NEW YORK GENOME CENTER®

COLUMBIA
COLUMBIA UNIVERSITY
DEPARTMENT OF
BIOMEDICAL INFORMATICS

Multi-domain rule-based phenotyping algorithms enable improved GWAS signal

Abigail Newbury, Ahmed Elhussein & Gamze Gürsoy

npj Digit. Med. **8**, 499 (2025).
https://doi.org/10.1038/s41746-025-01815-8

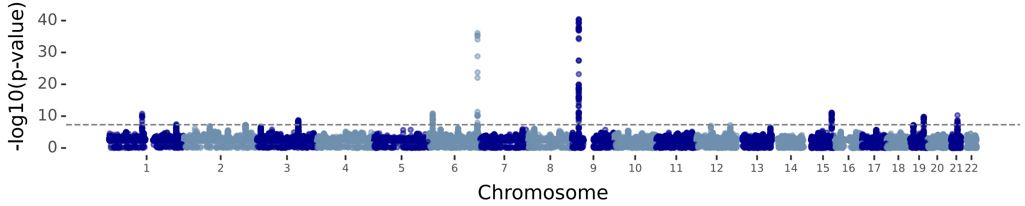
Genome-wide association studies link genomics and health data

Simple GWAS

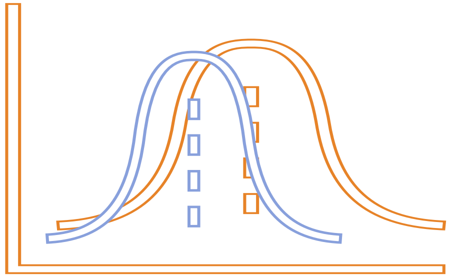
$$\boxed{y \sim \alpha + G\beta + Z\gamma + \epsilon}$$

Covariates Z: age, sex, principal components of G (represent genetic ancestry)

Identify associated SNPs



Predict individual disease risk



Case vs. control PRS distribution

[Uffelman et al. 2021]

Phenotype misclassification affects GWAS results

In a simple regression of a binary phenotype on a binary risk factor, we observe the following effect size:

$$\hat{\beta}_{obs} = (PPV + NPV - 1)\hat{\beta}_{true}$$

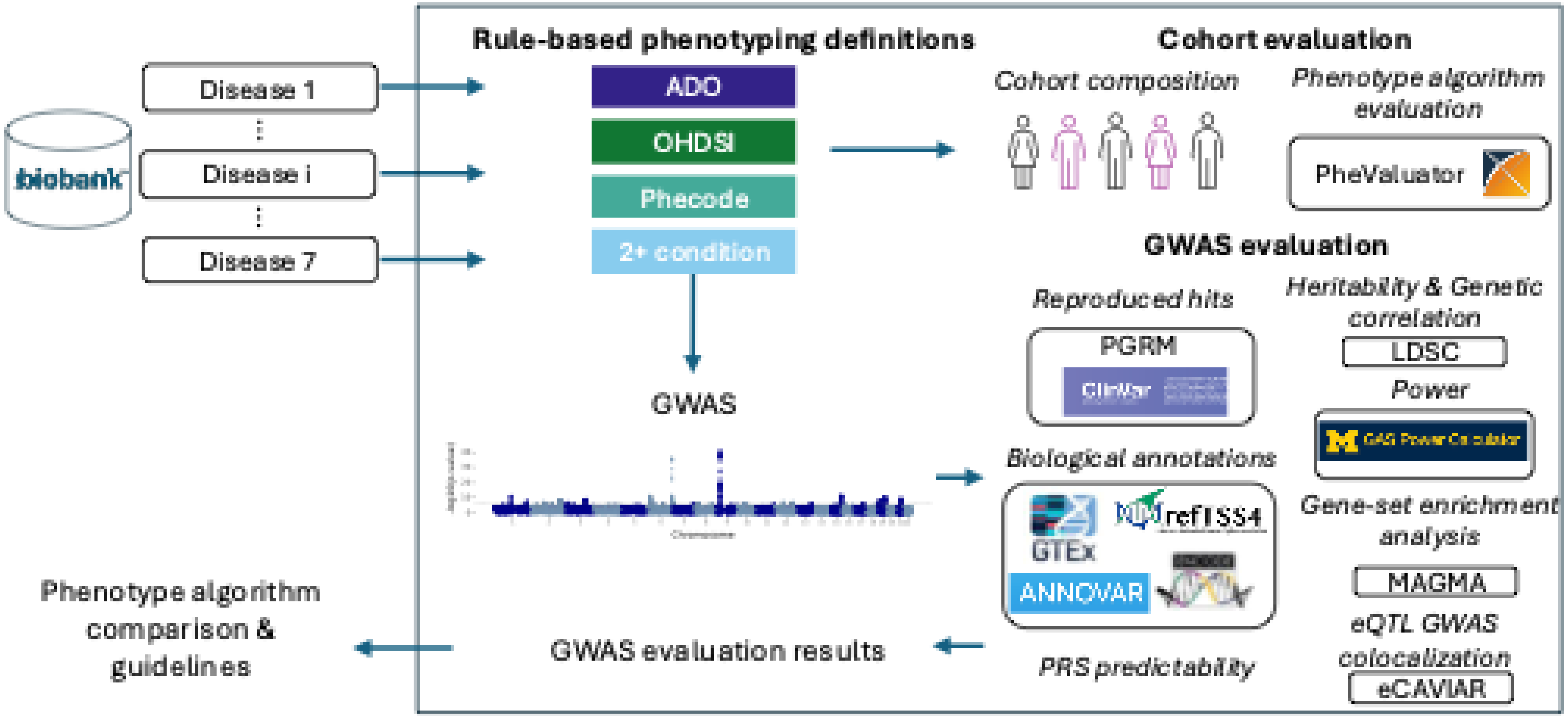
In regression models including covariates, it has been shown that

- Assuming perfect specificity, **effect sizes bias towards null** with decreased sensitivity (increase in false negatives)
- With imperfect specificity (increase in false positives), **effect sizes bias either towards or away from the null**

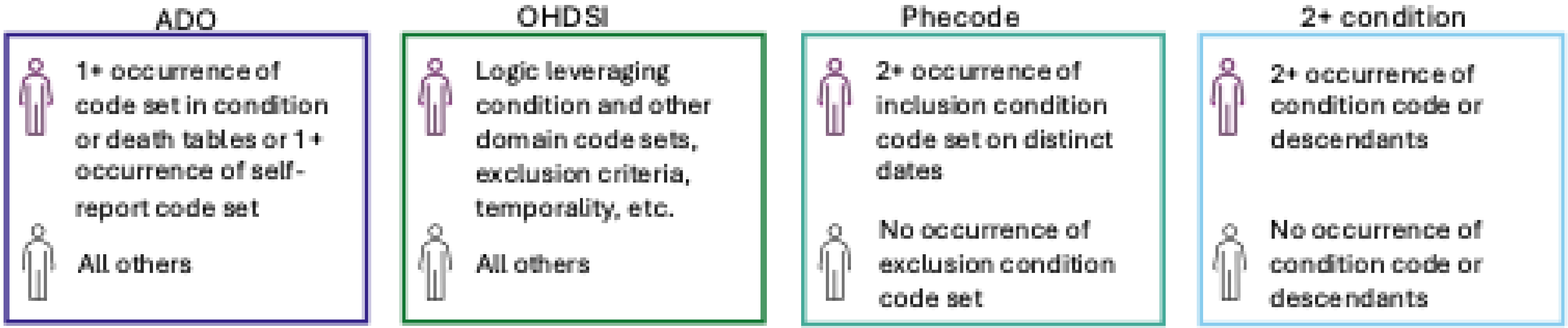
Accurate EHR phenotyping reduces Type I and II errors

[Duffy et al. 2004, Beesley et al. 2020]

Goal: to assess the **impact of various rule-based phenotyping algorithms on GWAS outcomes**, examining factors such as power, heritability, replicability, functional annotations, and polygenic risk score prediction accuracy.



Rule-based phenotyping algorithms with varying levels of complexity



ADO

1+ occurrence of code set in condition or death tables or 1+ occurrence of self-report code set

All others

OHDSI

Logic leveraging condition and other domain code sets, exclusion criteria, temporality, etc.

All others

Phecode

2+ occurrence of inclusion condition code set on distinct dates

No occurrence of exclusion condition code set

2+ condition

2+ occurrence of condition code or descendants

No occurrence of condition code or descendants

high complexity

high & **medium** complexity

medium complexity

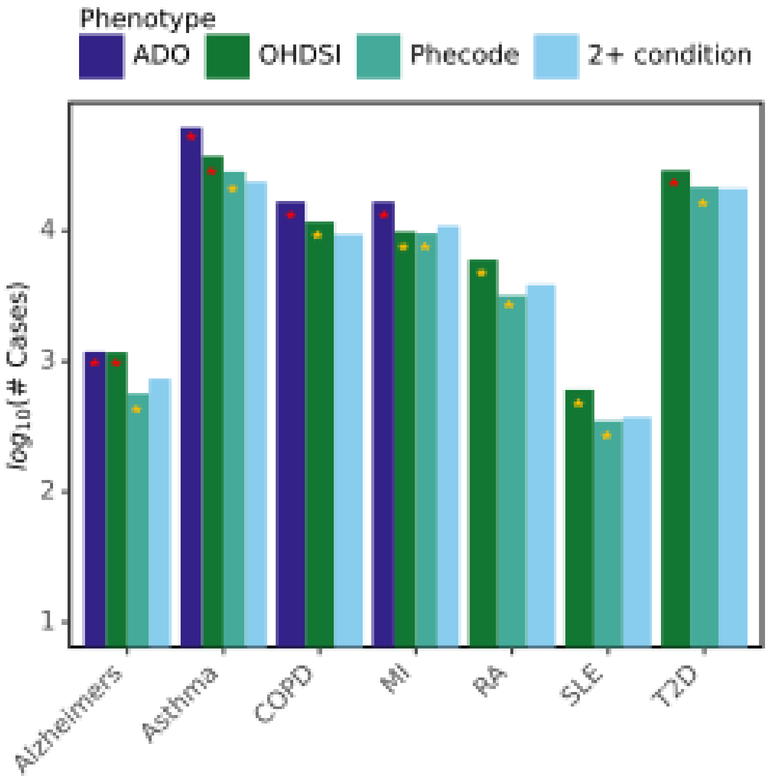
low complexity

High-complexity algorithms rely on a more **diverse set of data domains** to identify cases for cohort entry

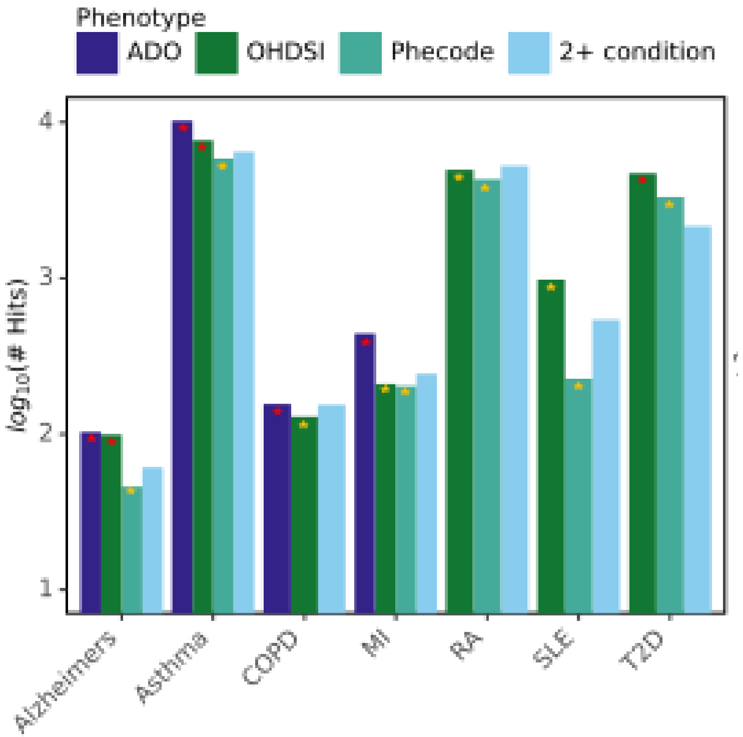
Algorithms found similar condition concepts in top index events

High complexity EHR phenotyping rules result in increased GWAS power

Number of cases by algorithm & disease



Number of GWAS hits by algorithm & disease

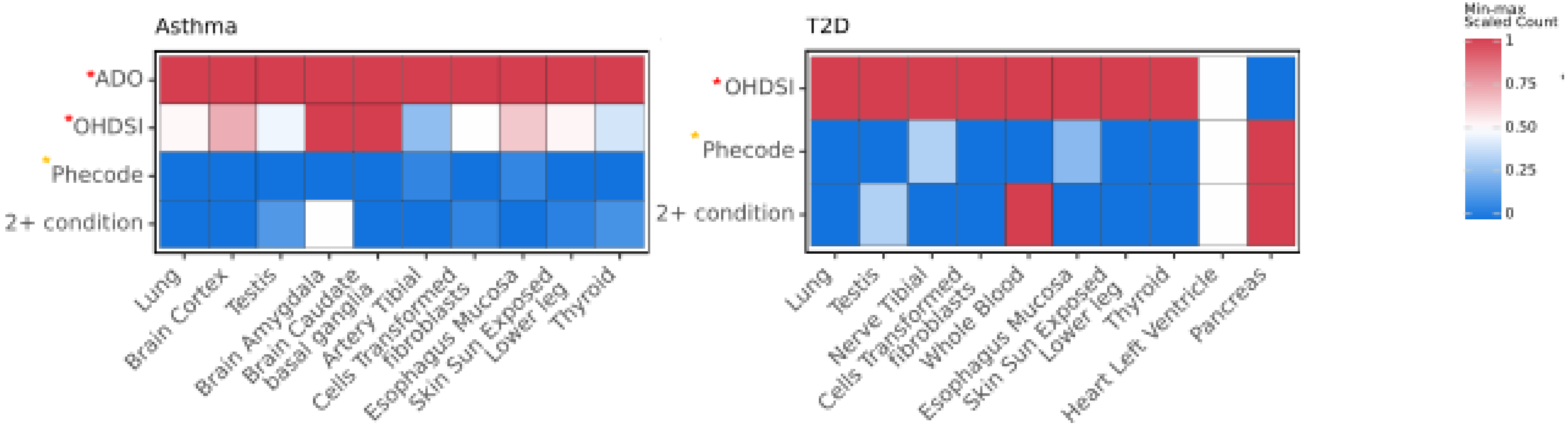


Cohorts created with the high complexity algorithms had the **highest number of cases**

High complexity algorithms generally found a **greater number of GWAS hits**

High complexity EHR phenotyping rules result in an increased number of coding and functional GWAS hits

Number of variants causal for disease and gene expression

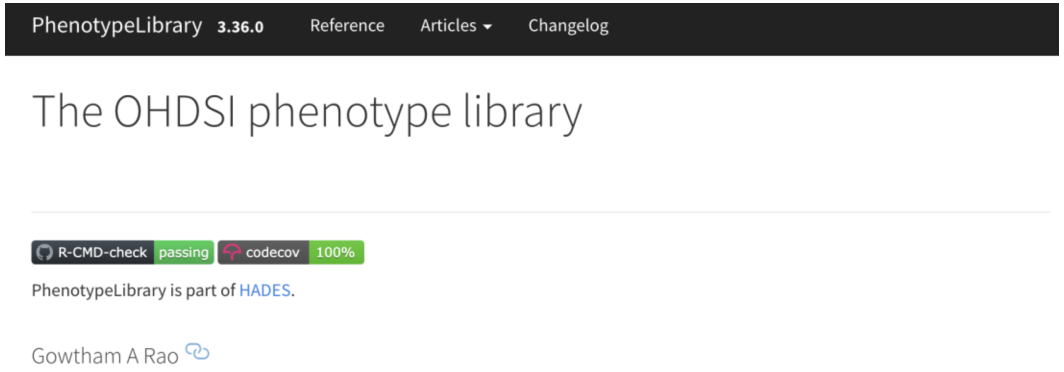


High complexity algorithms generally resulted in the **greatest number of colocalized variants**

High complexity algorithms generally resulted in **higher numbers of novel hits on the coding genome** (i.e., exons), including in exons of the most relevant genes for each disease

Key Takeaways

- **High complexity phenotyping algorithms generally improve GWAS outcomes**, including increased power, hits within coding and functional genomic regions, and co-localization with expression quantitative trait loci
- Biobank-scale GWAS can benefit from **phenotyping algorithms that integrate multiple data domains**
- Curated **repositories of complex, high-quality phenotyping algorithms** are essential to advance the understanding of disease etiology



The OHDSI phenotype library

PhenotypeLibrary is part of HADES.

Gowtham A Rao

OHDSI

OBSERVATIONAL HEALTH DATA SCIENCES AND INFORMATICS