Association of Polygenic Scores and Disease Prevalence in Global Populations

Hanyu Xiao

September 29, 2025

Abstract

This project investigates the correlation between polygenic scores (PGS) and the prevalence of multiple complex diseases like type 2 diabetes mellitus, chronic kidney disease, and breast cancer, and so on, in diverse human populations.

1 Methods

1.1 PGS Calculation

We downloaded 2504 individual genotype data from the publicly available 1000Genomes Project 2013 release data. The weights for the SNP included were derived from PGS Catalog. The catalog ID is listed in Table 1. The score is calculated using PGS-Calc on 1-22 chromosomes. The Python code is listed in the code file.

Table 1: List of diseases analyzed with their corresponding Polygenic Score (PGS) Catalog IDs.

Cause Name	**PGS ID**
Alzheimer's disease and other dementias	PGS004228
Asthma	PGS00178
Breast cancer	PGS000332
Chronic kidney disease	PGS00223
Chronic obstructive pulmonary disease	PGS00178
Diabetes mellitus type 2	PGS00511
Height	PGS002802
BMI	PGS000027

1.2 Phenotype Collection

Based on the metadata provided, we group the 26 populations into 21 countries and regions based on their ancestry information (see Table 2). This allowed us to align the genetic data with publicly available phenotype data from the Global Burden of Disease (GBD) database and gathered the disease prevalence values for each corresponding country and region at the year of 2013. Height is extracted from wiki-height) and BMI data is from wiki-BMI The R code is listed in the code file.

We have re-categorized the population affiliations that were unclear in the previous version. The 1000 Genomes populations ACB, ASW, and CEU were difficult to classify by ancestry. Our original approach used an "African union" or "European region", which was a custom definition that may have led to inconsistent differences in disease data. This updated version now uses the specific countries where these populations were sampled. The prevalence rate metric is the number of cases per 100,000 people.

- ACB stands for "African Caribbean in Barbados." We labeled the country as Barbados.

- ASW stands for "African Ancestry in Southwest USA." We labeled the country as the United States of America.

- CEU refers to "Utah residents with Northern and Western European ancestry". We labeled the country as the United States of America.

Table 2: Mapping of 1000 Genomes Project Population Codes to Super-populations and Countries/Regions.

Population Code	**Super-population**	**Country/Region**
ACB	AFR	Barbados
ASW	AFR	United States of America
BEB	SAS	Bangladesh
CDX	EAS	China
CEU	EUR	United States of America
CHB	EAS	China
CHS	EAS	China
CLM	AMR	Colombia
ESN	AFR	Nigeria
FIN	EUR	Finland
GBR	EUR	United Kingdom
GIH	SAS	India
GWD	AFR	Gambia
IBS	EUR	Spain
ITU	SAS	India
JPT	EAS	Japan
KHV	EAS	Viet Nam
LWK	AFR	Kenya
MSL	AFR	Sierra Leone
MXL	AMR	Mexico
PEL	AMR	Peru
PJL	SAS	Pakistan
PUR	AMR	Puerto Rico
STU	SAS	Sri Lanka
TSI	EUR	Italy
YRI	AFR	Nigeria

1.3 Phenotype Modeling

Note of Caution. At the time of this analysis, I did not have a full understanding of the PGS score calculator we use and the confidence in the statistical assumptions, and I was

not sure whether the chosen approaches were fully appropriate for modeling prevalence data. The results should therefore be interpreted with caution, and I would welcome further review of the methodology.

We investigated the relationship of PGS vs. quantitative traits with linear models and PGS vs. disease prevalence with logistic regression models. Disease prevalence was obtained as rates per 100,000 individuals, which were converted into probabilities for modeling.

1.3.1 Quantitative Traits Modeling

For quantitative traits (height, BMI), we applied a simple linear regression model. The phenotype values were standardized within each population, and the mean polygenic score (PGS) was used as the predictor:

$$Y = \beta_0 + \beta_1 \cdot \mathrm{PGS}_z,$$

where Y is the standardized quantitative phenotype, and PGS_z is the z-scored mean polygenic score for each population. The coefficient β_1 directly represents the expected change in phenotype (in standard deviation units) per +1 SD increase in PGS.

This approach provides a straightforward measure of association between PGS and continuous traits, without transformation of the outcome variable.

1.3.2 Disease Prevalence Modeling

Prevalence-to-probability. For each country c, reported prevalence per 100,000 individuals was converted to a probability

$$p_c = \frac{\text{cases per } 100{,}000}{100{,}000}.$$

Constructing binomial counts. Let n_c denote the country-specific sample size contributing to the regression. We formed grouped-binomial outcomes by scaling p_c to the sample size:

$$\mathrm{cases}_c = \lfloor n_c \cdot p_c \rceil, \qquad \mathrm{fails}_c = n_c - \mathrm{cases}_c,$$

where $\lfloor \cdot \rceil$ indicates rounding to the nearest integer.[1]

Predictor. For each country, we computed the standardized mean polygenic score (PGS) at the country level, $\mathrm{PGS}_{z,c}$, by z-scoring the country means.

Model specification. We fit a binomial generalized linear model (GLM) with a logit link to relate prevalence to PGS:

$$\mathrm{logit}(p_c) = \beta_0 + \beta_1 \cdot \mathrm{PGS}_{z,c}.$$

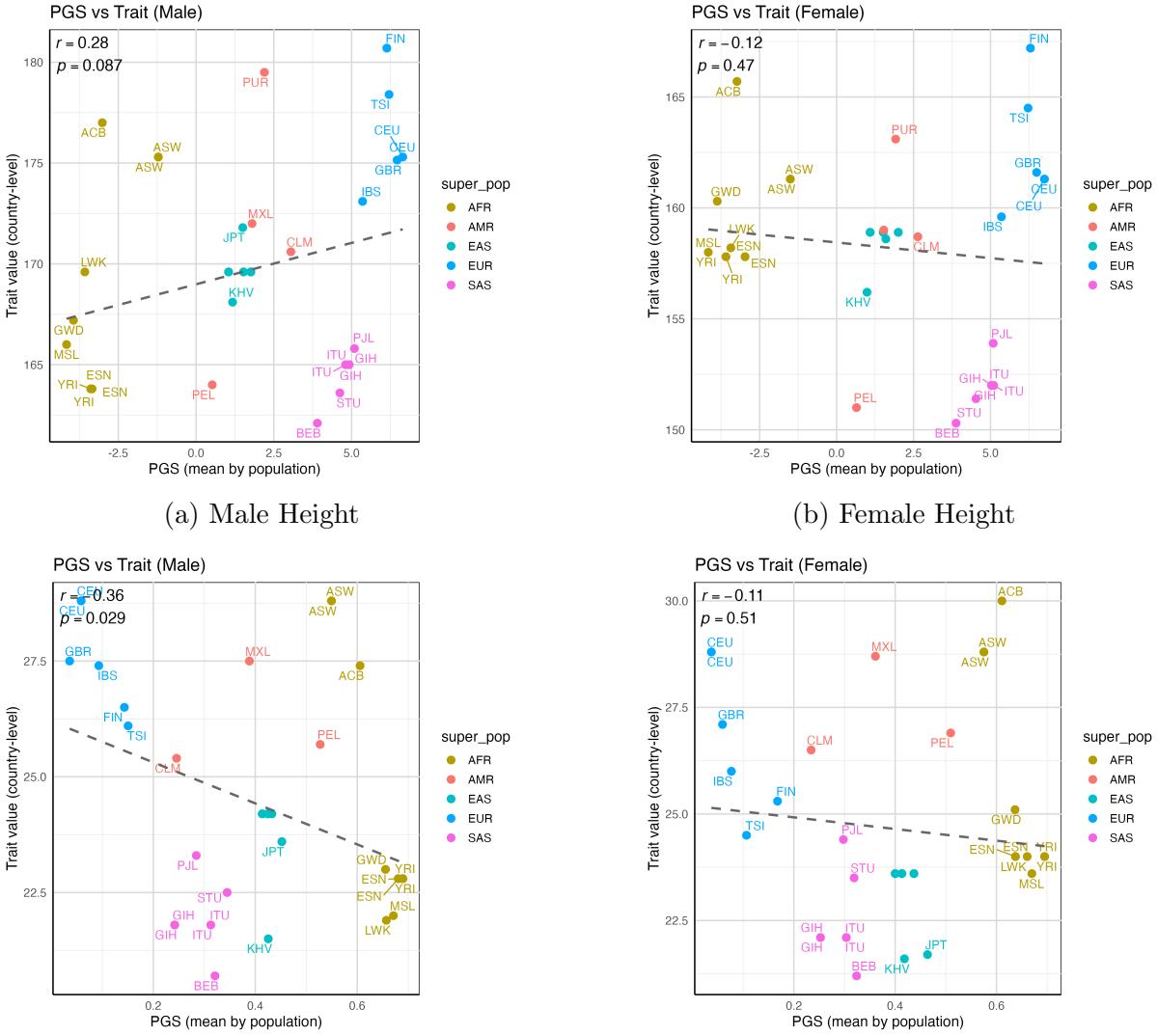
In implementation, the grouped-binomial outcome was provided as $\mathrm{glm}\big(\texttt{cbind}(\mathrm{cases}_c, \mathrm{fails}_c) \sim \mathrm{PGS}_{z,c},\ \mathrm{family} = \mathrm{binomial}(\mathrm{link} = \mathrm{logit})\big)$.

[1] To ensure valid binomial inputs, we truncated p_c to $(\varepsilon, 1-\varepsilon)$ if needed (with a small ε), and required $n_c \geq 1$.

2 Results

2.1 Quantitative Traits

In addition to the diseases, the relationship between mean population polygenic scores and country-level data for quantitative traits such as height and BMI is shown in Figure 1.



(a) Male Height

(b) Female Height

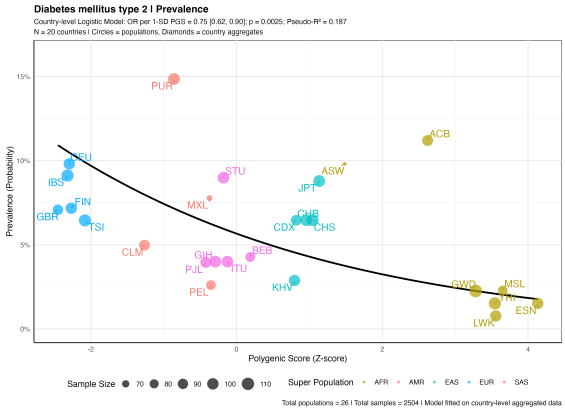
(c) Male BMI

(d) Female BMI

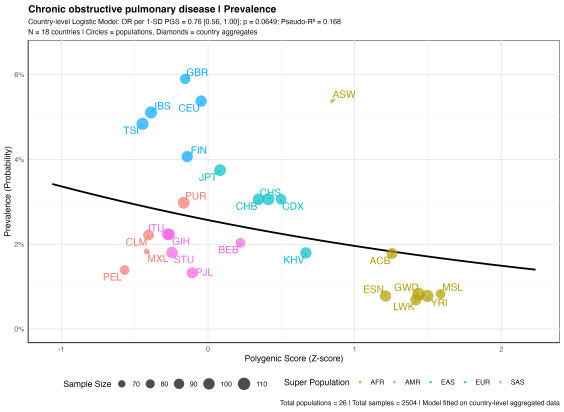
Figure 1: Scatter plots of mean population polygenic scores versus country-level data for quantitative traits, separated by gender. The top row shows the relationship for height, while the bottom row displays the relationship for BMI. Different colors represent different super-populations.

2.2 Disease Prevalence

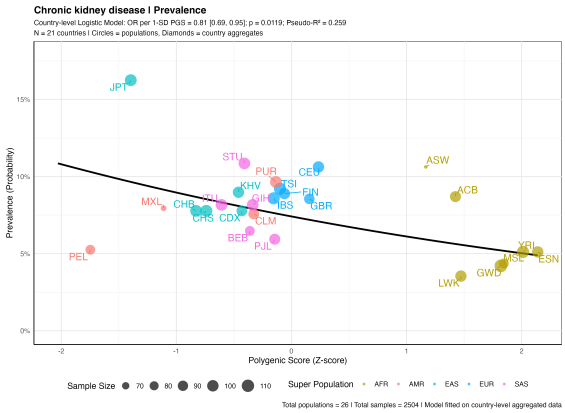
The results for the selected diseases are presented in Figure 2, which combines the individual scatter plots for each disease.



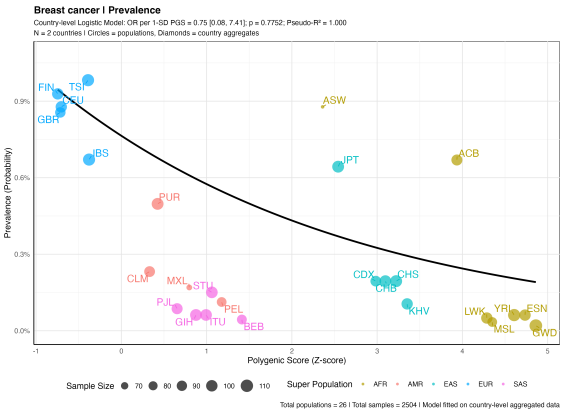
(a) Type 2 Diabetes Mellitus



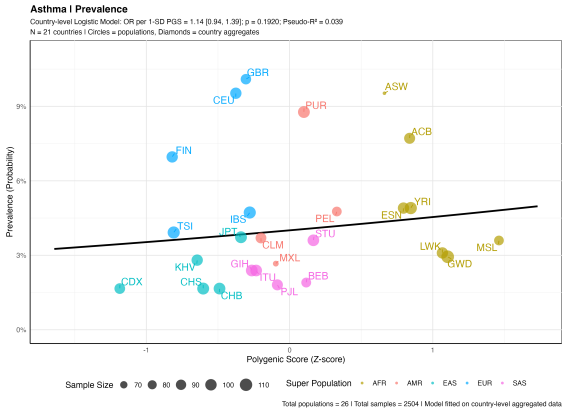
(b) Chronic Obstructive Pulmonary Disease



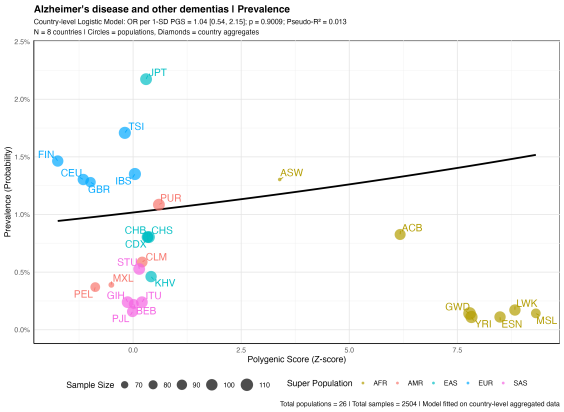
(c) Chronic Kidney Disease



(d) Breast Cancer



(e) Asthma



(f) Alzheimer's Disease

Figure 2: Scatter plots of PGS vs logit-transformed prevalence with fitted linear regression lines.

3 Discussion

Why do some traits have only positive / negative mean PGS values?

- Some traits show only positive or negative mean polygenic score (PGS) values due to choices in centering and allele coding. If scores are not centered to the discovery cohort (e.g., by subtracting $2p_{\text{ref}}$ at each SNP), the raw sum $\sum \beta_i G_i$ can have a non-zero mean, resulting in consistently positive or negative values across groups.

Additionally, allele flip errors, uneven SNP coverage, and GWAS meta-analysis offsets may introduce baseline shifts. In this project, we chose European-centric PGS, so there will likely be a systematic shift in PGS scores in other populations, which indicates that this population may have a higher/lower disease prevalence rate (positive/negative) than European populations. But this needs further analysis and validation.

Why do AFR samples always have the highest mean PGS?

- AFR samples often show the highest mean PGS because of allele frequency differences between AFR and EUR populations. If centering is performed using EUR allele frequencies, AFR samples naturally deviate upward or downward due to frequency mismatches. From an evolutionary perspective, the Out-of-Africa model predicts that African populations retain greater genetic diversity and older haplotypes, whereas non-African groups experienced founder effects and drift during serial bottlenecks after dispersal. These demographic histories lead to systematic differences in allele frequencies and LD patterns across ancestries, which can shift cross-ancestry PGS means when scores are centered on the European population.

Why didn't we see the positive correlation as expected?

- **Simplified PGS pipeline.** In this project we computed PGS using a relatively simple pipeline and did not apply downstream corrections or standardization. This may leave residual offsets across groups and reduce the comparability of score scales.
- **EUR-based weights and portability.** We used EUR-based PGS weights; limited cross-ancestry portability due to LD and MAF differences can attenuate signal or introduce systematic mean shifts when applying the score to non-EUR populations.
- **Prevalence rates vs. small n_c.** Although disease prevalence was sourced as rates per 100,000, several countries contributed relatively small total sample sizes n_c to the regression step. Our grouped-binomial counts (*cases*/*fails*) were derived by scaling these rates to n_c, so the resulting country-level prevalence used for modeling may be noisy and not fully rigorous for small n_c.

Together, these factors can weaken or obscure the expected correlation between mean PGS and population-level prevalence.

References

[1] Duncan, L. et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* **10**, 3328 (2019).

[2] Duschek, E. et al. A polygenic and family risk score are both independently associated with risk of type 2 diabetes in a population-based study. *Sci Rep* **13**, 4805 (2023).