Polygenic Risk Scoring for Autism Across Heterogeneous Populations

Katherine Grace Wasmer

University of Michigan

kwasmer@umich.edu

August 8, 2025

Submitted to the Rackham Graduate School in partial fulfillment of the requirements for the Master of Data Science under the supervision of Dr. Jonathan Terhorst.

Abstract

Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder that does not have a single cause, but it is estimated that 40-80% of risk factors are genetic. Genome-wide association studies (GWAS) provide information on common genetic markers that increase the risk of autism, and their summary statistics enable us to calculate the polygenic risk score (PRS). Through Bayesian regression and variational inference, we examine the portability of these scores across East Asian, Middle Eastern, and admixed American populations in N = 321 individuals (131 cases, 190 controls). For each ethnic group, we trained data on the respective UK Biobank samples to create linkage disequilibrium matrices.

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition defined by impairments in social communication and by the presence of restrictive and repetitive behaviors (RRBs). The diagnosis process often occurs in early childhood after noticeable delays in speaking, walking, motor coordination, and other crucial developmental milestones. The Center for Disease Control and Prevention currently reports that 3.2% of American children under 8 years old have autism, with a male-to-female ratio of 3.4:1 (Shaw, 2025). This respective prevalence and sex ratio have changed over time due to better diagnostic tools and a better understanding of autism.

While every individual with autism struggles with social interactions and RRBs, it is important to note that not everyone is affected by these symptoms to the same degree. Until 2013, the DSM-IV classified these variations with a "multi-categorical diagnostic system" (Rosen et. al., 2021) that encompassed Autistic Disorder, Asperger's Syndrome, and Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS). Due to the overlap between Asperger's, PDD-NOS, and high-functioning autism, the DSM-V consolidated these three subtypes into a Level 1 autism diagnosis. The DSM-V diagnostic criteria distinguishes autism by severity levels, with Level 3 individuals requiring the most support on a daily basis.

Autism does not have a single definitive cause, but it is influenced by a combination of genetic and environmental components that differ from person-to-person. Overall, autism is a highly heritable condition with concordance rates among monozygotic twins ranging from 50% (Zhang et. al., 2021) to 98% (Genovese et. al., 2023). The genetics of autism is a growing field of interest, with new discoveries made possible by advanced genotyping technologies. Genome-wide association studies (GWAS) have identified common genetic markers that may be associated with autism. We can quantify an individual's genetic predisposition for ASD by summing the number of risk variants across these markers and weighing them by their effect size. This summation returns the polygenic risk score (PRS)[1]. For the mathematics behind the PRS, please refer to the Methods & Materials section of this paper.

Polygenic risk scoring is highly accurate in predicting diseases primarily caused by common genetic variants (e.g., breast cancer, heart disease, etc.). It comes with limitations, however, with disorders that have complex causes. In conditions with unknown genetic and environmental interactions, the GWAS-identified variants are not typically causal. Moreover, the patterns of correlation between SNPs differ among global

[1]The polygenic risk score (PRS) is synonymous with polygenic score (PGS). Both terms are used in various scientific literature.

populations. Historically, GWAS samples are disproportionately European. Although researchers have made a concerted effort to diversify genomic studies in recent years, 90.53% of total GWAS participants are of unmixed European ancestry (Mills, 2020). The emphasis on European-based genetic markers in disease genomics is particularly problematic for individuals of sub-Saharan African (SSA) descent, who comprise 4.06% of GWAS participants, with the vast majority of individuals being from the African American or African Caribbean diaspora.

This issue in PRS portability can be explained by the research initiated by Cavalli-Sforza et. al. (1994), which connected the dots between population genetics and evolutionary biology. Investigation of linkage patterns and population structural differences allowed for hierarchical clustering of all global populations. Sub-Saharan Africans formed their own "branch" from all other populations, thus giving credence to the Out-of-Africa evolutionary theory. This means that linkage disequilibrium (LD) patterns are different in individuals of mainly SSA ancestry, so polygenic risk scoring from Eurasian groups are not as accurate.

Our paper studies the effectiveness of PRS of autism across underrepresented populations. We examine different approaches with the hope of improving the portability of PRS and increasing health equity among all ethnic groups.

Background & Literature Review

The literature on the genetic architecture of ASD is extensive and encompasses common variants and rare variants. These two categories can be further divided into inherited variants versus *de novo* mutations (genetic variants found in a child that are not found in either parent). The minor allele frequency (MAF) distinguishes common variants from rare variants; it is the proportion of individuals in a population who carry the second most common allele at a given genetic locus. Typically, rare variants have an MAF of 1% or less, while common variants have an MAF of at least 5%.

Common variants are the crux of GWAS findings. They rely on the common disease, common variant (CDCV) hypothesis. Groves et. al. (2019) performed a meta-analysis on GWAS data from Danish individuals in the iPSYCH biobank who were born after 1981 and had a formal ASD diagnosis. This study was the first of its kind to identify common variants associated with the disorder, and it is particularly influential in the sphere of autism genomics. The SNPs rs1620977, rs1452075, r325506, and rs10099100 were reported

to have a small but positive effect on developing ASD. Although these studies have are centered around northwestern European populations due to imbalanced GWAS data, common ASD variants were recently studied in the Han Chinese population (Lin et. al., 2023). Out of a sample of 757 families, they found that SNPs in *SORCS3* were linked to a higher autism risk in both European and Han Chinese populations. The minor allele at rs4307059 on Chromosome 5 was also found to be more prevalent in the cases versus the controls. This is a variant of the *MSNP1AS* gene; postmortem studies show that this gene is expressed more strongly in the cerebral cortices of patients with ASD.

While GWAS data relies on the CDCV, the importance of rare variants in the etiology of autism cannot be ignored. The common disease rare variant hypothesis (CDRV) suggests that in common disorders, rare variants are the driving genetic cause. The Simons Foundation Autism Research Initiative (SFARI) developed a database of SNPs associated with autism. The vast majority of these are rare single gene mutations. Moreover, studying rare variants is important in ASD, because finding statistically significant common variants requires an extremely large sample size to obtain the appropriate statistical power. The Michigan Genomics Initiative, for instance, only has 194 cases in their Freeze 6 dataset and reported rs55648134 on Chromosome 17 as the single statistically significant common variant. However, rs757021989 on Chromosome 1 was identified as a significant rare variant, but no other studies to date have replicated this finding. In Pisa, Italy, N = 127 families with autistic children were sampled and tested for rare copy number variants.

Studies on the X-linked genes corroborate the claim that males have a higher genetic predisposition to autism than females. Although women with lower support needs are often under-diagnosed due to differences in social behaviors, the number of recessive alleles on the X chromosome associated with autism support a gender imbalance in autism, even if the ratio is smaller than 4:1. In autism studies, it is common to stratify by sex and compare the findings to the non-stratified cohort.

In recent years, this literature also encompasses the polygenic architecture of specific autism symptoms. The heritability of autism may extend to social and emotional phenotypes. Warrier et. al. (2018) narrowed in on the genetics behind self-reported empathy and measured this trait with the empathy quotient (EQ).

Methods & Materials

Data Collection and Handling

We collected genetic data from N = 321 individuals from 3 different autism-related studies available on the Gene Expression Omnibus (GEO) database. Although the majority of samples could directly be parsed into Plink 1.9 for genotyping[2], the Almandil study did not include VCF or PLINK binary data, and therefore required a manual conversion from the Illumina microarray format (IDAT) into MAP and PED files. We developed a multi-step pipeline that integrated Python (3.12.1), R (4.4), and Bash.

In the first phase, we converted the IDAT file into a Plink compatible MAP file by merging information from the processed genotypes the Almandil study (2022) with the rsIDs for the Illumina Infinium human CytoSNP-850K iScan BeadChip. We mapped each SNP to its corresponding chromosome, base pair location, and reference/alternate alleles with the `biomaRt()` function from `Bioconductor` (3.12). This new data frame was joined with the genotype file in `pandas` on the chromosome and position columns to eliminate any SNPs not used in Almandil (2022). From this join, we were able to subset the information needed for a MAP file. In the second stage, we performed a second join with the annotations (i.e., manifest file) on the rsID value for the corresponding BeadChip. The genotype file was written in the TOP/BOT nomenclature, corresponding to "Allele A" and "Allele B", so we mapped these alleles to each SNP in the manifest file to obtain the actual nucleotide bases, i.e., adenine (A), cytosine (C), guanine (G), or thymine (T). This allowed us to derive a file in PED format, which could then be converted into a VCF file. Functions for converting IDAT to MAP and PED are included in the Supplementary Materials section.

Statistical Methods

Given a set of N individuals and M variants, let N_j represent a single individual, M_i a single variant, and $x_{ij} \in [0, 1, 2]$ the number of risk alleles carried at that variant. Let β_i denote the GWAS effect size of the given SNP. The polygenic risk score for a single person is therefore defined as the following:

$$PRS_j = \sum_{i=0}^{n} \beta_i x_{ij} \tag{1}$$

[2]One data set (Sakamoto et. al., 2021) had previously been converted into PLINK binary format with the gtc2vcf software. This data was converted and publicized on the Genarchivist forum by user "teepean" (2024).

The vector of all scores in the set is denoted by $\mathbf{y} = [PRS_1, \dots PRS_n]^T$, with corresponding estimates $\hat{\mathbf{y}} = [\widehat{PRS_1}, \dots, \widehat{PRS_n}]^T$. By extension, $\boldsymbol{\beta} = [\beta_1 \dots \beta_n]^T$, and $\mathbf{X} = \begin{bmatrix} x_{11} \dots x_{1n} \\ \vdots \\ x_{m1} \dots x_{mn} \end{bmatrix}$

In matrix form, it is more clear that this equation follows a simple linear regression $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where ϵ represents the absolute difference (i.e., residual) between $\mathbf{y}$ and $\hat{\mathbf{y}}$. The overall objective of this regression is to estimate a vector of effect sizes $\hat{\boldsymbol{\beta}}$ that is robust to changes in GWAS sample size or the addition of new variants. In other words, $\hat{\boldsymbol{\beta}}$ must stay consistent with changes in $\mathbf{X}$. The Bayesian approach to optimizing this value accounts for prior distributions of $\boldsymbol{\beta}$. In the context of polygenic scoring, we want to estimate the posterior distribution $P(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \theta)$, where θ represents "all fixed hyperparameters in the model" that do not have a prior (Zabad et. al., 2023). This true distribution of $\boldsymbol{\beta}$ is illustrated below by Bayes' Theorem.

$$\underbrace{P(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \theta)}_{\text{posterior}} \propto \underbrace{P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \theta)}_{\text{likelihood}} \cdot \overbrace{P(\boldsymbol{\beta}, \theta)}^{\text{prior}} \tag{2}$$

Although Bayesian polygenic scoring software has proven to be accurate and effective for smaller data sets, most modules rely on a Hidden Markov Model for inference. This process increases the O-complexity of the algorithm and is not computationally efficient. Zabad et. al. (2023) proposed an alterative form of the Bayesian PRS that implements variational inference instead of Hidden Markov Models. The Variational Inference of Polygenic Risk Scores (VIPRS) significantly improved the speed of the algorithm while also maintaing accurate scores across populations. We utilized the `viprs` and `magenpy` libraries to conduct polygenic risk scoring on each cluster of our sample.

Results

Our sample consisted of East Asian, Middle Eastern, and admixed American children, with most of them ranging from 5 to 8 years old. Although the Almandil and Sakamoto studies were sampled from homogeneous populations (Saudi and Japanese), the Zhu data set contained N = 56 American individuals from unknown ancestries. We estimated their admixture using the AEon module in Python (3.10); the specific proportions for each individual is listed in the Supplementary Materials section.

Table 1: Demographic information on the 321 samples used for polygenic risk scoring, stratified by study. For the Zhu study, the sex of the probands were not specified and were therefore designated N/A.

Study	**DNA Location**	**Ethnicities Included**	Cases			Controls		
			M	F	Total	M	F	Total
Almandil	Buccal cells	Saudi (MID)	0	22	22	0	51	0
Zhu	Whole umbilical cord blood	Admixed American (AMR)	N/A	N/A	30	N/A	N/A	26
Sakamoto	Peripheral blood	Japanese (EAS)	44	24	68	62	62	124

For each ethnic group, we trained data from the respective UK Biobank samples to create linkage disequilibrium (LD) matrices. GWAS summary statistics came from Matoba et. al. (2020), with 24,603 cases and a total sample of N = 58,794. For quality control, we adhered to the PRS guidelines provided by Choi et. al. (2020). For SNP subsetting, we set a lower bound of $p = 1 \times 10^{-6}$ and an upper bound of $p = 0.01$. We then used these summary statistics to fit our VIPRS models and compute LD matrices.

0.1 East Asian Results

The East Asian sample from the UK Biobank had 2700 individuals, all of which we used for the training of the VIPRS model. On average, the heritability of autism-related variants was approximately 0.03. The highest heritability was found on Chromosome 6, with $h^2_{snp} = 0.07$, and the lowest was on Chromosome 21 with $h^2_{snp} = 0.001$. After fitting the model, we computed the PRS for each chromosome and summed their scores for each individual.

	Count	**Mean**	**S.D.**	**Min**	**25%**	**Median**	**75%**	**Max**
Controls	124	0.290	0.139	-0.046	0.191	0.295019	0.385004	0.625420
Cases	68	0.292954	0.131014	0.017897	0.202560	0.263215	0.393555	0.627998

Table 2: Comparative statistics for the cases and controls in the Sakamoto study.

The distributions between the autistic and non-autistic individuals was very similar. The mean, median, and histogram in Figure 1 all suggest a normal distribution.

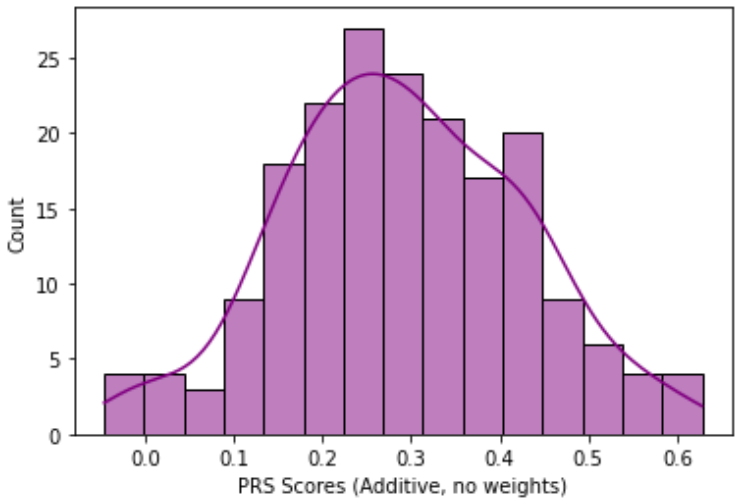

Figure 1: The distribution of polygenic risk scores for the Japanese cohort (N = 192).

0.2 Middle Eastern Results

0.3 Admixed American Results

Discussion

One of the limitations of our study was that we only performed the PRS of the 22 autosomes, while failing to take the sex chromosomes into account.

Generative AI Statement

Chat GPT 4.0 and Google AI were used to assist with debugging scripts, writing PLINK command lines, and parallelizing bash scripts to submit through SLURM. The source code can be found in the GitHub linked in the Supplementary Materials section, and all lines of code generated with AI are properly annotated. Generative AI was also used to help format figures, equations, and tables in LaTeX, but it was not used for the written portion of this paper.

Data Access

The linkage disequilibrium matrices for the UK Biobank samples can be found here: https://shz9.github.io/viprs/download_ld

References

Almandil, N. B., Maram Adnan Alismail, Hind Saleh Alsuwat, Abdulla AlSulaiman, Sayed AbdulAzeez, & J. Francis Borgio. (2023). Exome-wide analysis identify multiple variations in olfactory receptor genes (OR12D2 and OR5V1) associated with autism spectrum disorder in Saudi females. Frontiers in Medicine, 10. https://doi.org/10.3389/fmed.2023.1051039

Cavalli-Sforza, L. L., Paolo Menozzi, & Piazza, A. (1994). *The history and geography of human genes.* Princeton University Press.

Choi, S. W., Mak, T. S., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nature protocols, 15(9)*, 2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Lin, F., Jun Li, Ziqi Wang, Zhang, T., Lu, T., Jiang, M., Yang, K., Jia, M., Zhang, D., & Wang, L. (2023). Replication of previous autism-GWAS hits suggests the association between *NAA1, SORCS3,* and *GSDME* and autism in the Han Chinese population. *Heliyon, 10*(1), e23677. https://doi.org/10.1016/j.heliyon.2023.e23677

Mills, M.C and Rahal, C., (2020). 'The GWAS Diversity Monitor Tracks diversity by disease in real time'. *Nature Genetics, 52*, 242-243. doi: 10.1038/s41588-020-0580-y

Rosen, N. E., Lord, C., & Volkmar, F. R. (2021). The Diagnosis of Autism: From Kanner to DSM-III to DSM-5 and Beyond. *Journal of Autism and Developmental Disorders, 51(12)*, 4253–4270. https://doi.org/10.1007/s10803-021-04904-1

Sakamoto, Y., Shuji Shimoyama, Furukawa, T., Adachi, M., Takahashi, M., Mikami, T., Michito Kuribayashi, Osato, A., Tsushima, D., Saito, M., Ueno, S., & Nakamura, K. (2021). Copy number variations in Japanese children with autism spectrum disorder. *Psychiatric Genetics, 31(3)*, 79–87. https://doi.org/10.1097/ypg.0000000000000276

Shaw, K. A. (2025). Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years — Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022. *MMWR. Surveillance Summaries, 74(2).* https://doi.org/10.15585/mmwr.ss7402a1

teepean. (2024, July 11). *Copy number variations in Japanese children with autism spectrum disorder.* Genarchivist.net. https://genarchivist.net/showthread.php?tid=981&highlight=autism

Zabad, S., Gravel, S., & Li, Y. (2023). Fast and accurate Bayesian polygenic risk modeling with variational inference. The American Journal of Human Genetics, 110(5), 741–761. https://doi.org/10.1016/j.ajhg.2023.03.009

Zhu, Y., Gomez, J. A., Laufer, B. I., Mordaunt, C. E., Mouat, J. S., Soto, D. C., Dennis, M. Y., Benke, K. S., Bakulski, K. M., Dou, J., Marathe, R., Jianu, J. M., Williams, L. A., Gutierrez Fugón, O. J., Walker, C. K., Ozonoff, S., Daniels, J., Grosvenor, L. P., Volk, H. E., & Feinberg, J. I. (2022). Placental methylome reveals a 22q13.33 brain regulatory gene locus associated with autism. *Genome Biology, 23(1)*. https://doi.org/10.1186/s13059-022-02613-1

Supplementary Materials

Appendix A: Continental Ancestry Composition for Samples in Zhu et. al.