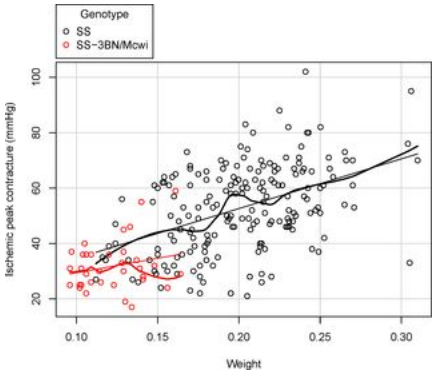
High-throughput multimodal automated phenotyping (MAP) with application to PheWAS

Department of Statistical Sciences
University of Toronto

SHAOHAN CHANG and TIMOTHY REGIS

November 30 , 2022

Introduction and Motivation



What is the problem being solved?

How to develop an algorithm which can phenotype patients accurately and efficiently?

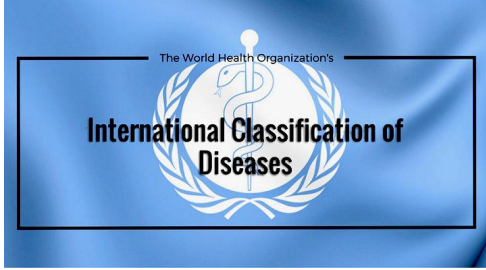
Why is it an important problem?

- High-throughput technologies is widely applied to dissect genomic and biologic architecture of complex human traits.

- Translation of 'omics' finding to improvement in patient care, key link to the high- throughput biologic data with detailed high-quality phenotypic data (challenging problem).

- Efficient and accurate algorithm to phenotyping is in urgent need.

What Previous Work Exists?



Approaches using International Classification of Diseases (ICD) codes only.

1. PheWAS approach.

- loss of power for association.

Approaches combining ICD codes and narrative features identified via natural language processing (NLP).

1. Rule-based manual curation

2. Machine learning-based supervised learning

- Labor intensive and limiting the feasibility of using

Approaches to improve the efficiency for developing phenotyping algorithms.(Reduce the level of human input required)

1. Active Sampling

2. Feature Refinement

- Curate silver standard labels or have variable accuracy.

Approaches that are fully automated.

1. Phenotyping algorithm based on normal mixture modeling using ICD codes only.

- Not provide the final classification of whether a participant had a specific phenotype necessary for clinical studies.

Methods

What methods were used

MAP procedure:

- ICD codes and NLP features corresponding to the target phenotype
- Annotation via unsupervised ensemble latent mixture modeling.

Identify the ICD code corresponding to the target phenotype

- **Defined by the investigator** or by **selecting a PheWAS code from the catalog** to use the associated ICD mapping.
- ICD feature.

Identify the NLP corresponding to the target phenotype

- **Main ICD feature** is extracted from the data set.
- Identify the medical concepts by **mapping** relevant clinical terms to the CUIs listed in the Unified Medical Language System
- **Combining all CUIs** mapped in the 3 steps gives a list of CUIs to represent each phecode.

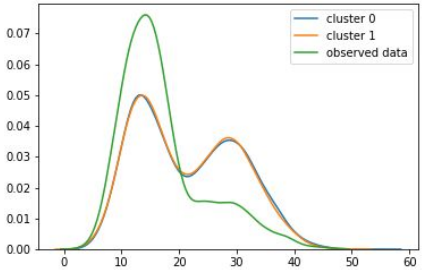
Methods

Annotation via unsupervised ensemble latent mixture modeling.

$$P(X_{\text{count}} = x) = \theta \frac{\left(\alpha \text{Note}_{\log} + \lambda_1\right)^x e^{-(\alpha \text{Note}_{\log} + \lambda_1)}}{x!} + (1-\theta) \frac{\left(\alpha \text{Note}_{\log} + \lambda_0\right)^x e^{-\left(\alpha \text{Note}_{\log} + \lambda_0\right)}}{x!},$$

- **Estimate the parameters in the model by EM algorithm.**
- **The prevalence estimate is then used as a threshold to assign a binary classification of whether a participant has the phenotype.**

Methods



What are the key assumptions of the methods?

- Poisson mixture model

How are the methods different from previous work?

- NLP with ICD codes for use in a high-throughput phenotype algorithm pipeline
- MAP is the ability to assign a threshold to provide yes/no classification for each phenotype.

Why is this approach a logical one to take ?

- ICD codes and narrative features identified via natural language processing (NLP) enhance the accuracy of phenotyping.
- Both ICD and NLP features automatedly, so the algorithm across institutions will not be labor intensive.

Results

Data:

EHR and Genetic Data Collected from 2 sources:

- Partners Biobank
  - Bio-data repository from Boston
  - 17,805 patients; 16 phenotypes

- Veteran's Affairs Million Veteran's Program (VA MVP)
  - Longitudinal cohort study on military exposure and genetics
  - 330,374 patients; PheWAS on 17 phenotypes

Assumptions:

Key distributional assumptions:

- Poisson Mixture Model
- Quasi-binomial model
- Log-normal model
    - All easily checkable

Not included in paper!

Accomplishments:

Primary Objectives:

- Develop an automated high-throughput phenotyping method
  - Focusing on PheWAS
- Develop a process for generating a binary classification of a phenotype's presence

- Success determined through comparisons to current methods in use

Comparisons

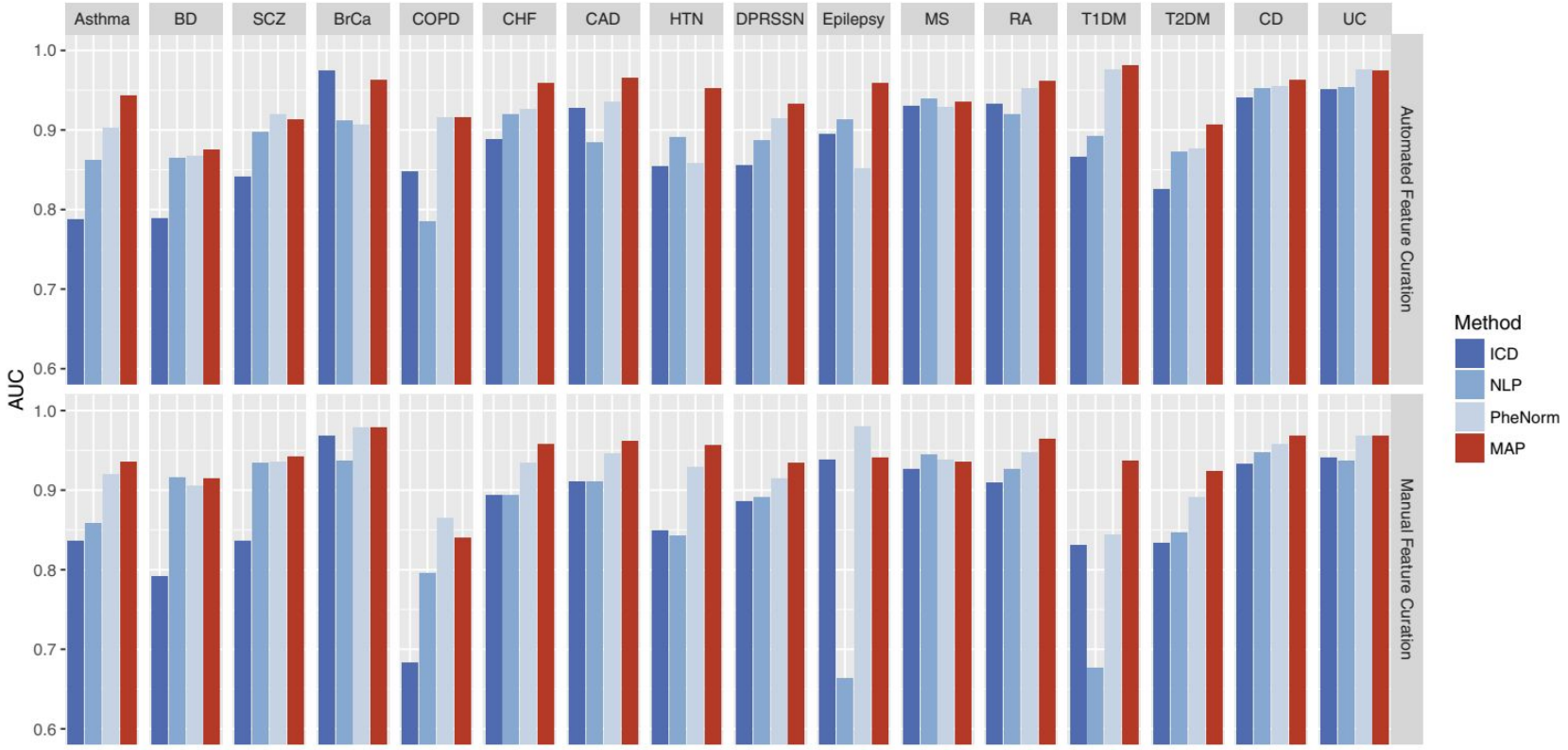


Figure 1. Top panel: Comparison of AUCs with gold standard labels for ICD-9 count, NLP, PheNorm, and MAP for 16 disease phenotypes using the MAP automated feature curation. Bottom panel: Comparison of AUCs for the 16 phenotypes features manually curated by domain experts.

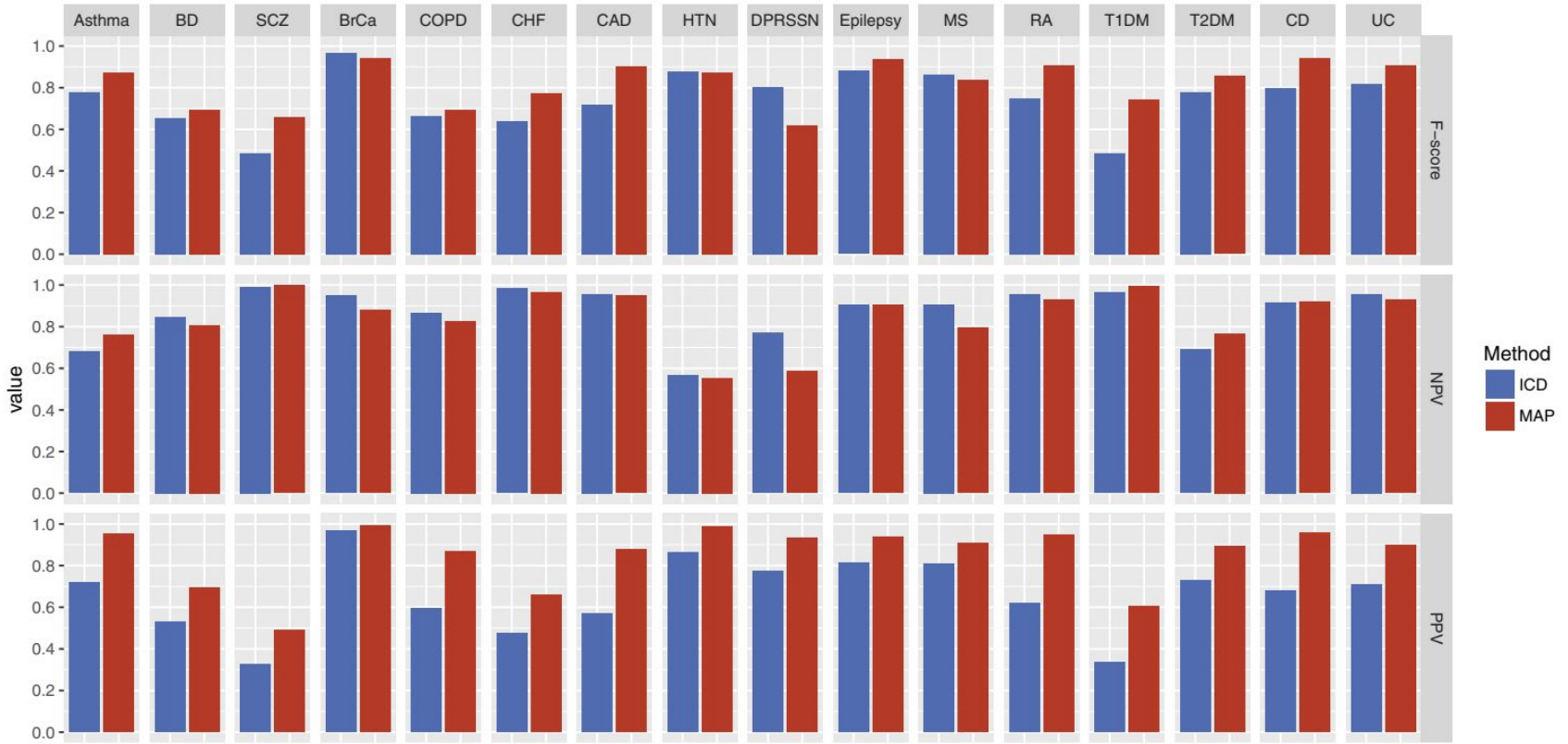


Figure 2. Performance of phenotype classification using MAP compared to ICD-9 codes for 16 phenotypes with gold-standard labels. (F-score, negative predictive value [NPV], positive predictive value [PPV, precision]).

phenotype	phecode	nCase (nControl)	OR (P-value)	
			MAP	ICD2
Ischemic Heart Disease	411	83772 (246602)	0.945 (3.89e-19)	0.955 (3.59e-12)
Coronary atherosclerosis	411_4	67720 (262654)	0.948 (2.27e-16)	0.952 (3.58e-12)
Abdominal aortic aneurysm	442_11	7894 (322480)	0.865 (3.29e-16)	0.868 (2.86e-16)
Aortic aneurysm	442_1	9946 (320428)	0.885 (6.7e-14)	0.895 (6.72e-13)
Other aneurysm	442	11349 (319025)	0.908 (4.21e-10)	0.911 (1.93e-10)
unspecified	411_8	49191 (281183)	0.961 (1.45e-07)	0.951 (8.4e-11)
Atopic/contact dermatitis due to other or unspecified	939	39400 (290974)	1.03 (2.24e-06)	1.03 (0.000869)
Peripheral vascular disease, unspecified	443_9	21438 (308936)	0.946 (7.89e-06)	0.953 (1.43e-05)
Rash and other nonspecific skin eruption	687_1	21994 (308380)	1.04 (8.9e-06)	1.02 (0.0483)
Chronic liver disease and cirrhosis	571	12195 (318179)	1.08 (1.38e-05)	1.05 (0.000256)
Atherosclerosis of native arteries of the extremities with intermittent claudication	440_22	5173 (325201)	0.912 (2.8e-05)	0.901 (1.54e-06)
Peripheral vascular disease	443	22944 (307430)	0.952 (3.22e-05)	0.958 (6.32e-05)
End stage renal disease	585_32	3990 (326384)	1.12 (3.29e-05)	1.06 (0.0196)
Myocardial infarction	411_2	18241 (312133)	0.953 (5.75e-05)	0.946 (2.22e-06)
Atherosclerosis of the extremities	440_2	7768 (322606)	0.932 (0.000367)	0.924 (9.98e-06)
Other mental disorder	306	26794 (303580)	1.03 (0.00095)	1.05 (7.34e-06)
Atherosclerosis	440	10193 (320181)	0.959 (0.01)	0.937 (3.04e-05)

Supplementary Table 3: PheWAS of IL6R (rs2228145) using phenotypes defined using the MAP and the standard approach (ICD), where either the MAP p value or the ICD p value passes the Bonferroni threshold of 3.73x10E-5 adjusting for 1342 tests. Reported also were the number of cases and controls for each phenotype

Conclusion

Key Summary:

3 Main Contributions:

- An automated method for selecting specific NLP features for a phenotype
- An automated method for combining NLP features and ICD codes relating to a phenotype
- And a method for generating a binary classification of a phenotype's presence

Primary Takeaway:

- MAP algorithm is "more efficient, robust, and accurate," than competing methods, and can provide a significant improvement to high-throughput phenotyping

Confidence:

Results are **highly convincing**

- Achieved general all-around significance
- Outperformed all other options in almost every case
- High-degree of robustness benefits new data

Validation

- Compared Full model to simplified versions
    - Full model still outperforms

Contributions:

MAP algorithm provided multiple unique and novel discoveries, as well as improvements.

Two previous papers:

- Yu et al 2015 - **AFEP**
  - Failure: Large-scale data
- Yu et al 2017 - **PheNorm**
  - Failure: Robustness and classification

PheWAS Results:

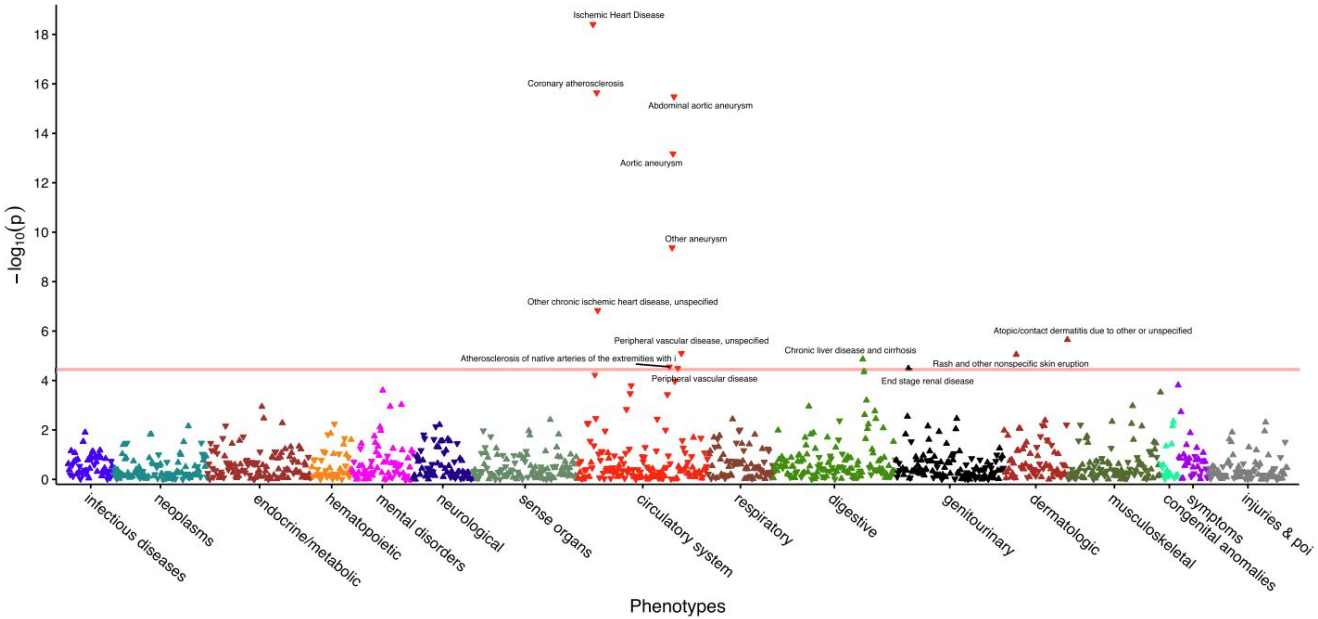


Figure 3: PheWAS results using MAP-defined phenotypes for the IL6R SNP. Phenotypes significantly associated with IL6R after Bonferroni correction are annotated.