Polygenic Risk Scoring for Autism Across Populations

Katherine Grace Wasmer

University of Michigan

kwasmer@umich.edu

August 15, 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 insight on common genetic markers that increase the risk of autism, and GWAS summary statistics allow for the calculation of the polygenic risk score (PRS) of autism. Through Bayesian regression and variational inference, we assess the portability of the PRS in autism across East Asian, Middle Eastern, and admixed American populations in N = 320 individuals (119 cases, 201 controls). We predicted the polygenic risk scores with the `viprs` library and evaluated the performance of the model by calculating the area under the receiving operator curve (AUROC) and determining which individuals scored at the lower 5th percentile and upper 95th percentile. Our method returned low portability across populations; only the admixed American population obtained above-average metrics. This was likely a consequence of the low genotyping rate of the sample combined with imputation errors of rare variants. Future research in this field could look at the integration of rare variants and common variants in autism to improve the predictive power of PRS.

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 acknowledge that not everyone is affected by these symptoms to the same degree. Until 2013, the DSM-IV classified these differences in severity with a "multi-categorical diagnostic system" (Rosen et. al., 2021) that included 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 now distinguishes autism by severity levels, with Level 3 individuals requiring the most support on a daily basis. Recently, the term "profound autism" was coined to emphasize the struggles faced by those with the highest support needs.

Autism does not have a single definitive cause, but it is influenced by a combination of genetic and environmental factors that differ from person-to-person. Overall, autism is a highly heritable condition with concordance rates among monozygotic twins ranging from 50-90% (Zhang et. al., 2021). Some report this rate to be as high as 98% (Genovese et. al., 2023). The genetics of autism is a growing field, with new discoveries made possible by advanced genotyping technologies. Genome-wide association studies (GWAS) have identified common genetic markers 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

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

complex causes. In conditions with unknown genetic and environmental interactions, the GWAS-identified variants are not typically causal. These variants, which are also called single nucleotide polymorphisms (SNPs), have distinct patterns of correlation that differ between global populations. 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.

Cavalli-Sforza et. al. (1994) conducted research on population structural differences that highlights the importance of PRS portability across ethnic groups, especially for those with substantial SSA ancestry. This study connected the dots between population genetics and evolutionary biology through hierarchical clustering. They measured the genetic distances between a carefully-selected set of global populations. By creating a phylogenetic tree with a common root, they found that Sub-Saharan Africans formed a distinct branch from all other groups, thus giving credence to the Out-of-Africa evolutionary theory. Different alleles at nearby loci are often associated with one another; this phenomenon is known as linkage disequilibrium (LD). LD patterns differ based on this hierarchical clustering, so polygenic risk scoring tested on mainly Eurasian samples will not be as accurate when applied to African populations.

Our paper evaluates the portability of polygenic risk scoring for autism across 3 underrepresented populations. We implement statistical learning methods for predicting PRS values, with the long-term goal of improving health equity for all individuals with ASD and their families.

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 and is defined as the proportion of individuals in a population who carry the second most common allele at a given genetic locus. Rare variants have an MAF of 1% or less, while common variants have an MAF of at least 5%.

Common variants are the basis of GWAS findings. They rely on the common disease, common variant

(CDCV) hypothesis, which assumes that conditions with high prevalence are caused by common variants. 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. Other studies have expanded upon these findings, but they are typically centered around northwestern European populations due to imbalanced GWAS data. However, common autism 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, and the vast majority of these SNPs are rare single gene mutations. Finding statistically significant common variants requires an extremely large sample size to obtain the appropriate statistical power, so studying rare variants is often more informative. Bacchelli et. al. (2020) backs up this claim in a study conducted on families with autistic children in Pisa, Italy. N = 127 families were genotyped with the Illumina PsychArray chip and tested for rare copy number variants. One rare CNV in this study was found at *VPS13B*. This gene contributes to the central nervous system and functionality of the eye. Mutations may be associated with Cohen Syndrome, which is a rare genetic disorder that is characterized by developmental delays, problems with vision, and often includes intellectual disability.

Studies on the X-linked genes support 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. Mendes et. al. conducted an X-wide association study (XWAS) on SFARI-reported genes that suggested a strong or moderate link to autism. *ENOX2* (a protein

found in human cancer cells) was found to be significant in females with autism, even after a Bonferroni correction and sanity testing. The study also found that loci in the *FGF13* region had a higher MAF in males than females. This gene regulates the adult central nervous system.

In recent years, this literature has expanded to 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. A GWAS was performed on N = 46,861 samples from 23andme, which mapped their genetic results to their Empathy Quotient (EQ) results. Rolland et. al. (2023) also conducted a meta-analysis on GWAS data to estimate effects of 185 loss-of-function (LoF) genes. This study also utilized brain imaging data from the U.K. Biobank in tandem with the GWAS to determine if there were any structural differences in the brain between LoF carriers and non-carriers. Brain scans often accompany GWAS, XWAS, or other large-scale genomic analyses on autism. The visualization of gene dosages in different regions of the brain offer biological explanations for various autism symptoms. Although many genes have strong links to autism, accurate polygenic risk scoring still poses a challenge.

Methods & Materials

Data Collection and Handling

We collected genetic data from N = 320 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/shell scripting.

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.

[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).

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 through a Bash script. Functions for converting IDAT to MAP and PED are included in the Data Availability section. We also set our default reference assembly to GRCh37 and lifted over any files that were in GRCh38 to maintain consistency among SNPs and base pairs. For this liftover, we used the `CrossMap` script[3]. To avoid reference/allele mismatches, we normalized each converted data set to the GRCh37/hg19 fasta file. We limited our analysis to the 22 autosomes.

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:

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

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

[3]Source code for the CrossMap conversion can be found at `https://github.com/KatherineWasmer/SpectrumPRS/blob/main/useful-shell-scripts/hg38ToHg19.sh`

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 maintaining accurate scores across populations.

We utilized the `viprs` and `magenpy` libraries to conduct polygenic risk scoring on each cluster of our total sample. We first downloaded summary statistics from SPARK, which were published in Matoba et. al. (2020) and publicly available as a TSV file. Adhering to the PRS guidelines provided by Choi et. al. (2020), we set the lower bound of heterozygosity at $p = 1 \times 10^{-6}$ and an upper bound of $p = 0.01$. Any SNPs with a minor allele frequency ≤ 0.01 were removed. This was the threshold recommended by Zabad et. al. (2023) in the VIPRS documentation. The last step in our quality control pipeline was to lift the genomic coordinates over from GRCh38 to GRCh37, so that the GWAS statistics matched the genotype data.

Each individual was assigned to a cluster based on continental-level ancestry from the 1000 Genomes project. The possible categories were European (EUR), Central & South Asian (CSA/SAS), African (AFR), East Asian (EAS), admixed American (AMR) and Middle Eastern (MID). For cases with unreported ethnicities, we estimated admixture with the AEon module in Python (3.10), which included all of these categories except for Middle Eastern. We defined "admixed American" as any individual who had $\geq 10\%$ non-European ancestry from a single continental category. Once we established the different clusters, we ran a 5-fold cross validation (CV) stratified by diagnosis and grouped by ancestry. We used the `StratifiedGroupKFold` object from the Scikit-learn Python library (Pedregosa et. al., 2015) for performing this CV. To obtain an unbiased estimate for all samples, we predicted the PRS of the out-of-sample fold from training on the other 4 folds.

VIPRS computes the PRS on a per-chromosome basis, so we calculated a weighted summation

$$PRS_{weighted} = \sum_{i=1}^{n} w_i PRS_i \tag{3}$$

for $i \in [1, 22]$, where w_i represents the the number of variants assessed on Chromosome i divided by the total number of variants assessed on all autosomes. For tuning hyperparameters θ, we conducted a grid search with a set of pseudo-randomly generated values for each parameter, with a consistent seed across all models. The purpose of this grid search was to find the combination of biologically feasible hyperparameters that minimized the Kullback-Leibler (KL) divergence, which measures the difference between an estimated posterior distribution and its true distribution. This value can be optimized by maximizing the evidence lower bound (ELBO) of the log marginal likelihood. We fixed our heritability (h^2_{snp}) hyperparameter at 70% with a standard error of 5% based on concordance rates between twins. This value was used to generate a plausible grid search for the precision of the prior of effect sizes ($\tau\beta$) and the residual variation ($\sigma\epsilon$). We also fine-tuned the proportion of causal variants (π). Logically, we would not expect a large amount of the SNPs to genuinely have a causal effect on autism, due to linkage disequilibrium. Before running our polygenic risk scoring models, we performed LD pruning with PLINK 1.9 in windows of 50 SNPs.

Results

0.1 Population Information

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	29	N/A	N/A	26
Sakamoto	Peripheral blood	Japanese (EAS)	44	24	68	62	62	124

Table 1: Demographic information on the 320 samples used for polygenic risk scoring, stratified by study. For the Zhu study, the sex of the probands were not specified on the GEO database and were therefore designated N/A.

Our sample consisted of East Asian, Middle Eastern, and admixed American children, with most of them ranging from 5 to 8 years old. This is consistent with the time of diagnosis for many children with autism (*Autism Statistics and Facts*, 2023). Although the Almandil and Sakamoto studies were sampled from homogeneous populations (Saudi and Japanese), the Zhu data set contained N = 55 American individuals from unknown ancestries. We estimated their admixture using the AEon module in Python (3.10). We also computed a principal components analysis on all of our samples to determine the best number of clusters. Our

projection of the first two principal components revealed 3 distinct clusters corresponding to the 3 different studies.

The size of the convex hull for each cluster (see Figure 1) indicated the level of in-group heterogeneity. The Saudi and Japanese hulls were relatively uniform in shape and diameter. Both of these hulls are nearly circular with heavily compressed data points; both populations had an equidistant diameter (d) of 0.008 units across the x-axis (PC1). Given that the first principal component captures the largest amount of variance, these measurements reflect a low level of genetic diversity among the respective populations. Research into the population structure of Saudi Arabia indicates that each region of the country is highly endogamous, with the exception of metropolitan areas such as Riyadh (Kassab et. al., 2020). Data for the Saudi sample were collected in Khobar, which is a city situated on the Persian Gulf in the eastern part of the country. Therefore, we would expect the girls in this study to be genetically similar. The Japanese population from the main islands are strongly homogeneous (Saw et. al., 2015); since the sample was collected from Hirosaki, we would also expect a similar level of genetic relatedness.

The American cluster was more spread out than the other two. All 55 individuals in the American cluster scored a percentage of sub-Saharan African (SSA) ancestry with AEon, with admixture ranging from from 6% to 49%. The differing degrees of SSA ancestry are illustrated in the first principal component with $d = 0.02$, which is 2.5 times greater than the variation in the other two clusters.

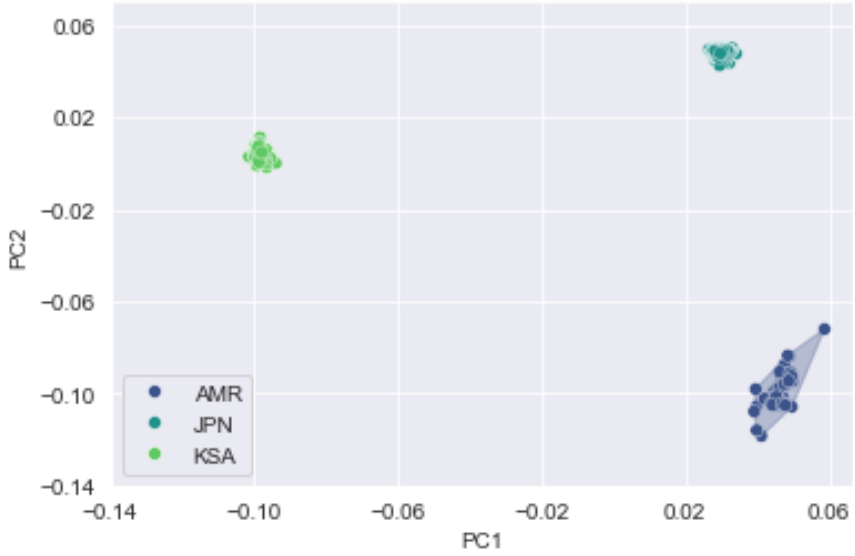

Figure 1: The first 2 dimensions of the principal components analysis on N = 320 samples, labeled by country code.

0.2 Model Performance

Before tuning any of the hyperparameters, we ran a model with the 320 samples to determine a baseline ELBO and preliminary performance metrics. The area under the receiving operator curve (AUROC) was our most important measurement to evaluate. If we were to use these scores as a predictor for autism, an AUROC of exactly 0.5 would mean that our model had the same predictive power as random chance, so we aimed to get a value above this threshold.

After hyperparameter tuning, our ELBO increased from $-\infty$ to an average of -22094.35 across all folds and chromosomes, therefore decreasing the KL divergence. We evaluated our model with the `viprs.eval` module and obtained an overall AUROC of 0.541. This result suggests that our model performed slightly better than random chance, which is also supported by the increase in ELBO. We examined the lower 5th percentile and the upper 95th percentile to evaluate portability and estimate precision-recall rates. For the lower 5th percentile, 62.5% of the sample did not have ASD. Of particular interest was the ethnic distribution; all but 1 of the individuals were from the Japanese cohort. The outlier was admixed American individual A105308.

For the upper 95th percentile, the populations were more evenly distributed, but only 4 of the 16 samples had autism. The two admixed American individuals (A108411 and A108187) both had an autism diagnosis. More surprisingly, the 6 Saudi samples in this upper percentile (CDY11, PDY28, PDY34, PDY41, PDY50, and HBEM1) were all non-autistic. These unexpected results prompted us to evaluate the summary statistics at the population level. For the Japanese and American clusters, the cases scored higher mean PRS scores, but it is worth noting that the confidence intervals are wide and overlap with the means of the control groups. It is unlikely that the differences in means between the cases and controls are statistically significant.

Population	Diagnosis	Min.	Mean (95% CI)	Median
AMR	Control	**-0.065**	-0.017 (-0.064, 0.031)	0.012
	ASD	-0.086	**-0.013** (-0.067, 0.042)	**-0.011**
JPN	Control	**-0.085**	-0.026 (-0.081, 0.030)	**-0.025**
	ASD	-0.111	**-0.018** (-0.074, 0.038)	-0.048
KSA	Control	-0.040	**0.004** (-0.031, 0.040)	**-0.001**
	ASD	**-0.010**	-0.004 (-0.011, 0.001)	-0.005

Table 2: PRS summary statistics, aggregated by case/control status and population. The higher summary statistic of each pair is in bold font.

We also investigated the evaluation metrics for the stratified samples. The admixed American group was

the only group to score higher than 0.5 for the AUROC; they scored an AUROC of 0.556 and a precision of 0.633. The Saudi cohort had extremely poor precision (0.233) and AUROC (0.034). The Japanese sample scored in between these two groups, with a precision of 0.423 and an AUROC of 0.415.

Discussion

Our study explored the use of Bayesian regression with VIPRS to assess the portability of polygenic risk scoring in autism. Even with stratified and grouped cross validation, the PRS distributions differed across the groups. Although the admixed American group scored above-average metrics, it would be disingenuous to use this model as a predictor of autism. We only explored bi-allelic locations from the GWAS, even though there are numerous CNVs that have a causal effect on autism.

One problem we faced during the modeling phase was the difference in SNP counts between the 3 data sets. Out of the 3, the Saudi cohort had the lowest number of variants per chromosome. Further inspection of the Almandil study revealed that only the exome was sequenced. The exome, or the protein-coding regions of the genome, only make up 1-2% of the total genome. This could be a possible explanation for the poor precision and AUROC scores for the Saudi cohort. Only 243,345 SNPs were used, which isn't enough to study common variants. In comparison, the Japanese data set had 20 times as many SNPs. However, the PRS did not translate well across the Japanese population either. Zabad et. al. (2023) acknowledges that the VIPRS algorithm still does not work well across multiple ancestries, even though there are ancestry-specific linkage disequilibrium matrices available for harmonizing.

Future research into polygenic risk scoring of autism could take rare variants into account. Since rare variants have a higher penetrance, Williams et. al. (2024) proposed a method called RICE (polygenic Risk predictions Integrating Common and rarE variants) that includes ensemble learning for common variants (RICE-CV) and association testing for rare variants (RICE-RV). This study included European, African, East Asian, South Asian, Latin American, and Middle Eastern populations. The portability significantly improved for the African cohort. A similar approach was taken to predict the PRS for 5633 children with autism (Bourque et. al.), but the AUROC was only 0.63, which was 9 percentage points higher than the AUROC we reported.

Although there is no cure for autism, a diagnosis and early intervention can significantly improve the

quality of life of a child with ASD. Improving the PRS of autism with integration of rare variants, CNVs, and de novo mutations is a topic of interest. Although it's not possible to obtain 100% accuracy in these scores, learning this information early on can help families prepare for raising a child on the spectrum. As autism biobanks such as SPARK and MSSNG increase in size, the ability to identify autism through predictive machine learning may very well be possible.

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. Relevant source code can be found in the GitHub linked in the Data Availability 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 any portion of the written report.

Data and Code Availability

Code Repository (may be subject to occasional updates): `https://github.com/KatherineWasmer/SpectrumPRS`

GWAS data from SPARK (all populations): `https://bitbucket.org/steinlabunc/spark_asd_sumstats/src/master/SPARK_update_model1QC.tsv.gz`

GEO Accession IDs for samples used:

- American: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178204`
- Japanese: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144918`
- Saudi: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE221098`

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

Autism statistics and facts. (2023). Autism Speaks; Autism Speaks Inc. https://www.autismspeaks.org/autism-statistics-asd

Bacchelli, E., Cameli, C., Viggiano, M. et al. An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray. Sci Rep 10, 3198 (2020). https://doi.org/10.1038/s41598-020-59922-3

Bourque V, Schmilovich Z, Huguet G, et al. Genomic and Developmental Models to Predict Cognitive and Adaptive Outcomes in Autistic Children. JAMA Pediatr. 2025;179(6):655–665. doi:10.1001/jamapediatrics.2025.0205

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

Ch. Kassab, A., Alaqeel, H.F.M., Messaoudi, S.A. et al. Population data and genetic diversity analysis of 17 Y-STR loci in Saudi population. *Egypt J Forensic Sci* **10**, 30 (2020). https://doi.org/10.1186/s41935-020-00205-3

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

Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze S, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh PR, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. Nature Genetics 48, 1284–1287 (2016).

Genovese, A., & Butler, M. G. (2023). The Autism Spectrum: Behavioral, Psychiatric and Genetic Associations. Genes, 14(3), 677. https://doi.org/10.3390/genes14030677

Khubrani, Y. M. Y. (2020). Genetic Diversity and Population Structure of Saudi Arabia (Version 1). University of Leicester. https://doi.org/10.25392/leicester.data.11799858.v1

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

Mai, A. S., Yau, C. E., Tseng, F. S., Foo, Q. X. J., Wang, D. Q., Tan, E. K. (2023). Linking autism spectrum disorders and parkinsonism: clinical and genetic association. *Annals of clinical and translational neurology, 10(4*), 484–496. https://doi.org/10.1002/acn3.51736

Matoba, N., Liang, D., Sun, H., Aygün, N., McAfee, J. C., Davis, J. E., Raffield, L. M., Qian, H., Piven, J., Li, Y., Kosuri, S., Won, H., Stein, J. L. (2020). Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. *Translational psychiatry, 10*(1), 265. https://doi.org/10.1038/s41398-020-00953-9

Mendes, M., Chen, D. Z., Engchuan, W., Leal, T. P., Thiruvahindrapuram, B., Trost, B., Howe, J. L., Pellecchia, G., Nalpathamkalam, T., Alexandrova, R., Salazar, N. B., McKee, E. A., Alfaro, N. R., Lai, M. C., Bandres-Ciga, S., Roshandel, D., Bradley, C. A., Anagnostou, E., Sun, L., & Scherer, S. W. (2024). Chromosome X-Wide Common Variant Association Study (XWAS) in Autism Spectrum Disorder. *medRxiv : the preprint server for health sciences*, 2024.07.18.24310640. https://doi.org/10.1101/2024.07.18.24310640

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

Pedregosa, F., Buitinck, L., Louppe, G., Grisel, O., Varoquaux, G., Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, 19(1), 29–33. https://doi.org/10.1145/2786984.2786995

Rolland, T., Cliquet, F., Anney, R.J.L. et al. Phenotypic effects of genetic variants associated with autism. Nat Med 29, 1671–1680 (2023). https://doi.org/10.1038/s41591-023-02408-2

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

Saw, WY., Liu, X., Khor, CC. et al. Mapping the genetic diversity of HLA haplotypes in the Japanese populations. *Sci Rep* **5**, 17855 (2015). https://doi.org/10.1038/srep17855

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

Williams, J., Chen, T., Hua, X., Wong, W., Yu, K., Kraft, P., Li, X., & Zhang, H. (2024). *Integrating Common and Rare Variants Improves Polygenic Risk Prediction Across Diverse Populations.* https://doi.org/10.1101/2024.11.05.24316779

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