A use case of OHDSI’s ATLAS tool in a biobank-scale GWAS pipeline

Craig C. Teerlink[1,2], Hamid Saoudian[1,2], Richard Boyce[3], Philip S. Tsao[4,5], Kyle M. Hernandez[6], Victoria Zaksas,[6] Pieter Lukasse,[6] Andrew Prokhorenkov,[6] Noah Metoki-Shlubsky,[6] Robert L. Grossman[6], Scott L Duvall[1,2]

[1] VA Informatics and Computing Infrastructure, [2] University of Utah School of Medicine, [3] University of Pittsburgh Department of Biomedical Informatics, [4] VA Palo Alto Health Care System, [5] Stanford University, [6] Center for Translational Data Science, University of Chicago

Background

Infrastructure and data governance limitations have prevented widespread use of Million Veteran Program (MVP) data among the research community and has emerged as a critical barrier to success. Due to participant consent restrictions, MVP data cannot be distributed out to research teams. The VA Data Commons was introduced as a solution for scaling up appropriate access to MVP data and significantly boost computational capabilities. The VA Data Commons is a cloud-based analytic environment that allows VA-credentialed research teams to securely access and perform genome-wide association studies (GWAS) using MVP data. Our goal is to provide a “no-code” environment for users to create cohorts and run GWAS and break down barriers between phenotype specialists and genotype specialists. As such, we are interested in taking advantage of existing open-source tools to implement the GWAS pipeline. To this end, we have incorporated Observational Health Data Sciences and Informatics (OHDSI)’s ATLAS tool for phenotype and covariate selection, the University of Chicago’s G en3 platform for administration and infrastructure, and the University of Washington’s GENetic EStimation and Inference in Structured Samples (Genesis) R package for genome-wide association.

Methods

Most GWAS environments require hand-coding of phenotypes and covariates and use an analytic file as input to the GWAS software. As an alternative strategy, we incorporated an instance of the open-source ATLAS tool to allow users to define dichotomous phenotypes and covariates from raw data, which are then delivered to the GWAS software for computation.

Results

ATLAS is an ideal tool for the VA Data Commons GWAS pipeline for several reasons: it allows users to take advantage of repeatable computable phenotypes that can be shared/validated in other settings, it provides the possibility of creating complex phenotypes for users who may be unfamiliar with this process, and it allows users to interact with data while still preserving privacy and not providing direct visibility to the data.

Conclusion

As the VA Data Commons is made available to VA and non-VA-credentialed users in the near future, we anticipate that users will have powerful phenotyping capability due to the incorporation of the ATLAS tool, which will optimize wide-spread utilization of the MVP data set.