Capstone Proposal - Calculating Polygenic Risk Scores for Autism Spectrum Disorder Across Populations

Katherine G. Wasmer, Data Science M.S.

April 2025

1 Project Overview

1.1 Background

Autism spectrum disorder (ASD) is a complex condition that does not have a single cause. Current studies estimate that 40-80% of risk factors for ASD are genetic. Genome-wide association studies (GWAS) have identified specific common genetic markers that are associated with autism. We can quantify the likelihood of an individual having autism by summing the number of risk variants across all of these markers and weighing them by the marker's effect size. This summation is called the *polygenic risk score* (PRS).

Using the PRS to predict the risk of autism, however, comes with its limitations. These variants are not usually causal, and the patterns of correlation between these genetic variants differ among different ancestral groups. Moreover, GWAS samples are heavily European, which means that the PRS values are not as accurate in other populations. My capstone project focuses on studying the accuracy of PRS for autism across different populations.

1.2 Project Supervision

Dr. Jonathan Terhorst will be supervising my capstone due to his experience in statistical population genetics. His research lab focuses on evolutionary biology, and he implements mathematics, statistics and computer science to study this topic–all of which are essential components of data science.

- I plan to enroll in STATS 750 (Directed Reading) to obtain credit for this independent study, since Dr. Terhorst is part of the statistics department. I have coordinated with Tiffany Comfort, who will issue permission for me to take this capstone in the spring/summer semester.
- I aim to work at least 9 hours a week between May 6 and August 15, which translates to 3 semester hours. (Before the spring/summer semester, I will

have completed at least 25 semester hours of graduate coursework and 19 s.h. of advanced graduate level courses, so 3 s.h. should be sufficient for my capstone.)

- I will email Dr. Terhorst once a week to update him with my progress. We will also plan Zoom meetings or in-person meetings as needed.

1.3 Deliverable

- By the end of the semester, I will have written a scientific paper that includes the following research objectives:

    1. **Literature review** - In the Introduction section, I will include an intensive literature review on the genetics of autism. I plan to cite studies from universities, autism science organizations, and other reputable institutions to gain a full understanding of the expository research in this field. I will include genome-wide association studies that identify specific genes associated with autism.
    2. **PRS across populations** - I will study the current PRS of autism across different ethnic groups, which expands upon the autism-related genes from the literature review. Since many of these studies are heavily Euro-centric, I will collect GWAS and phenotypic data from available biobanks and calculate polygenic risk scores while accounting for population stratification.
    3. **Using the PRS for diagnosing and treating autism** - Although there is no cure for autism, a diagnosis and early intervention can significantly improve the quality of life of children with ASD. Using the PRS to predict whether or not an individual has autism–possibly through predictive machine learning–can be particularly useful in helping families obtain an early diagnosis, especially in communities where autism is underdiagnosed.

- My paper will include the following sections:

    - Introduction and Literature Review (see point 1)
    - Materials & Methods (see the Data Science Applications section)
    - Results - based on my findings, I will also provide supplementary data visualizations.
    - Conclusion & Discussion - here, I will discuss the implications of my findings and integrate the expository and novel research on the genetics of autism.
    - Appendix - I will provide supplementary code written in Python and/or R, since both have scientific libraries that are useful for genomics.

2 Data Science Applications

- The methods for calculating the PRS would be entirely ***regression based***, given its mathematical definition:

$$PRS_j = \sum_i = \beta_i x_{ij} \tag{1}$$

  where j represents a given individual, β_i denotes the effect size at the i^{th} genetic marker, and x_{ij} denotes the number of risk alleles at the i^{th} marker for individual j.

- I would be analyzing large GWAS data sets from the following resources (may be subject to change):

  1. TOPMed - I have access to this data set through the University of Michigan
  2. Genes for Good
  3. The David Reich Lab (Harvard)
  4. IGSR (The International Genome Sample Resource/1000 Genomes)

- When accounting for population stratification, I will identify outlying samples by performing a principal components analysis (PCA) and visualizing the data. This can be done with the Scikit-learn library in Python.

- Due to the large data sets, I would probably need to use the Great Lakes computing cluster.

- Relevant Coursework

  1. STATS 500 - Statistical Learning I: Linear Regression (Fall 2023)
  2. BIOSTAT 626 - Machine Learning for Health Sciences (Winter 2024)
  3. SI 618 - Data Manipulation & Analysis (Winter 2024)
  4. CSE 598-006 - Causality & Machine Learning (Fall 2024)
  5. BIOSTAT 625 - Computing with Big Data (Fall 2024)
  6. BIOSTAT 666 - Statistical Models and Numerical Methods in Human Genetics (Winter 2025)

Approved: Faculty advisor Alex Tsodikov, professor of Biostatistics