2 Training-Validation-External Generalization Workflow

From a statistical inference perspective, the construction and evaluation of PRS/PGS should follow a staged process of estimation-regularization-prediction-validation, with the core objective of obtaining a stable and interpretable predictive function for the target population while avoiding overfitting. This workflow not only involves data partitioning and model selection, but also directly relates to the definition of the statistical target (estimand) and its consistency across different data domains.

2.1 Data splitting and internal validation

To ensure unbiased model development and evaluation, a three-stage framework is typically adopted: training, validation, and testing sets. The training set is used for effect estimation and model fitting; the validation set is used for hyperparameter tuning and recalibration (e.g., p-value thresholds, shrinkage strength, and global scaling factors); and the test set is reserved for one-time performance evaluation (Sima et al., 2024). This process corresponds to the "effect size → shrinkage → PRS" stage in the statistical inference pipeline and represents the key point at which model bias enters.

In both human genetics and crop breeding contexts, particular attention must be paid to sample structure dependence. Related individuals (e.g., families, lines, or close relatives) should be assigned to the same data split to avoid information leakage. At the same time, the distribution of phenotypes and key covariates (e.g., sex, batch effects, ancestry principal components) should be comparable across splits. For studies with limited sample sizes, nested cross-validation or leave-group-out strategies (e.g., by population, environment, or experimental site) can improve estimation stability and reduce selection bias (Lennon et al., 2024).

In the preprocessing stage, genotype-phenotype harmonization must be strictly reproducible, including alignment of reference alleles and genomic coordinates, removal of ambiguous variants, and the use of LD reference panels and allele frequency estimates matched to the target population. Different methods reflect distinct modeling assumptions in their tuning strategies: C+T relies on grid search over p-value thresholds and LD parameters, whereas LD-aware and Bayesian approaches adjust prior strength or LD block structure for shrinkage estimation (Wang et al., 2023; Sima et al., 2024). During validation, only linear recalibration (e.g., slope and intercept adjustment) should be permitted. Once the model is frozen, any form of re-tuning ("information leakage") must be strictly avoided to ensure independence of testing and external evaluation.

2.2 External generalization and cross-population evaluation

External generalization is the core step in evaluating PRS portability. It is fundamentally a statistical inference problem under domain shift, assessing whether a predictive function estimated in the training data can maintain stable performance across different ancestries, environments, or technical platforms (Ruan et al., 2021). This stage corresponds to the "PRS → prediction" step and represents the primary point where cross-population failure occurs.

The standard workflow includes independent quality control and allele alignment for external datasets, selection of appropriate LD reference panels based on principal component analysis or ancestry inference, and computation of PRS on a consistent scale. Performance evaluation should then be conducted independently in the external dataset, including discrimination, calibration, and utility assessment, while documenting differences in phenotype definitions and measurement error (Kachuri et al., 2024). Numerous studies have shown that PRS trained in European populations typically experience a 40-60% reduction in predictive accuracy when applied to non-European populations. Multi-ancestry training or methods (e.g., PRS-CSx, CT-SLEB) can partially recover performance, in some cases reaching approximately 80% or more of the source population performance (Duncan et al., 2019; Jung et al., 2025).

From a statistical perspective, this performance decay can be understood as an estimand mismatch caused by differences in LD structure, allele frequency spectra, and effect size distributions across populations. To disentangle structural differences from environmental effects, it is recommended to design multiple external