A Comparison of Stroke Classifiers Leveraging Hospital Billing Codes versus Natural Language Processing

Danielle L. Mowery PhD MS1,2, Brent Hill PhD2, Mingyuan Zhang MS3, Wendy W. Chapman PhD1,2, Lisa Cannon-Albright PhD MStat4, Jennifer J. Majersik MD MS5,6

1IDEAS Center George E. Wahlen Veteran Affairs Medical Center; 2University of Utah HealthCare; 3Division of Biomedical Informatics, University of Utah; 4Division of Genetic Epidemiology, Department of Internal Medicine, Division of Vascular Neurology, Department of Neurology, University of Utah, Salt Lake City, UT

Introduction

Objective: Develop an NLP-based stroke classifier that accurately predicts stroke and its types by encoding and reasoning with predictors from the clinical notes within the electronic health record (EHR).

In large-scale epidemiological studies, ICD-9 billing codes can misclassify many non-stroke patients as stroke cases resulting in varied precisions (ischemic: 66-95% and hemorrhagic: 65-89%)1. More precise classification might be achieved using information extracted from the clinical notes using natural language processing (NLP) and machine learning (ML). Therefore, we developed an NLP-based stroke classifier and compared its performance to an ICD-9 billing code-based stroke classifier, both using ML. This pilot study is a first step toward NLP-based stroke phenotyping to support large-scale pedigree studies leveraging EHR data and the Utah Population Database.

Methods

Dataset: We randomly sampled 400 patients hospitalized in the Neurocritical Care Unit at the University of Utah Hospital from 2008–2015 with hand- adjudicated discharge diagnoses: 200 ischemic stroke, 100 hemorrhagic stroke (non-traumatic intracranial and subarachnoid), and 100 other acute neurological disorders. We then queried EPIC EHR data for corresponding ICD-9 billing codes and radiology notes.

Approach: We developed an NLP approach: remove stop words, encode n-grams (n=1-4; e.g., 2-gram = “intraventricular hemorrhage”), select most informative n-grams using Chi-square feature selection (top 1% n-grams ranked by p-value association to each class), and train a linear support vector machine using the most informative n-grams and 5-fold cross validation. For each class, we compared the average precision produced by an NLP-based stroke classifier versus ICD-9 billing code-based stroke classifier. We used both primary and secondary ICD-9 billing codes.

Results

In Figure 1, we depict a word cloud of the most informative n-grams for discerning hemorrhagic stroke from ischemic stroke and other neurological disorders. The most informative n-grams describe symptoms (“headache”), signs (“hypertensive”), radiological findings (“hematoma”), diagnoses (“intraparenchymal hemorrhage”), procedures (“ventriculostomy”), and medications (“nimodipine”) among other information types. In Figure 2, compared to the ICD-9 billing code-based stroke classifier, the NLP stroke classifier produced comparable recall (not shown; all classes > 95.0) and promising precision: ischemic stroke: 70.1, hemorrhagic stroke: 85.3, and other neurological disorders: 81.2.

Figure 1. Hemorrhagic stroke word cloud of n-grams. The n-gram p-values were transformed using –log 10. Therefore, the font size is scaled to the p-value representing the association to a hemorrhagic stroke diagnosis.

Figure 2. Precisions according to approach: ICD-9 billing codes and NLP

Conclusion

- NLP stroke classifier produces precisions within known performances of epidemiological studies1
- A more complex approach is required to produce comparable ICD-9 precisions, which performed better than anticipated.
- ICD-9-based precisions are not likely to precisely identify pre-EPIC stroke cases locally and stroke cases at other institutions
- We are enhancing our NLP stroke classifier to leverage features from a knowledge-engineered stroke knowledge base including stroke-related concepts, their synonyms, and their relationships2

Acknowledgements/References

Study Funded By: Neuroscience Initiative (NI) Collaborative Pilot Project Grant. This project has received ethical approval from the University of Utah Institute Review Board.

Correspondences: danielle.mowery@utah.edu

References
