Association of Polygenic Scores and Disease Prevalence in Global Populations

Hanyu Xiao

September 19, 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	PGS00403
Asthma	PGS00178
Breast cancer	PGS00404
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.

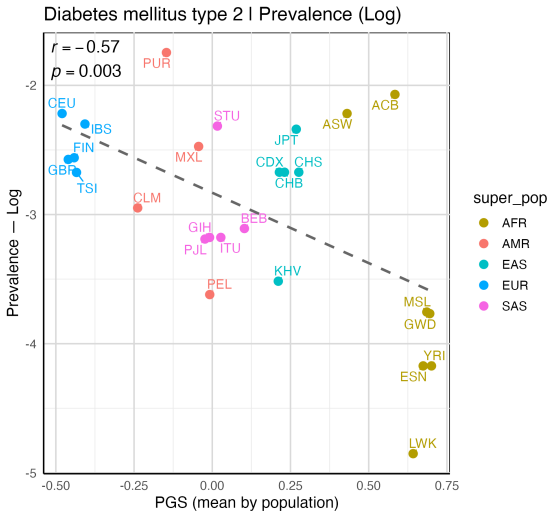
Table 2: Mapping of 1000 Genomes Project Population Codes to Super-populations and Countries/Regions.

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

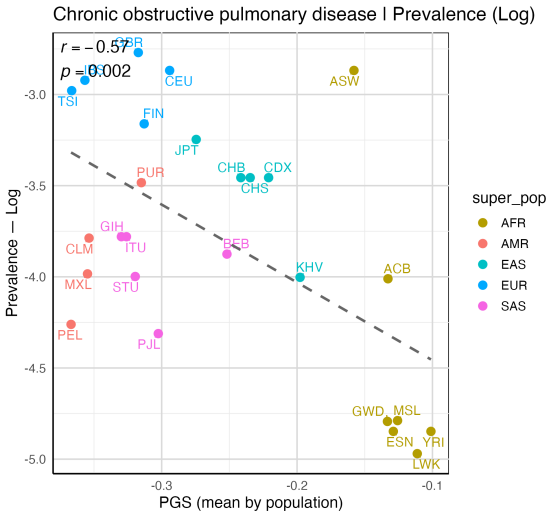
2 Results

2.1 Disease Prevalence

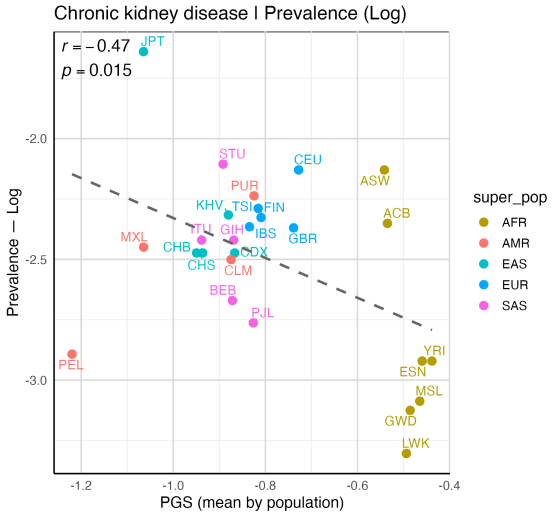
The results for the selected diseases are presented in Figure 1, which combines the individual scatter plots for each disease.



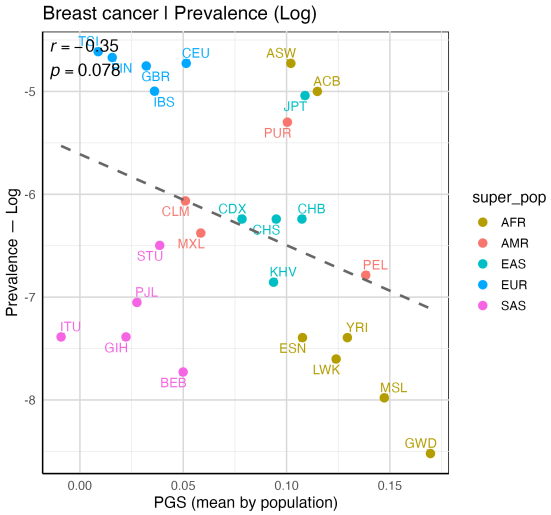
(a) Type 2 Diabetes Mellitus



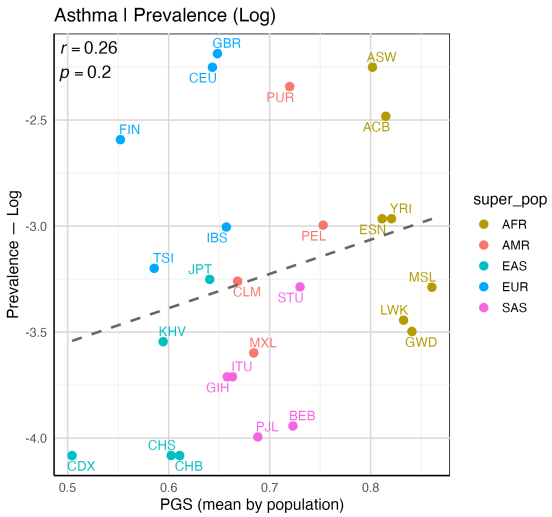
(b) Chronic Obstructive Pulmonary Disease



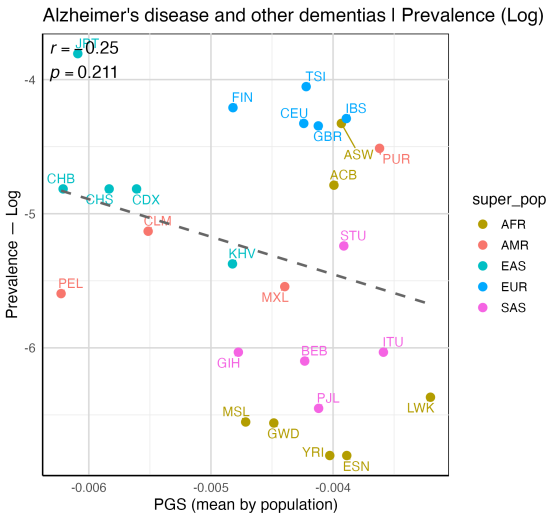
(c) Chronic Kidney Disease



(d) Breast Cancer



(e) Asthma

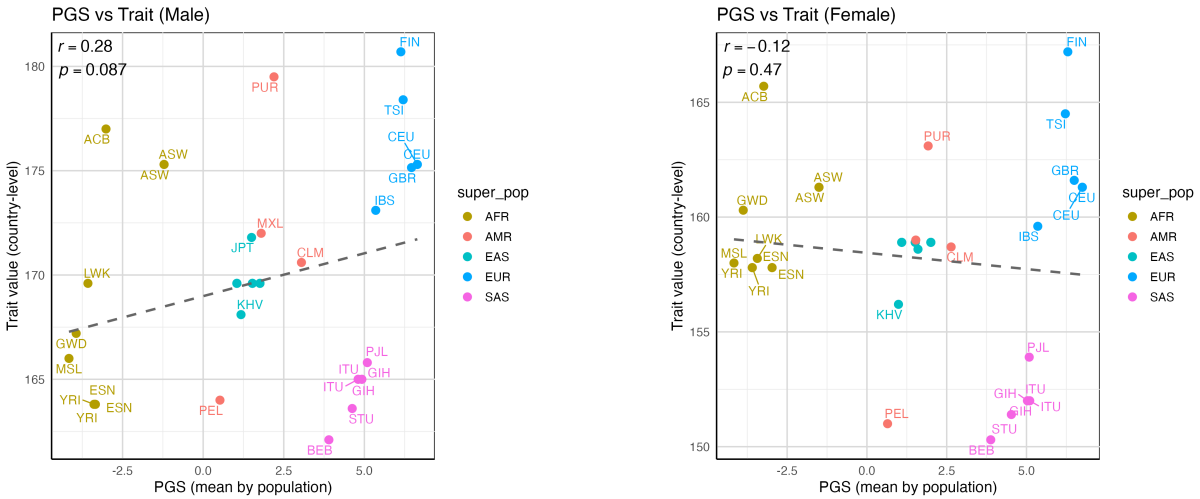


(f) Alzheimer's Disease

Figure 1: Scatter plots of mean population polygenic scores versus country-level prevalence data for multiple diseases. Each subplot shows the relationship for a specific disease, with different colors representing different super-populations.

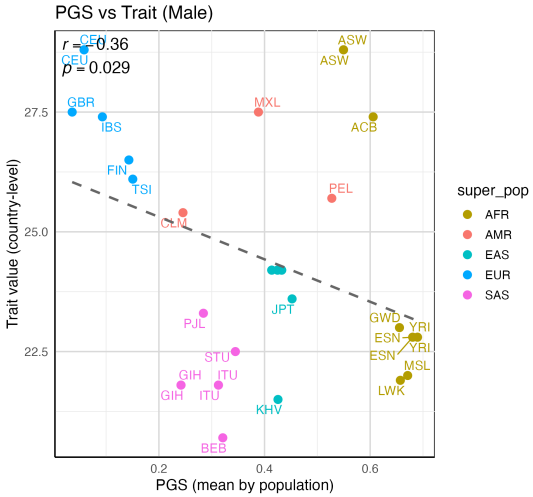
2.2 Quantitative Traits

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 2.

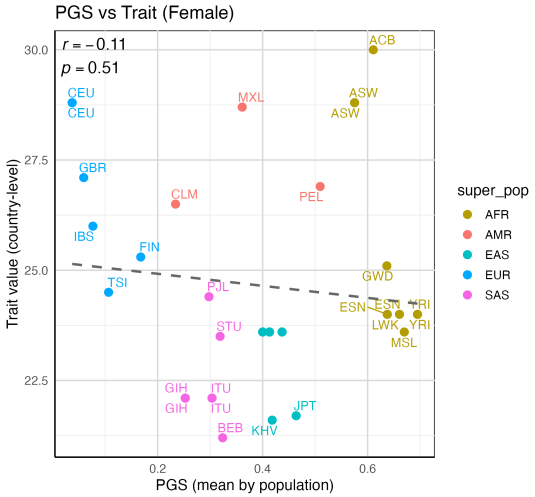


(a) Male Height

(b) Female Height



(c) Male BMI



(d) Female BMI

Figure 2: 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.

3 Discussion

1. Why do some traits have only positive / negative mean pgs values?
2. Why do AFR samples always have the highest mean pgs?
3. Why didn't we see the positive correlation as expected?

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).