GGSB PRELIM QUESTION # 3

Question:
It is often argued that one potential utility of GWAS is to develop the ability to accurately predict an individual's disease risk from their genotype/genome sequence, using so called "polygenic scores" (PGS). Significant recent attention has focused on factors which may cause PGS to be more accurate for some individuals than others, or in some cases to be downright misleading. Specifically, prediction accuracy of a PGS is typically best in groups closely resembling the GWAS sample from which it was constructed, with accuracy declining when applied to less similar groups. This is often described as the issue of PGS "portability".

A) Assume that you have been given a panel of individuals (the "GWAS sample") who have all been measured for a particular quantitative/complex phenotype of interest, as well as relevant covariates (age, sex, etc.), and that all individuals have been genotyped across a common set of genetic markers. Assume that you have also been given genotypes, covariates, and phenotypes for a second panel of individuals (the "prediction sample"). Assume that the prediction sample is very similar to the GWAS sample (in terms of ancestry, environment, and other relevant factors), such that portability is not a concern. Explain how you could use the GWAS sample and the prediction sample together to develop a polygenic score that was maximally predictive among individuals of ancestral and environmental backgrounds resembling those of the GWAS and prediction samples. Try to be as detailed as possible about each step along the way, but do not worry about specific pieces of software one might use in each step, but rather on what is accomplished in each step and why. As part of your answer, write down the equation for an individual's polygenic score.

B) Identify major factors (besides stratification, see below) which could be responsible for low portability of PGS among groups.

C) Perhaps one of the most serious issues we deal with in GWAS is environmental stratification, which occurs when an environmental variable that affects the phenotype is correlated with one or more major axis of variation in ancestry within the GWAS sample. Early in the GWAS era (i.e. mid-2000s) there was a great deal of concern that stratification could (and occasionally did!) lead genome wide significant associations that were in fact false positives generated by stratification.

Since then, an array of statistical methods have been developed to control for stratification, and sampling designs have been updated to minimize its impact. The current status in the field is that two seemingly contradictory observations coexist:
1) It is typically agreed that the majority of genome wide significant hits from well powered GWAS represent real associations.

2) biases due to stratification remain a potentially significant problem for the application of PGS, even when the PGS has high predictive value in samples closely related to the GWAS sample.

Explain how the two observations above can both simultaneously be true?

Useful References:

Predicting Polygenic Risk of Psychiatric Disorders
https://www.sciencedirect.com/science/article/pii/S000632231832119X

Pitfalls of predicting complex traits from SNPs
https://www.nature.com/articles/nrg3457

Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland
https://www.sciencedirect.com/science/article/pii/S0002929719301879