Title: Prediction of Post-Stroke AF in ESUS Patients is Enhanced by Combining Expert-Derived Predictors and Embedding of Full Diagnostic Codes using Pre-Trained Hypergraph Neural Networks

Yuhua Wu[1], Fadi B Nahab[1], Yi Ge[2], Yuzhang Xie[1,2], Hassan O Aboul-Nour[3], Carl Yang[1], and Xiao Hu[1]

[1]Emory University, Atlanta, GA, United States
[2]University of California, Berkeley, University Avenue and, Oxford St, Berkeley, CA, United States.
[3]Neurology and Neurosurgery, University of Kentucky, Lexington, KY, United States.
[*]Address correspondence to: Xiao.hu@emory.edu

1 Introduction

Atrial Fibrillation (AF) occurs in about one-fourth of patients with embolic stroke of unknown source (ESUS). Accurate prediction of post-stroke AF upon discharge from an index stroke admission informs a personalized post-stroke monitoring strategy of AF and interventions. While clinical risk scores predict AF, machine learning (ML) models have shown superior performance. However, traditional ML approaches only use expert-derived predictors available in an electronic health record (EHR) and thus may miss variables that would potentially increase the accuracy of prediction.

2 Aims

This study aims to enhance AF prediction by augmenting expert-derived predictors with an unbiased selection of full diagnostic codes and medication histories up to index strokes. Through embedding learning with hypergraph neural networks, we generate compact representations of high-dimensional data to improve prediction accuracy by capturing complex feature interactions.

3 Methods

We analyzed data from 510 ESUS patients (55.3% female, mean age 61.4 years) from 2015 to 2023 at Emory Healthcare. We focus on experiments using a logistic regression (LR) model to predict AF from different sets of features. At baseline, we use 58 clinically motivated predictors, including comorbidities characterized by 17 ICD codes manually extracted based on literature, and 41 other features extracted from lab results, echocardiographic and ECG. To directly model the full history of comorbidities and medications, another baseline uses the full 1530 ICD codes plus the 41 other features (1571 in total). In contrast, the embedding method uses the full 1530 ICD codes to generate condensed, informative embedding vectors (32- dimensional), eventually getting 32+41=73 features. To generate the embedding, a hypergraph neural network was trained on a larger stroke cohort (n=7956) to model the interactions between the 1530 ICD codes. A nested cross-validation approach was employed within 5-fold splits, and ROC-AUC scores were recorded.

4 Results

Among 510 ESUS patients, 107 (21.0%) developed AF (mean age 67.9 years, 57% female). We compared the performance of LR model with different features from ICD codes (Table 1). The results show that the learned 32-dim embedding vectors improves the prediction of post-ESUS AF.

Table 1: Performance of Embedding Methods for Post-ESUS AF Prediction

Features	**Number of Features**	**AU ROC**
Baseline models	58	0.580 ± 0.034
1530 ICD codes	1571	0.606 ± 0.041
32-dim Embedding	73	0.705 ± 0.042

5 Conclusion

The embedding technique can significantly enhance predictive performance by integrating comprehensive medical information, maximizing the use of available data for improved outcomes.