Population Effects in GWAS Based Risk Scores

Ross DeVito

ABSTRACT

Polygenic risk scores (PRS) use the findings of genome-wide association studies (GWAS) to predict risk for a trait or disease. While these scores seem poised to drastically improve future clinical care, current PRS perform significantly better for members of the discovery population of the original GWAS. This ancestry-based inequity of care is a major ethical obstacle, especially as the current Eurocentric bias in GWAS would exacerbate existing health disparities. In this context, population effects are differences in risk score performance resulting from either the discovery population used in the underlying GWAS or the population of the sample being scored. In this paper, differences in score distributions between populations were used to detect these population effects in existing risk scores for schizophrenia and height without ground truth data.

Introduction

Genome-wide association studies use large samples of corresponding genotypes and phenotypes to find associations between genetic variations, typically single-nucleotide variants (SNVs), and a trait or disease. The results of these studies can be used to form polygenic risk scores, which provide an estimate of predisposition for the complex trait. These scores are of great interest for their potential to improve clinical care through more accurately quantifying risk, allowing for better targeting of treatments, and serving as common biomarkers. PRS have even been shown to be more accurate than current clinical models for breast cancer, prostate cancer, and type 1 diabetes, but only when the target person being scored is from the population that composed the underlying GWAS's sample (the discovery population)[1].

When the target's population is not represented in the discovery population, PRS are shown to be sifnificantly less accurate[1,2]. This performance variation based on population is not surprising and can be explained in part by genetic differences in minor allele frequencies and linkage disequilibrium structure between populations. These differences lead to a violation of an implicit assumption made by the GWAS's underlying statistical model: that individuals being scored are from the same population as the discovery population[3]. Both theoretical simulations and empirically studies confirm this result and suggest that performance decreases as genetic divergence from the discovery population increases[1–3].

This decrease in performance presents a significant obstacle to clinical use, as it would lead to inequity in care based on ancestry. PRS currently have a clear Eurocentric performance bias due to the relative overrepresentation of European descent samples in GWAS discovery populations (16% of global population, but about 79% of all GWAS participants)[1]. Additionally, European sample populations tend to be the largest and fastest growing, which gives them greater statistical power. The relative share and size of non-European samples has stagnated since 2014, with minimal growth in samples with African or Hispanic/Latino ancestry[1]. With the lack of progress in this area, it is important to understand the implications.

The goal of this project is to examine existing GWAS based risk scores to look for signs of population effects. Here, this can mean the differences in score based on sample population or the impacts discovery population has on performance, either overall or by target population. Samples from the 1000 Genomes Project were scored using PRS from the NHGRI-EBI GWAS Catalog for studies with both single and multi-super population discovery samples. In the absence of ground truth phenotype data, heuristics and variance were used as proxies to evaluate performance. The expected correlation between decreased performance and increased divergence from the discovery population was also checked for each study.

Methods

GWAS and Risk Score Data

The NHGRI-EBI GWAS Catalog provides summary statistics and background metadata for all published studies that meet the inclusion criteria[4]. The summary statistics provide the locations, risk alleles, and weights for statistically significantly associated sites required to generate PRS. The weights are either odds-ratios or beta values based on whether the underlying study is for a case/control or quantitative trait. The metadata was used to determine the population makeup of each GWAS's discovery sample.

Genotype and Population Reference Panel

Selected PRS were used to score all samples from the 1000 Genomes Project[5]. Only biallelic SNVs were considered and the GRCh38 reference assembly was used. This included 73,257,633 SNVs for 2,548 samples. By super population, 660 samples are African (AFR), 347 Ad Mixed American (AMR), 504 East Asian (EAS), 503 European (EUR), and 489 South Asian (SAS).

To evaluate the relationship between genetic divergence from the discovery population and score performance, a measure of how genetically different populations are is needed. As in previous studies[3,6], the fixation index (F_{ST}) was used to measure this divergence. F_{ST} was calculated pairwise between all populations and super populations using the 1000 Genomes samples.

Generating Scores

The GWAS_Scorer Python library[1] was created as part of this project to allow for easy scoring of samples based on PRS from the GWAS Catalog. The library takes in the samples' genotype data in VCF format, then preprocesses it to HDF5 to allow for faster runtimes and lower memory usage when generating scores. Once this is done, scores can be quickly be generated for all preprocessed samples by providing the desired study's ID (the "STUDY ACCESSION" field in the catalog) and paths to the preprocessed genotype data and required GWAS Catalog files. The library can also use the preprocessed HDF5 data to calculate all pairwise fixation indices between populations. Python functions and command line tools are provided for this preprocessing, scoring, and F_{ST} calculation. These tools should work with all VCF data that includes the required fields.

Evaluating Scores

Even without ground truth phenotype data, aspects of PRS performance can still be inferred by comparing distributions of scores between populations. Accuracy can be estimated by how well real world expectations hold in the predicted scores, while variance can be used to proxy a PRS's variance explained by population.

Heuristics

For diseases or traits where the relative prevalences or distributions between populations are known, we can look to confirm that this relationship holds true in the predicted scores. With schizophrenia, the first disease to be examined here, the prevalence has shown to be similar across populations[2,7]. To test that this was true for a study's predicted risk scores, the scores were grouped by super population and a Kruskal–Wallis test was used to determine whether all groups' scores were drawn from the same distribution, as would be expected. A Conover–Iman test was also used to determine pairwise if super populations' scores were from the same distribution. These results acted as a proxy for accuracy, while also showing which target populations' scores appeared out of line with expectations. Of particular note was which target populations had different distributions than the discovery population's super population, as the discovery population's distribution is the most likely to be correct.

For the second trait, height, the heuristic test was if the European height cline held in predicted scores. This would be true if the combined distribution of predicted heights for FIN, CEU, and GBR population samples was significantly greater than the distribution for TSI and IBS samples[2], as determined with a one sided Mann–Whitney U test.

Variance

When grouped by population or super population, the relative variances of predicted risk scores can be used as a proxy for variance explained for each group. Equal variance between populations' scores is a positive indicator that a PRS has similar variance explained for all target populations. Equality of score variance for all super populations, as well as pairwise between them, was evaluated for each study using Brown–Forsythe tests.

Variance can also be used in conjunction with the fixation index to see if there is a correlation between divergence from the discovery population and decrease in variance explained. Pearson product-moment correlation was used to quantify the strength and significance of the relationship.

[1] github.com/RossDeVito/GWAS_Scorer

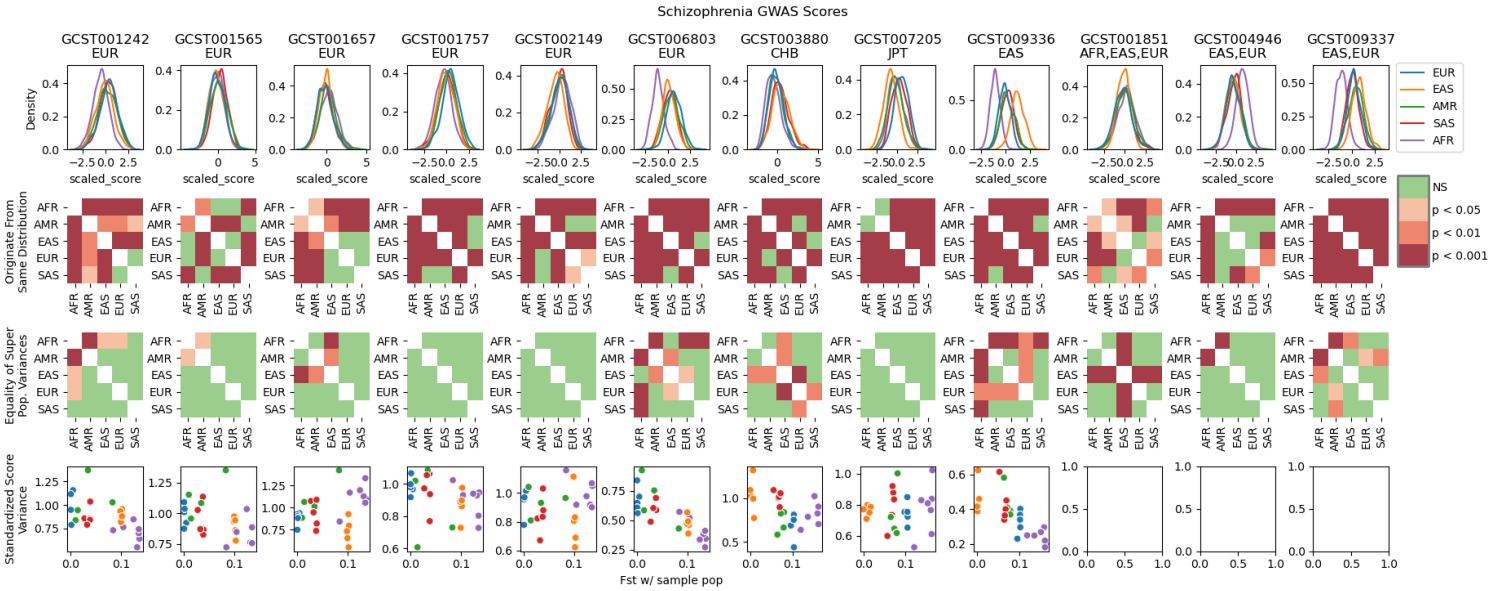

Figure 1. Schizophrenia risk score distributions and tests. P-values Holm–Bonferroni adjusted by pairwise test. The top row shows each super population's score distribution. The next two rows down show the pairwise Conover–Iman then Brown–Forsythe test results between super populations. The bottom row plots each population's variance against its F_{ST} with the discovery population.

Study ID	Sample Pop.	Date	Num. SNVs	All Same Dist. (K-W p-val)	Pairwise Same Dist.	All Same Var. (B-F p-val)	Pairwise Same Var.	Pearson r	Pearson p-val
GCST001242	EUR	2011-09-18	25	False (2.69E-67)	2	False (0.00261)	9	-0.491	0.108
GCST001565	EUR	2012-06-12	14	False (2.09E-17)	4	True (0.0834)	10	-0.379	0.451
GCST001657	EUR	2012-09-01	9	False (2.33E-21)	4	False (1.48E-4)	8	0.262	1.0
GCST001757	EUR	2012-12-01	16	False (1.33E-44)	2	True (1.0)	10	-0.323	0.756
GCST002149	EUR	2013-08-25	27	False (2.83E-38)	3	True (1.0)	10	0.075	1.0
GCST006803	EUR	2018-02-26	83	False (7.94E-241)	1	False (4.37E-8)	6	-0.804	8.99E-06
GCST003880	CHB	2016-12-06	9	False (2.70E-63)	2	False (1.36E-4)	6	-0.474	0.131
GCST007205	JPT	2018-10-03	13	False (4.75E-108)	1	True (1.0)	10	0.090	1.0
GCST009336	EAS	2019-11-18	21	False (∼0.0)	1	False (1.28E-12)	4	-0.741	0.000166
GCST001851	AFR, EAS, EUR	2013-02-01	12	False (1.0E-14)	5	False (6.71E-10)	6	x	x
GCST004946	EAS, EUR	2017-10-09	217	False (5.64E-137)	4	False (0.00642)	9	x	x
GCST009337	EAS, EUR	2019-11-18	198	False (1.26E-270)	0	False (1.36E-4)	7	x	x

Table 1. Schizophrenia studies. P-values are Holm–Bonferroni adjusted.

Results

Schizophrenia

PRS based on 12 GWA studies were evaluated (table 1). Six of these studies had European discovery populations, three were East Asian (either CHB, JPT, or EAS), and three had discovery samples that included multiple super populations. As is typical for case/control studies, the PRS output was an odds-ratio. The score used for analysis was the log of this odds ratio standardized by study.

Heuristic: Risk Scores from Same Distribution

The prevalence of schizophrenia is roughly similar across populations, so we would expect the scores for all super populations to be drawn from the same distribution[2,7]. However, this was not found to be the case for any of the studies. Studies with discovery samples including multiple super populations had on average more pairs of super populations with similar distributions, but the number of PRS tested here is too small to make a general conclusion.

Equal Variance

Half of the European based PRS and one of the three with an East Asian discovery population had equal score variance between all super populations, indicating they are most likely to have similar variance explained across all populations. When the GWAS discovery sample was European, the African super population most often had pairwise differences in variance with other super populations. It is a three way tie between AFR, EAS, and EUR for most pairwise differences when the discovery population was East Asian.

Relationship with Genetic Divergence

For both European and East Asian discovery population based risk scores, the most recent studies were the only ones to have a statistically significant correlation between population divergence from the discovery sample and variance. In both cases these studies also included the largest number of risk alleles for their respective sample populations.

Study ID	Sample Pop.	Date	Num. SNVs	Height Cline (M-W p-val)	All Same Var. (B-F p-val)	Pairwise Same Var.	Pearson r	Pearson p-val
GCST000175	EUR	2008-07-01	50	False (0.0435)	True (0.479)	10	-0.0181	1.0
GCST000176	EUR	2008-09-12	18	False (1.0)	False (3.25E-25)	2	-0.841	1.30E-6
GCST000174	EUR	2008-09-17	19	False (1.0)	False (0.00175)	8	-0.639	0.00617
GCST000372	EUR	2009-04-22	19	False (1.0)	False (0.00326)	8	-4.92E-01	0.117
GCST000817	EUR	2012-10-24	179	False (1.0)	False (0.00784)	9	0.195	1.0
GCST001956	EUR	2013-09-12	83	True (0.000120)	False (1.84E-11)	5	-0.802	1.39E-5
GCST002647	EUR	2015-09-24	173	False (1.0)	True (0.479)	10	-0.316	1.0
GCST004212	GBR	2017-06-21	29	False (1.0)	False (4.98E-148)	1	0.563	0.0332
GCST005908	EUR	2018-07-19	18	False (1.0)	True (0.826)	10	0.0581	1.0
GCST008163	EUR	2019-07-17	505	False (1.0)	False (1.40E-50)	3	0.467	0.162
GCST000611	JPT	2010-03-25	24	False (1.0)	False (3.02E-12)	6	-7.69E-01	6.54E-5
GCST001885	CHB	2013-06-01	11	True (0.000437)	False (2.82E-7)	6	-0.6	0.0153
GCST002702	EAS	2015-07-15	124	False (1.0)	False (0.000766)	6	2.22E-01	1.0
GCST008839	JPT	2019-10-16	527	False (1.0)	False (0.000169)	7	-1.05E-01	1.0
GCST001263	AFR	2011-11-08	32	False (1.0)	True (0.405)	10	-0.228	1.0
GCST001290	AFR	2011-11-25	7	False (1.0)	False (1.93E-13)	5	-0.799	1.54E-5
GCST008053	Non-EUR	2019-07-02	201	False (1.0)	False (1.59E-7)	6	x	x
GCST008904	AFR,AMR,EAS,EUR	2019-10-23	20	False (1.0)	False (0.00136)	7	x	x

Table 2. Height studies. P-values are Holm–Bonferroni adjusted.

Height

Eighteen PRS for height were examined: ten with European discovery populations, four East Asian, two African, and two with multiple super populations represented (table 2). For one of the two multi-population studies the discovery population included samples from the four non-European super populations, while the other included all but the South Asian super population.

Heuristic: European Height Cline

The heights of Northern Europeans were predicted to be statistically significantly greater than Southern Europeans by just one European and one East Asian discovery population PRS. This performance is particularly underwhelming for the EUR based studies, as we would expect these to perform best on a test that only evaluates performance scoring targets who are members of the discovery population.

Equal Variance

Three European and one African sample population based study showed equal variance for all super populations, implying similar variance explained. The AFR super population again had on average the most pairwise differences when the study's sample population was EUR, with an average of 1.9 differences per study followed by EAS with an average of 1.7. With an East Asian discovery population, the AFR and AMR super populations have the most pairwise differences with averages of 2.75 and 2 per study respectively. Only one of the African based studies showed pairwise differences between super

population variances. Here, the AFR target samples were shown to have significantly greater variance than all other groups but Ad Mixed American.

Relationship with Genetic Divergence

A significant negative correlation was found between fixation index and variance for three EUR, one EAS, and one AFR based PRS. Of these, four used relatively low number of SNVs (7, 18, 19, and 24). The fifth, which used 83 SNVs, was the only score with a European discovery population to correctly predict the height cline. The strong European performance combined with the correlation between decreased performance and genetic divergence from Europeans suggests this score may be overfit to Europeans.

Discussion

The tests used here allow us to look for population effects in GWAS based risk scores using just genotype samples with population labels. With validation of the methodology of using heuristics and variance to proxy accuracy and variance explained, these types of tests could be useful tools for detecting population effects in risk scores, even without access to ground truth phenotype data. The reduced availability and scale of paired genotype and phenotype data for non-European populations makes this capability especially useful.

Combining performance information across more studies and perhaps more diseases and traits at once has the ability to increase our power to detect and describe population effects, but this approach comes with additional challenges. Factors such as the underlying distributions of the traits by population, differences in sample size and availability by population, and the age and methodologies of individual studies could confound results. Despite this added difficulty, larger meta-analysis of this type would be better powered to detect the true effect sizes of both target and discovery population based effects.

References

1. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591, DOI: 10.1038/s41588-019-0379-x (2019).
2. Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *The Am. J. Hum. Genet.* **100**, 635–649, DOI: 10.1016/j.ajhg.2017.03.004 (2017).
3. Scutari, M., Mackay, I. & Balding, D. Using genetic distance to infer the accuracy of genomic prediction. *PLOS Genet.* **12**, 1–19, DOI: 10.1371/journal.pgen.1006288 (2016).
4. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005–D1012, DOI: 10.1093/nar/gky1120 (2019). 30445434[pmid].
5. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74, DOI: 10.1038/nature15393 (2015).
6. Guo, J. *et al.* Global genetic differentiation of complex traits shaped by natural selection in humans. *Nat. communications* **9**, 1865–1865, DOI: 10.1038/s41467-018-04191-y (2018). 29760457[pmid].
7. Ayuso-Mateos, J. L. Global burden of schizophrenia in the year 2000: Version 1 estimates.

Author Contribution and Reflection

I really enjoyed this project and spent probably more time on it than I did all the problem sets together. The GWAS_Scorer library I wrote made things much easier in the end, but took a solid chunk of time in the beginning. The combination of the paper from the journal club and the paper you suggested by the same author[2] were a really good starting points and helped a lot with finding other relevant papers. Most of the actually evaluation of scores in this paper happens on a by PRS basis. I had originally hopped to do more pooling many studies together to look more explicitly for things like the association between a sample populations performance and the discovery population, but it would have been hard to control for all the potential influences outside of population. It was already a little hard to distinguish some noise from signal without ground truth.