PRECISION HEALTH
UNIVERSITY OF MICHIGAN

Michigan Genomics Initiative Freeze 5 Polygenic Scores

Brett Vanderwerff[1*], Lars G. Fritsche[1], Emily Bertucci-Richter[1], Snehal Patil[1,2], Matthew Zawistowski[1], Michael Boehnke[1], Xiang Zhou[1], and Sebastian Zöllner[1,3]

[1]Department of Biostatistics and Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA. [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA. [3]Department of Psychiatry, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA

*To whom correspondence regarding data preparation should be addressed: brettva@umich.edu

1 Data Description

Polygenic scores (PGS) summarize an individual's genetic liability of a trait. In the simplest setting, PGS are calculated as the weighted sum of the number of trait associated alleles observed in the individual with weights chosen from GWAS effect size estimates. PGS have many possible applications in research including cohort stratification[1,2], identifying shared genetic background between traits[3], Mendelian randomization[4], and testing for gene–environment interactions[5].

In this release, we offer PGS constructed for 6 binary phenotypes represented in the Global Biobank Meta-analysis Initiative (GBMI) PGS analysis, each with SNP-based heritability at the liability scale > ≈ 5%[6]. The traits are primary open angle glaucoma (POAG), thyroid cancer (ThC), abdominal aortic aneurysm (AAA), gout, chronic obstructive pulmonary disease (COPD), and asthma[6]. These PGS are constructed based on prediction models trained on the leave-MGI-out summary statistics available from the GBMI and are available for 70,266 MGI participants included with the release of Freeze 5 (Table 1).

	European	African	West Asian	East Asian	Central/South Asian	Native American
Phenotype	(n=60,959)	(n=4,436)	(n=1,883)	(n=1,426)	(n=963)	(n=599)
POAG	366	69	16	12	4	3
ThC	1,119	52	58	27	9	12
AAA	1,335	66	35	7	2	3
gout	2,915	285	94	58	25	15
COPD	6,142	410	110	21	22	26
asthma	11,149	1,127	340	206	141	103

Table 1. Cohort properties. The total sample sizes (in parentheses) and the numbers of cases for each ancestry group and phenotype. We inferred genetic ancestry European, African, West Asian, East Asian, Central/South Asian and Native American using ADMIXTURE and the Human Genome Diversity Panel samples and super-population labels as reference. POAG: primary open angle glaucoma; ThC: thyroid cancer; AAA: abdominal aortic aneurysm; COPD: chronic obstructive pulmonary disease.

2 Data Evaluation

Significant ancestry signals have been described in PGS generated elsewhere for schizophrenia, thus we sought to evaluate the ancestry component contained within PGS generated in MGI[7]. We evaluated the distribution of PGS across the majority genetic ancestry groups inferred in MGI. The difference between the mean PGS of the ancestry groups was significant with $p < 2 \times 10^{-16}$ for each of the 6 phenotypes (Figure 1). Considering the limited sample sizes in MGI for Central/South Asian, East Asian, Native American, and West Asian ancestries, we focused our subsequent evaluations on the European and African ancestries only. We recommend users exercise caution using PGS from the unevaluated ancestry groups as clear ancestry differences exist and the evaluations of Europeans and Africans may not be a reflection of or informative for the data quality in other ancestry groups.

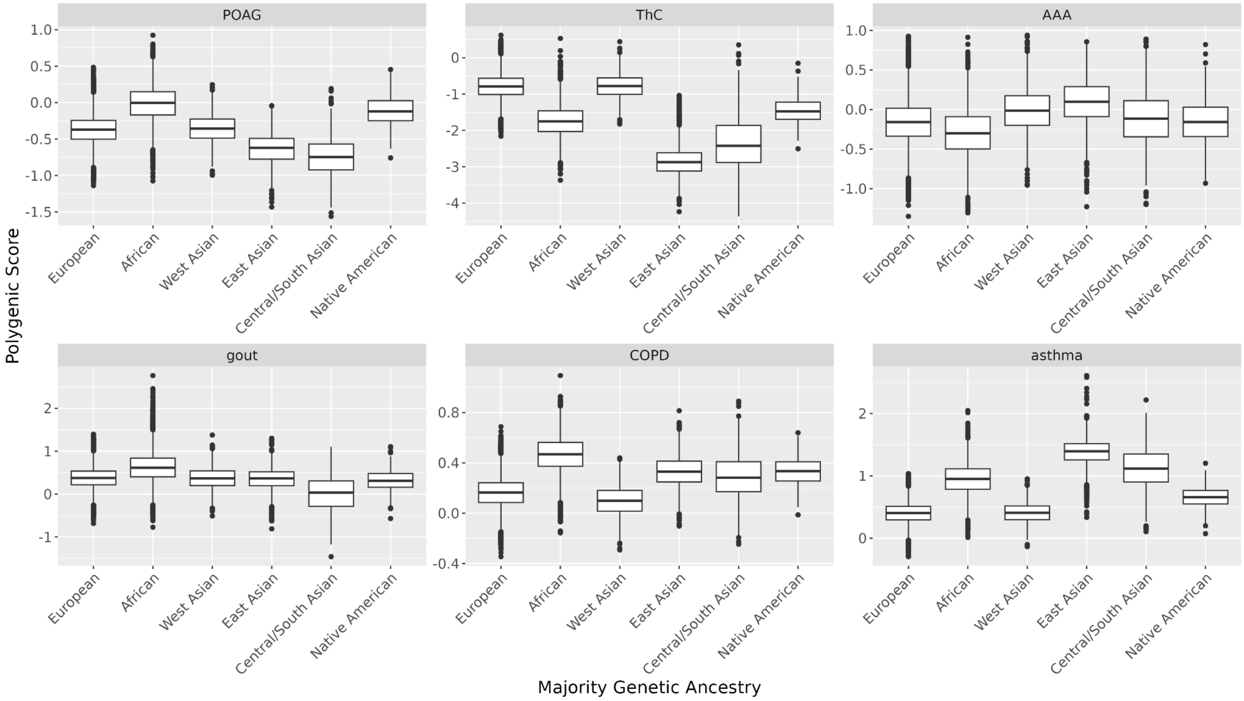

Figure 1. Polygenic score distributions by ancestry. We inferred genetic ancestry European, African, West Asian, East Asian, Central/South Asian and Native American using ADMIXTURE and the Human Genome Diversity Panel samples and super-population labels as reference. Outliers beyond 1.5 times the interquartile range are plotted as single points. POAG: primary open angle glaucoma; ThC: thyroid cancer; AAA: abdominal aortic aneurysm; COPD: chronic obstructive pulmonary disease.

We summarized the predictive accuracy of each PGS by measuring R^2 on the liability scale ($R^2_{liability}$) which ranged from 0.0144 (AAA) - 0.0543 (POAG) in Europeans and 0.000201 (AAA) - 0.0220 (POAG) in Africans. Despite the small $R^2_{liability}$ values, all phenotypes except abdominal aortic aneurysm in Africans had significant ($p < 0.05$) predictive accuracy in MGI (Figure 2).

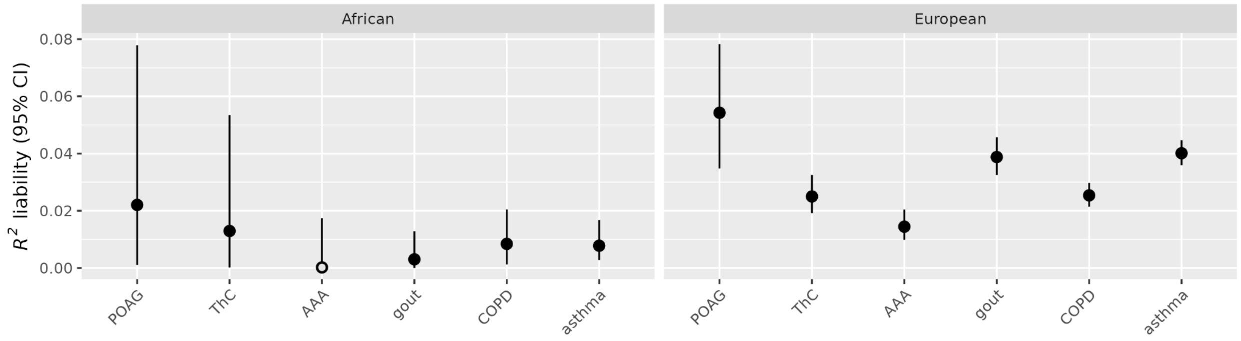

Figure 2. *R*^{2} on the liability scale. 95% confidence intervals (CI) are derived from bootstrapping with 1000 replicates. POAG: primary open angle glaucoma; ThC: thyroid cancer; AAA: abdominal aortic aneurysm; COPD: chronic obstructive pulmonary disease. Closed circles denote Nagelkerke's R^2 p < .05.

We compared odds ratios (OR) for each PGS decile to the lowest decile. In Europeans, the largest significant OR was 5.58 for thyroid cancer (p= 3.37 x 10^{-24}) and each phenotype showed a clear trend of increasing OR with PGS decile. In Africans, the largest significant OR was for gout (2.39, p=3.63 x 10^{-3}) and the OR of the 10th decile was consistently higher than the 1st but the differences between these deciles were consistently smaller compared to that observed in Europeans. (Figure 3).

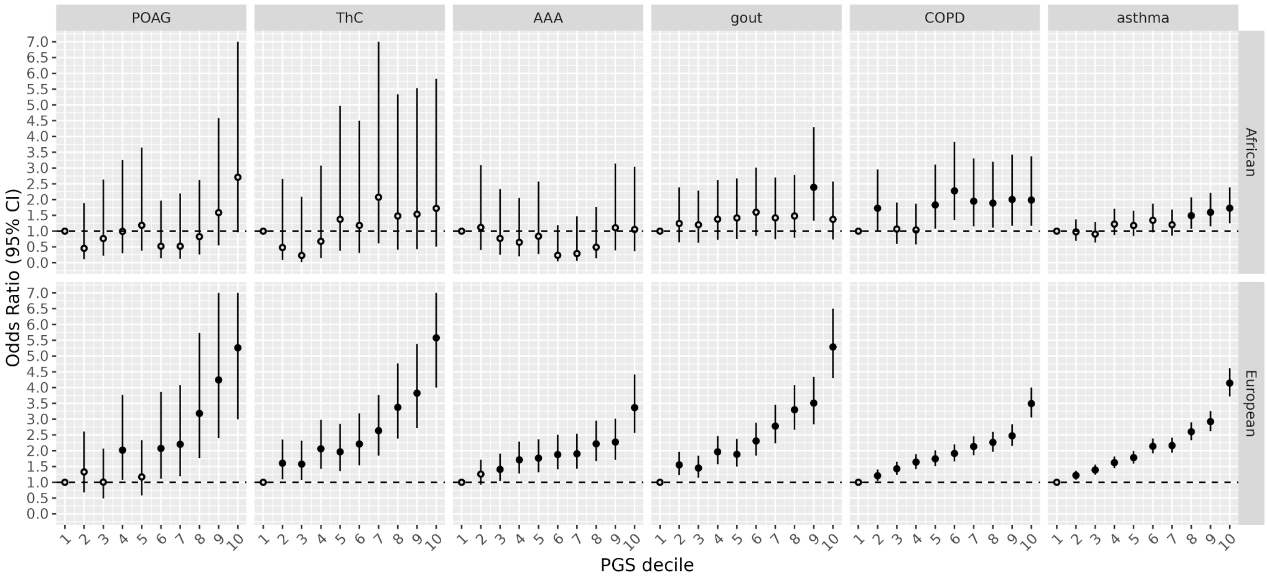

Figure 3. Odds ratio by PGS decile. 95% confidence intervals (CI) are derived from bootstrapping with 1000 replicates. The dashed black lines indicate an odds ratio of 1. POAG: primary open angle glaucoma; ThC: thyroid cancer; AAA: abdominal aortic aneurysm; COPD: chronic obstructive pulmonary disease. Closed circles denote odds ratio p < .05.

We inspected the distribution of disease prevalence across select age bins and PGS strata. The largest differences in disease prevalence between PGS strata within a single age bin was 0.199 (bottom 10%) vs 0.369 (top 10%) for asthma in Africans aged 50-70 years and 0.125 (bottom 10%) vs 0.369 (top 10%) for asthma in Europeans aged 30-50 years (Figure 4). We note a trend toward reduced asthma prevalence among aged European participants, which has been described in a European population elsewhere[8].

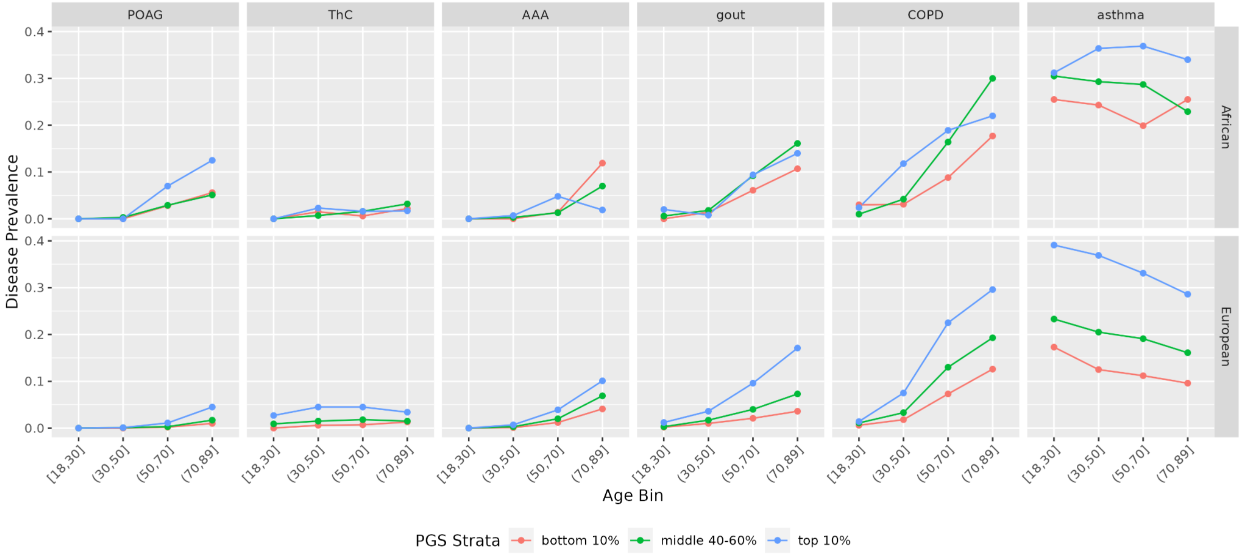

Figure 4. Disease prevalence by age and PGS group. Age calculated as of January 1st 2022 for living participants and at date of death for now-deceased participants. Participants aged > 89 years are masked as 89. POAG: primary open angle glaucoma; ThC: thyroid cancer; AAA: abdominal aortic aneurysm; COPD: chronic obstructive pulmonary disease.

3 Methods

3.1 PGS Construction

Genotyping, quality control, imputation, and ancestry inference for the MGI sample is described elsewhere[9]. We constructed PGS using multi-ancestry leave-MGI-out summary statistics available from the Global Biobank Meta-analysis Initiative (GBMI) as reference (Table S1)[10]. We removed variants flagged by the GBMI as having strand flip or different allele frequency compared to gnomAD before estimating posterior SNP effect sizes using PRS-CS v1.0.0 (auto model)[11,12]. We used the effective sample size of the GBMI genome-wide association study and LD reference panels constructed from UK Biobank data available from the PRS-CS github repository as PRS-CS arguments. We computed posterior SNP effect sizes separately for each ancestry group using an ancestry matched LD reference panel with the exception of using the East Asian LD panel for Central/South Asians and the European LD panel for West Asians. We filtered hard call autosomal genotypes imputed in MGI from the TOPMed reference panel to exclude variants with minor allele frequency < 1% or imputation Rsq < 0.3 before computing PGS using the “–score” function of plink v1.9[13].

3.2 PGS Evaluation

We inferred phenotypes for MGI participants using the createPhenotypes function of the R PheWAS package (v0.99.5-5) using International Classification of Diseases (ICD) 9 and 10 diagnosis codes[14]. We tested the significance of differences in mean PGS between majority genetic ancestry groups using a one-way ANOVA. For each phenotype we tested for association with the PGS by first calculating Cox and Snell R^2 by comparing a base model fit with the phenotype as outcome and covariates for: array + recruiting study + sex + age + first 10 genetic principal components and a full model fit including a covariate for the PGS[15]. We then performed Nagelkerke's modification to Cox and Snell R^2 to calculate the proportion of variance explained by the PGS on the liability scale[16,17]. We calculated p values for Nagelkerke's R^2 based on the difference between the deviance of the base and full model and 95% confidence intervals based on bootstrapping with 1000 replicates. We compared odds ratios between each PGS decile and the bottom decile as reference. We performed all PGS evaluations on a set of participants unrelated to the 2nd degree or closer (KING v2.2.7)[18].

4 Supplementary Information

Phenotype	Ancestry Composition	Cases	Controls	# Training SNPs
POAG	EAS:18%,AFR:1.9%,AMR:0.8%,EUR:79%,SAS:0.3%	45,549	2,433,237	905,804
ThC	EAS:18.2%,EUR:80.5%,AMR:0.7%,AFR:0.6%	9,285	2,616,100	905,845
AAA	EAS:14.9%,AFR:0.3%,EUR:84.4%,AMR:0.4%	14,223	2,324,198	905,792
Gout	EAS:23.7%,AFR:2.2%,EUR:71.5%,AMR:1.2%,SAS:1.5%	55,786	2,490,276	904,645
COPD	EAS:23.2%,AFR:2.2%,EUR:71.8%,AMR:1%,SAS:1.8%	134,499	2,297,115	904,587
Asthma	EAS:20.5%,AFR:2%,EUR:74.1%,AMR:1.1%,SAS:2.2%,MID:0.1%	243,978	2,624,009	904,489

Table S1. PGS training data. Sample size and ancestry composition of the leave-MGI-out summary statistics from the Global Biobank Meta-analysis Initiative (GBMI). Data are adapted from Wang et al.[6] # Training SNPs represents the number of sites that overlap the GBMI summary statistics and the MGI imputation that were used to derive posterior SNP effect sizes.POAG: primary open angle glaucoma; ThC: thyroid cancer; AAA: abdominal aortic aneurysm; COPD: chronic obstructive pulmonary disease;

EUR, European; AMR, admixed American; MID, Middle Eastern; CSA, Central and South Asian; EAS, East Asian; AFR, African.

5 References

1. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* **104**, 21–34 (2019).
2. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219–1224 (2018).
3. Parodi, L. *et al.* Shared genetic background between SARS-CoV-2 infection and large artery stroke. *Int J Stroke* 17474930221095696 (2022) doi:10.1177/17474930221095696.
4. Bond, T. A. *et al.* Exploring the causal effect of maternal pregnancy adiposity on offspring adiposity: Mendelian randomisation using polygenic risk scores. *BMC Medicine* **20**, 34 (2022).
5. Mullins, N. *et al.* Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. *Psychol Med* **46**, 759–770 (2016).
6. Wang, Y. *et al.* Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genomics* **3**, 100241 (2023).
7. Curtis, D. Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr Genet* **28**, 85–89 (2018).
8. Akmatov, M. K. *et al.* Comorbidity profile of patients with concurrent diagnoses of asthma and COPD in Germany. *Sci Rep* **10**, 17945 (2020).
9. Vanderwerff, B. *et al.* Michigan Genomics Initiative Freeze 5 Genome-Wide Genotypes. https://precisionhealth.umich.edu/wp-content/uploads/sites/67/2022/11/freeze5_white_paper.pdf.
10. Wang, Y. *et al.* Global biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. 2021.11.18.21266545 Preprint at https://doi.org/10.1101/2021.11.18.21266545 (2022).

11. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019).

12. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

13. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

14. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).

15. Cox, D. R. & Snell, E. J. *Analysis of Binary Data, Second Edition*. (CRC Press, 1989).

16. NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).

17. Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology* **36**, 214–224 (2012).

18. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).