A Collaborative Approach to Phenotype Development: Framing Conceptset Discovery as a Machine Learning Problem

Amelia J. Averitt[1], Shreyas A. Bhave[1], Kai Chen[1], RuiJun Chen[1], Noemie Elhadad[1], Thomas Falconer[1], George M. Hripcsak[1], Xinzhuo Jiang[1], Krishna S. Kalluri[1], Andrew S. Kanter[1], Junghwan Lee[1], Cong Liu[1], Anna Ostropolets[1], Chao Pang[1], Adler Perotte[1], Victor Rodriguez[1], James R. Rogers[1], Fabricio Sampaio Peres Kury[1], Ning Shang[1], Matthew E. Spotnitz[1], Tony Y. Sun[1], Casey Ta[1], Liana Tascau[1], Nicholas P. Tatonetti[1], Phyllis M. Thangaraj[1], Runsheng Wang[1], Chunhua Weng[1], Karthik Natarajan[1], Patrick B. Ryan[1]

[1] Columbia University Medical Center, New York, NY

Background

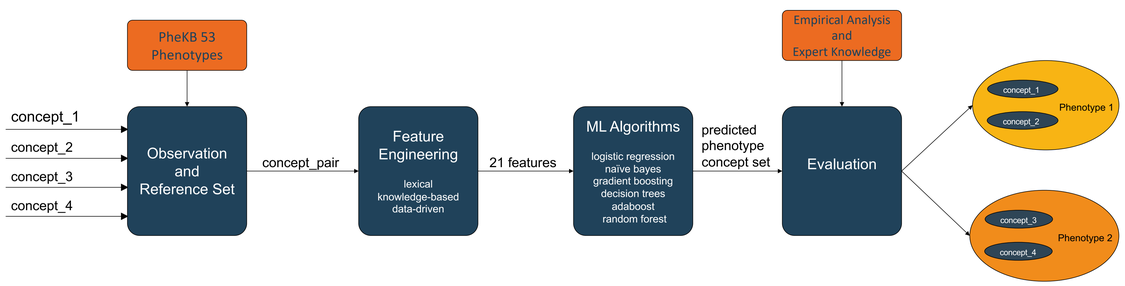
Phenotype development and evaluation remains a challenge for the entire research community. Despite substantial progress on bespoke solutions for individual disease states, there is still a major gap that limits the progress of generating reliable evidence from observational healthcare data: a comprehensive phenotype library that contains human-readable and computer-executable logic to identify cohorts of persons that satisfy one or more inclusion criteria for a duration of time, with characterizations and evaluations that provide context for how the cohort definitions perform against a network of databases, which can be re-used to instantiate populations for use in observational analyses. The OHDSI Phenotype Library workgroup is developing an infrastructure to allow storage and maintenance of such phenotype entries. The Columbia Department of Biomedical Informatics (DBMI) Phenotype workgroup aims to contribute to the OHDSI Phenotype Library efforts by developing an automated solution for phenotype development that can be scaled to produce reliable and validated cohort definitions.

Methods

We sought to develop machine learning algorithms that can address the problem of 'conceptset construction', by answering the **question**:

"Which structured concepts (in any domain: condition/procedure/drug/measurement) belong together as a group when trying to define an inclusion criteria within a rule-based phenotype definition, or a candidate feature in a probabilistic-based phenotype?"

To frame this as a machine learning algorithm, we defined 4 key groups: Unit of observation and reference set, feature engineering, algorithms and evaluation. Members of the phenotype working group joined each of these four subgroups based on interest and expertise.



Contact: kn2174@cumc.columbia.edu

Subgroups

- **Unit of observation and reference set**: We generated the reference set from 53 validated PheKB phenotypes by extracting disease, procedure and measurement codes and grouping them based on the clinical entities that these phenotypes contained. We then made positive and negative controls and removed the concepts that belong to the same hierarchical tree in OMOP vocabularies.
- **Feature engineering**: Using the reference set with positive and negative controls, this subgroup generated features that can be further used in learning algorithms. Three broad groups of features were created: *lexical, knowledge-based, and data-driven features.*
- **Algorithms**: We modeled features with the Python scikit-learn package including logistic regression, naïve bayes, gradient boosting, decision trees, adaboost and random forest. Models were evaluated on random and phenotype-aware training and test sets.
- **Evaluation**: The evaluation subgroup, comprised of clinical experts, will develop methods to validate phenotypes by a combination of empirical analysis and expert knowledge.

Conclusions

This first study by the Columbia DBMI phenotype working group was accomplished in a month and demonstrates the impact a department-wide collaboration can yield. The development of machine learning models have the potential to provide a scalable and fully automated method for creating an entire library of conceptsets to support phenotype development. Additionally, the resulting model can be used as part of a recommender system to guide researchers in the concept set creation process, or could be applied to construct 'first-order' cohort definitions that can be characterized and evaluated across the OHDSI network. While preliminary results are promising, further collaborative work to refine the reference set, expand the feature engineering, train additional algorithms, and facilitate further clinical evaluation may yield additional value to the research community.

Take a picture to download the full paper