In human studies, environmental exposures are often time-dependent (e.g., smoking, diet, physical activity), and can be modeled using logistic regression or survival analysis:

$$h(t \mid PGS, E(t)) = h_0(t)\exp\left(\beta_g \cdot PGS + \beta_e \cdot E(t) + \beta_{ge} \cdot PGS \cdot E(t)\right)$$

Within a statistical modeling framework, the incorporation of G×E effects extend the predictive function from one that depends solely on genetic information to one that is explicitly conditioned on environmental context. In this setting, PRS no longer represents a marginal genetic effect averaged across environments; rather, it acts as a modifier of environmental effects, allowing the relationship between environmental exposure and phenotype to vary systematically across different levels of genetic risk. From this perspective, gene-environment interaction can be understood as heterogeneity in environmental response patterns across strata of polygenic scores, whereby the functional form of the environment-phenotype relationship becomes dependent on PRS. Consequently, prediction shifts from a marginal formulation to a conditional, context-dependent representation (Figure 2).

However, this process is susceptible to multiple sources of bias, including measurement error in environmental variables, time misalignment, and confounding induced by gene-environment correlation (rGE). Therefore, stratified analyses, instrumental variable approaches, or negative control designs are recommended, along with replication in independent cohorts to ensure robustness of interaction effects (Kachuri et al., 2024; Sima et al., 2024).

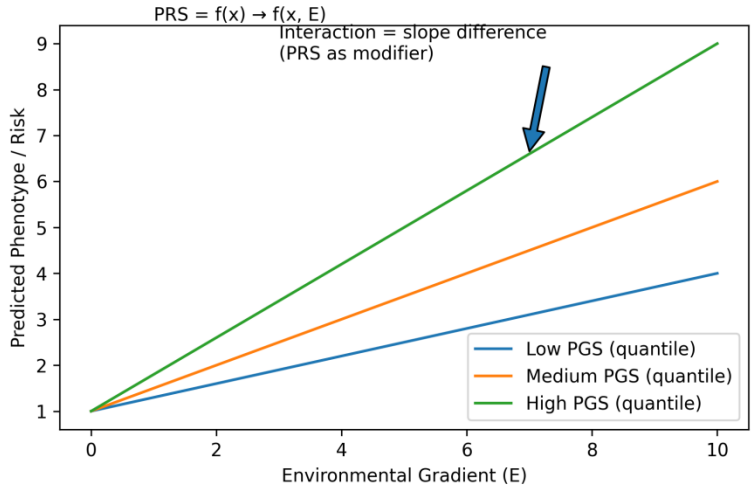

Figure 2 Conditional prediction under gene-environment interaction (G×E): PRS as a function modifier

Note: Illustration of gene-environment interaction (G×E) as differences in slopes across environmental gradients. Each line represents a quantile of polygenic scores (PGS), showing that the effect of environmental exposure on phenotype depends on genetic background. This reflects a conditional predictive function, where PRS modifies the relationship between environment and outcome (PRS = f(x) → f(x, E)).

3.2 Statistical and machine learning models

Within traditional statistical frameworks, joint models are typically formulated as linear or generalized linear models incorporating genetic, environmental, and interaction effects:

$$\eta = \alpha + \beta_g \cdot PGS + \beta_e^{\top} E + \gamma^{\top}(PGS \circ E) + C^{\top}\theta$$

where C represents covariates such as age, sex, ancestry principal components, and batch effects.

In high-dimensional settings, to prevent overfitting, regularization methods such as ridge regression, elastic net, or group lasso are commonly used to impose hierarchical shrinkage on interaction terms, balancing model complexity and predictive stability. In MET designs, leave-group-out cross-validation (e.g., by site or year) should be employed to prevent environmental information leakage.