Understanding Model Misspecification in Polygenic Risk Score (PRS) Methods

Xiaoyu Wang

Polygenic risk scores (PRS) are widely used in clinical genetics and show strong potential for personalized medicine. Many methods have been developed to construct PRS from large genome-wide association studies (GWAS) using multiple linear regression frameworks. However, because individual-level data are often unavailable due to privacy constraints, researchers must rely on GWAS summary statistics. In this setting, the linkage disequilibrium (LD) matrix becomes central for modeling SNP correlations. A mismatch between the assumed and true LD structures can lead to model misspecification—where the assumed model cannot fully represent the true data-generating process (DGP).

Most PRS models assume correct specification, yet this is rarely true in practice. Misspecification may reduce prediction accuracy or produce misleading inference. Existing strategies attempt to mitigate its effects through quality control or by integrating additional biological and population information. We define compatibility as the ability of the assumed model to generate data consistent with observations; incompatibility occurs when this assumption fails.

We distinguish two types of misspecification: fundamental and practical. Fundamental misspecification stems from intrinsic modeling limitations such as using an external LD reference panel that differs from the target population or combining inconsistent GWAS data. Many Bayesian approaches assume the model approximates the true DGP, but when sample sizes are small, priors may dominate inference; when large, the likelihood dominates, and misspecification in the likelihood can bias results. Practical misspecification arises from computational or numerical approximations, including algorithmic simplifications or the use of approximate LD matrices. These issues can appear as convergence failures rather than explicit model errors.

Among all sources, mismatched LD reference matrices remain the primary driver of misspecification. Recent research emphasizes robust likelihood formulations that maintain validity even when LD panels are imperfect. Projection-based corrections and rescaled likelihoods improve stability but are often computationally demanding.

In this study, we propose a robust and computationally efficient likelihood-based approach that directly addresses LD mismatch. By connecting theoretical analysis with simulations where the true DGPs are known, we demonstrate a clear link between Kullback–Leibler (KL) divergence and prediction accuracy, providing a principled understanding of model misspecification in PRS frameworks.