A comparison of machine learning methods for automated EHR phenotyping of stroke patients

Phyllis M. Thangaraj, MPHi,1,2,3, Joseph D. Romano, MPHi,2, Fernanda Polubriaginof, MD MA2, Benjamin Kummer, MD2,5, Nicholas Giangreco, BS1,2,3, Mitchell Elkind, MD, Ph.D, Nicholas P. Tatonetti, PhD, MD1,2,3

Columbia University, New York, NY, USA. 1Integrated Program in Cellular, Molecular, and Biomedical Studies, 2Department of Biomedical Informatics, 3Department of Systems Biology, 4Department of Physiology, 5Department of Neurology

Background

- Stroke is a leading cause of death and disability in the world. (1)
- Accurate identification of stroke is essential for research studies
- Phenotyping from the EHR can be difficult, time consuming, and incomplete (Figure 1)
- We apply machine learning algorithms to automatically identify acute ischemic stroke patients

Objectives

Goal: Systematic method of stroke identification by comparison of models

MODEL DEVELOPMENT
- Build integrated phenotyping system
- Reduce size of Gold Standard required
- Reduce feature size
- Explore robustness of models

APPLICATIONS OF MODELS
- Refine Gold Standard
- Visualize the cohorts identified
- Determine whether probabilities can be used to identify cohorts for genotyping

Data

- Columbia University Medical Center Clinical Data Warehouse
- Patients included from 1985-2016
- In-patient admission data

Methods

- Columbia University Medical Center Clinical Data Warehouse (~5.3 million patients)
- Gold Standard (450k)
- No Stroke (2.6 million)
- No CV (2.3 million)

Results

PHENOTYPE CLASSIFICATION 1: STROKE VS. NOT STROKE

- Accuracy: LR: 0.75-0.99, RF: 0.74-0.98, AB: 0.86-0.99
- Metrics of the model show good performance for identifying acute ischemic stroke patients (Figure 3)
- Model-derived stroke cohorts were defined by probability scores in the test set above the probability threshold at the maximum F1-score

PHENOTYPE CLASSIFICATION 2: STROKE VS. OTHER CV DISEASE

- Accuracy: LR: 0.75-0.98, RF: 0.62-0.95, AB: 0.74-0.98
- Models with CV disease but not stroke as controls performed the best precision and recall

REDUCTION OF FEATURES BY HIERARCHICAL COLLAPSING

- Feature reduction from ~50000 features to ~50000 features depending on hierarchy level
- Refinement of Gold Standard test patients with the lowest probability scores, 53% were false negatives
- This suggests that the Gold Standard can be curated by our models.

Conclusions

- Different cohorts of stroke patients can be highlighted by different models
- The gold standard could potentially be refined by model probability stratification
- Topological clustering can visualize feature enrichment correlation with multi-dimensional model probability

Future Plans

- Evaluate performance of models with reduced features by hierarchical collapsing
- Evaluate robustness of models
- Explore correlations between model probabilities and severity of stroke, and use as quantitative trait proxy in genetic studies

References

4. The gold standard for stroke can be achieved by different methods.
5. The gold standard could potentially be refined by model probability stratification.