Computational Phenotyping with Limited Data and Winning the Partners Healthcare Biobank Challenge

Prithwish Chakraborty[1*] , Michal Ozery-Flato[2*] , Kristen Severson[1*] , Eryu Xia[3*], Mohamed Ghalwash[1], Eleftheria Pissadaki[4], Jing Mei[3], Fei Wang[5], Kenney Ng[1], Amar Das[1], Jianying Hu[1], and Daby Sow[1]

IBM **Research**

[1]IBM Research, USA; [2]IBM Research, Israel; [3]IBM Research, China; *Authors contributed equally

Overview

Computational Phenotyping of Disease in real-world setting

- Data for diseases often limited
  - clinical annotations are expensive and time-consuming
  - significant amount of data not annotated but of potential use

IBM Research placed 1st (tied) in the first Partners Healthcare Biobank Disease Challenge

- Open challenge (50 teams from industry and academia)
- 50 teams from industry and academia entered
- Evaluated both on quantitative performance (33%) and other qualitative measures such as interpretability and visualizations

Data Availability

The challenge leveraged real patient data

- Partners Healthcare biobank contains information from 80K patients and includes electronic health records (EHRs) and health survey information
- EHRs contain information related to diagnosis, lab tests, medications, procedures, and vital signs

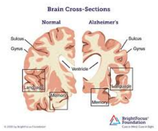
Motivation

Robust phenotyping is important for many studies

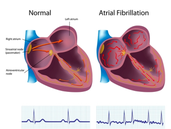
- ICD codes are noisy indicators of disease states
- Robust phenotyping is needed for problems such as claim processing, prognostic models, and observational studies
- Manual phenotyping is expensive – medical experts need 30 min – 6 hr per patient.

Diseases of Interest

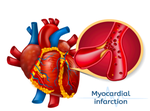
Alzheimer's Disease



Atrial Fibrillation



Myocardial Infarction



Migraine Headaches

Phenotype	# Concept Codes	# Patients	# Positive Labels	# Negative Labels
AD	18	2,369	15	60
AFIB	13	10,894	52	23
MI	85	8,360	34	41
MHA	125	12,721	56	19

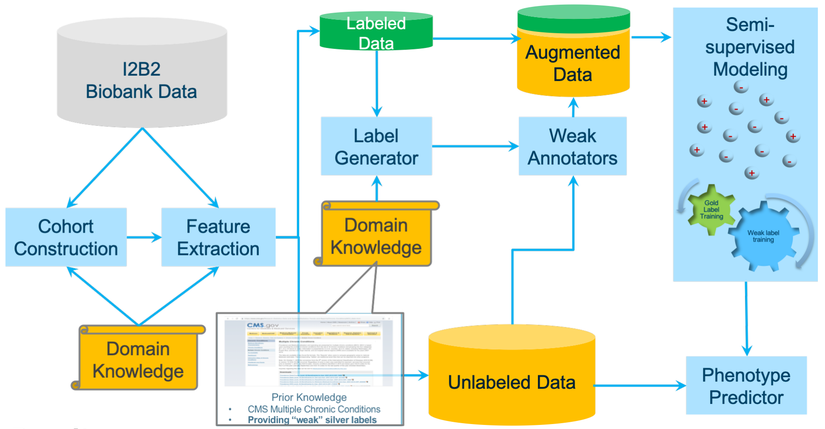
Part 1: Computed Phenotypes

Key Challenges:

- Each disease had only 75 labeled patients
- Challenge ran for four weeks
- Computational resources were set to 4vCPU, 92 GB of RAM and 1 TB of shared space per team

Approach Overview:

- Feature engineering leveraging domain knowledge
- Generation of weakly labeled samples
- Semi-supervised learning algorithm

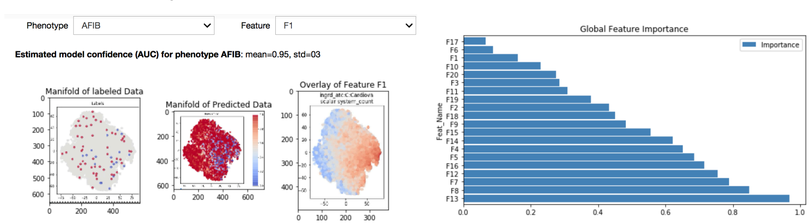


Results:

- Estimated AUC using k-fold cross validation analysis

Phenotype	Estimated AUC
AD	0.960 ± 0.003
AFIB	0.917 ± 0.012
MI	0.873 ± 0.016
MHA	0.895 ± 0.014

- Interpretability via interactive visualizations and analysis of feature importance



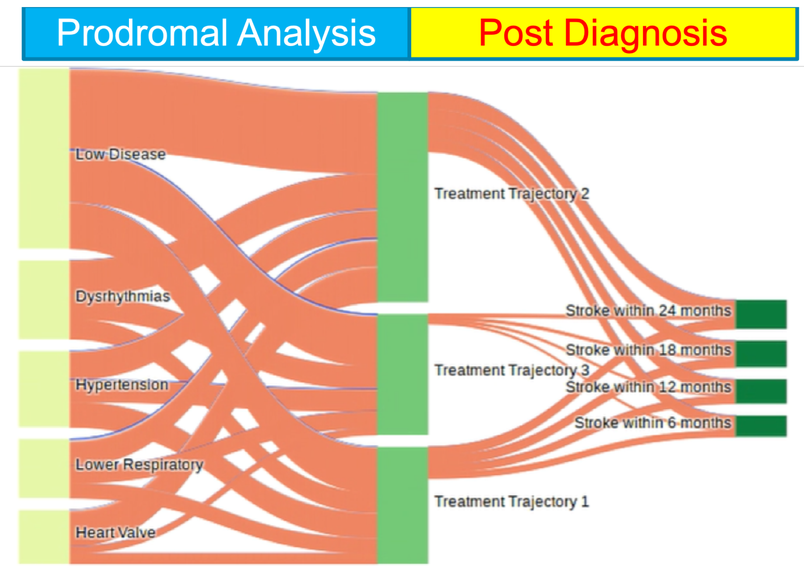
Part 2: Patterns of Anticoagulant Use in Atrial Fibrillation

Trajectory of prescription prevalence may offer insight into patient outcomes

- AFIB is associated with a 3- to 5-fold increase in stroke
- Warfarin and other anticoagulants are long-term medications which decrease risk of ischemic stroke in AFIB patients
- Prevalence of ischemic stroke admission varies between anticoagulant treatment trajectories
- Prevalence of chronic conditions also varies between treatment trajectories

Prodromal analysis may reveal signs of disease prior to first AFIB inpatient visit

- Diagnosis codes were used to perform clustering prior to first AFIB in-patient visit
- Five discovered clusters have varying representation in the treatment trajectories
- Interactive visualization allows user-driven exploration



Watch the presentation here