output the top ten significant features in the study, which were age, the number of admissions, length of stay, readmission days, comorbidities such as complications of surgical and medication care, and Diseases of Esophagus, Stomach, and Duodenum, and some procedures related with incision, excision, and occlusion of vessels.

Conclusion

Using the MIMIC-III database, we developed accurate ML models and identified relevant features to predict the individualized risk of readmission with acute VTE within 30 days of discharge. The ML model developed in this study can be used as a tool to better predict patient outcomes with VTE for improved preventive care.