

# Multi-trait and Trans-ethnic PRS Models for Complex Traits and Diseases

NUU-SSU G-LAMP International Workshop  
October 2nd, 2026 (Fri)

Department of Statistics and Actuarial Science, Soongsil University  
Wonil Chung, Ph.D.

# Lab Introduction

# Biostatistics and Data Science Lab



# Research Objectives

## Developing Multi-ethnic PRS methods and Conducting Meta GWAS for Multiple Populations

### [Research Objective 1] Developing Multi-ethnic PRS Method using Penalized and Bayesian Approach

- Improve Multi-ethnic PRS by integrating GWAS summary statistics from diverse populations
- Extend CTPR to develop cross-ethnic PRS that leverage LD difference across multiple populations

### [Research Objective 3] Development of MR Models for Causal Association with Survival Time

- Develop novel Survival MR methods to identify diseases that causally affect survival time
- Compare Cox models and additive models



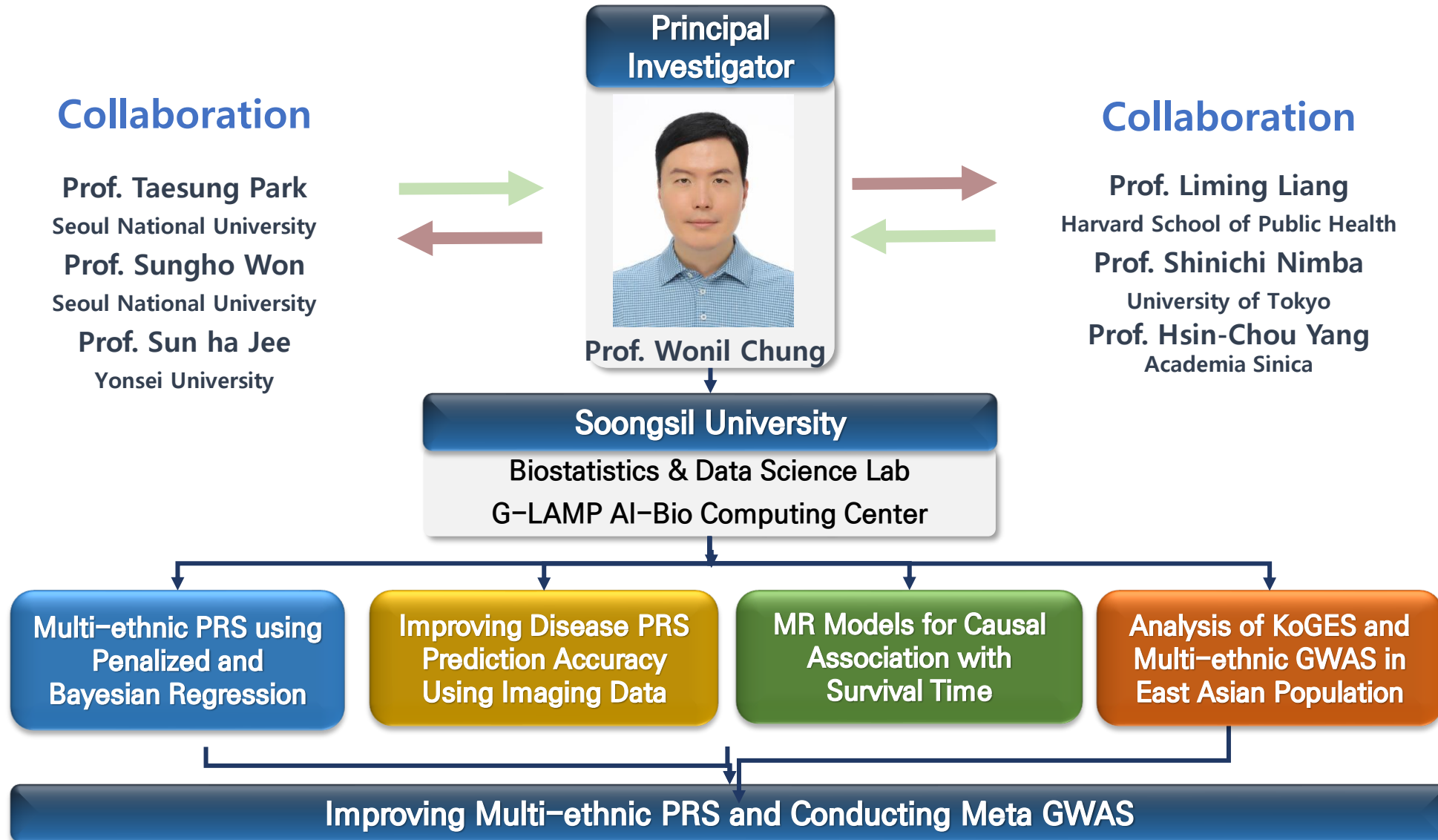
### [Research Objective 2] Improving Polygenic Prediction Accuracy Using Imaging Data

- Develop PRS methods using deep-learning approaches such as Transformers and transfer learning
- Improve disease PRS prediction accuracy by incorporating MRI imaging data

### [Research Objective 4] Analysis of 161K KoGES Data and Multi-ethnic GWAS in East Asian Populations

- Analyze 161K KoGES data for 125 phenotypes and conduct meta-analysis with KCPS-II data
- Conduct East Asian population GWAS and multi-ethnic GWAS

# Collaboration



# Outline

## <Part I> Multi-trait and Trans-ethnic PRS Methods

- Background and Introduction
- Methods and Results
- Simulation Study
- Discussion and Conclusion

## <Part II> Survival MR Methods and Real Data Analysis

- Background and Introduction
- Causal Effect Estimation via CoxPH and Additive Models
- Survival MR for Gene Expression on Prognosis in Colorectal Cancer

# Part I: Multi-trait and Trans-ethnic PRS methods

# Introduction

# Introduction

- Polygenic Risk Scores (PRS)
  - PRS represents an individual's **genetic susceptibility** to complex traits and diseases by aggregating the effects of numerous SNPs.
  - It has effectively quantified genetic risk for diseases such as T2D, CAD, and asthma.
- Challenge of Ancestral Bias
  - Most PRS models are developed using European (EUR) populations, leading to a significant loss of predictive power in non-EUR groups.
  - This discrepancy is driven by differences in **LD patterns** and **allele frequencies** across distinct ethnic backgrounds.

# Introduction

- Multi-trait PRS Approach
  - This strategy leverages GWAS summary statistics from multiple correlated traits to enhance predictive power.
- Trans-ethnic PRS Approach
  - This method utilized various ethnic GWAS data to build better predictive models.
- Selection Priority Criteria
  - The primary objective is to develop a new criterion that leverages both **multi-trait** and **trans-ethnic** strategies.
  - This approach utilizes only correlated traits and ethnicities based on their **genetic correlation** to achieve better predictive power of PRS.

# Background

# Overview of PRS

- Heuristic PRS Methods

- PRS are calculated as  $PRS = \sum_j X_j \beta_j$  for the genotype  $X_j$  and the effect size  $\beta_j$  for  $SNP_j$ .
- Early PRS methods focused on a limited number of GWAS significant SNPs
- LD Clumping + P-value Thresholding (**CT**) approach was introduced to improve accuracy by filtering linked SNPs based on genotypic correlations ( $r^2$ ) and P-values.

- Sophisticated PRS Methods

- Bayesian methods (e.g., **LDpred**, **PRS-CS**) and penalized regression (e.g., **lassosum**) have emerged to better account for genetic architecture.
- Individual-level genotype data can also be utilized through **GBLUP** method for more refined risk estimation.

# Overview of PRS

- Data Availability and Performance Disparity
  - The growth of large-scale biobanks (**UKB, BBJ, TWB, CKB, KoGES**) has increased the availability of GWAS summary statistics.
  - As of 2025, approximately **77.7%** of GWAS participants are of European (EUR) ancestry and 20.6% are of East Asian (EAS) ancestry, showing a heavy data bias.
  - Despite this bias, many GWAS significant SNPs identified in EUR populations are replicable in non-EUR groups, suggesting shared genetic components.

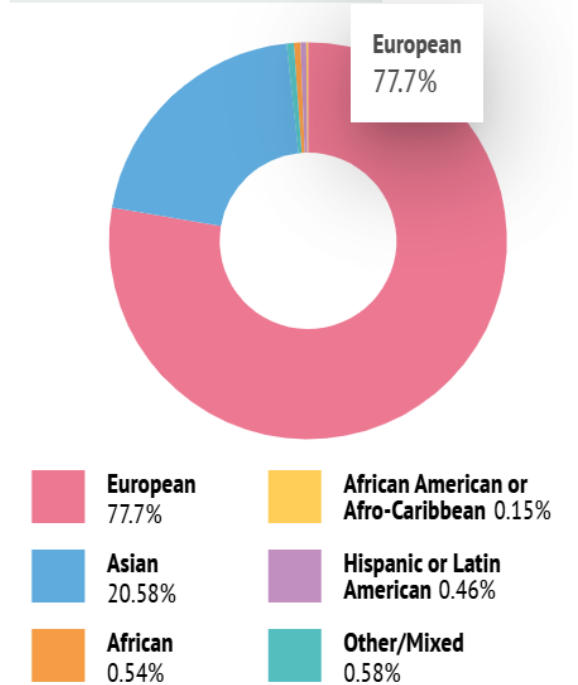
## GWAS Diversity Monitor

### Participants by ancestry

Discovery Stage

Click to show associations discovered

All parent terms



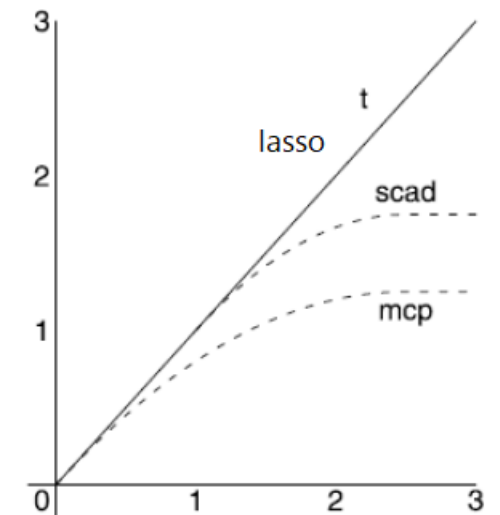
« < 2025 > »

# Multi-trait PRS

- **MTAG** (Turley et al. Nat Genet 2018) and **wMT-SBLUP** (Maier et al. Nat Comm 2018)
  - **MTAG** (Multi-Trait Analysis of GWAS) leverages the generalized method of moments (GMM), assuming SNPs share a variance-covariance structure across traits.
  - **wMT-SBLUP** (Weighted Multi-Trait Summary Statistics BLUP) extends GBLUP model to summary data.
- **MPS** (Krapohl et al. Mol Psychiatry 2018) and **PRSsum** (Kurniansyah et al. Nat Comm 2022)
  - **MPS** (Multi-Polygenic Score) uses multiple PRSs as independent predictors in joint models like Elastic Net.
  - **PRSsum** approach simplifies the process by summing individual PRSs with equal weights to form a single, representative genetic index.

# Multi-trait PRS

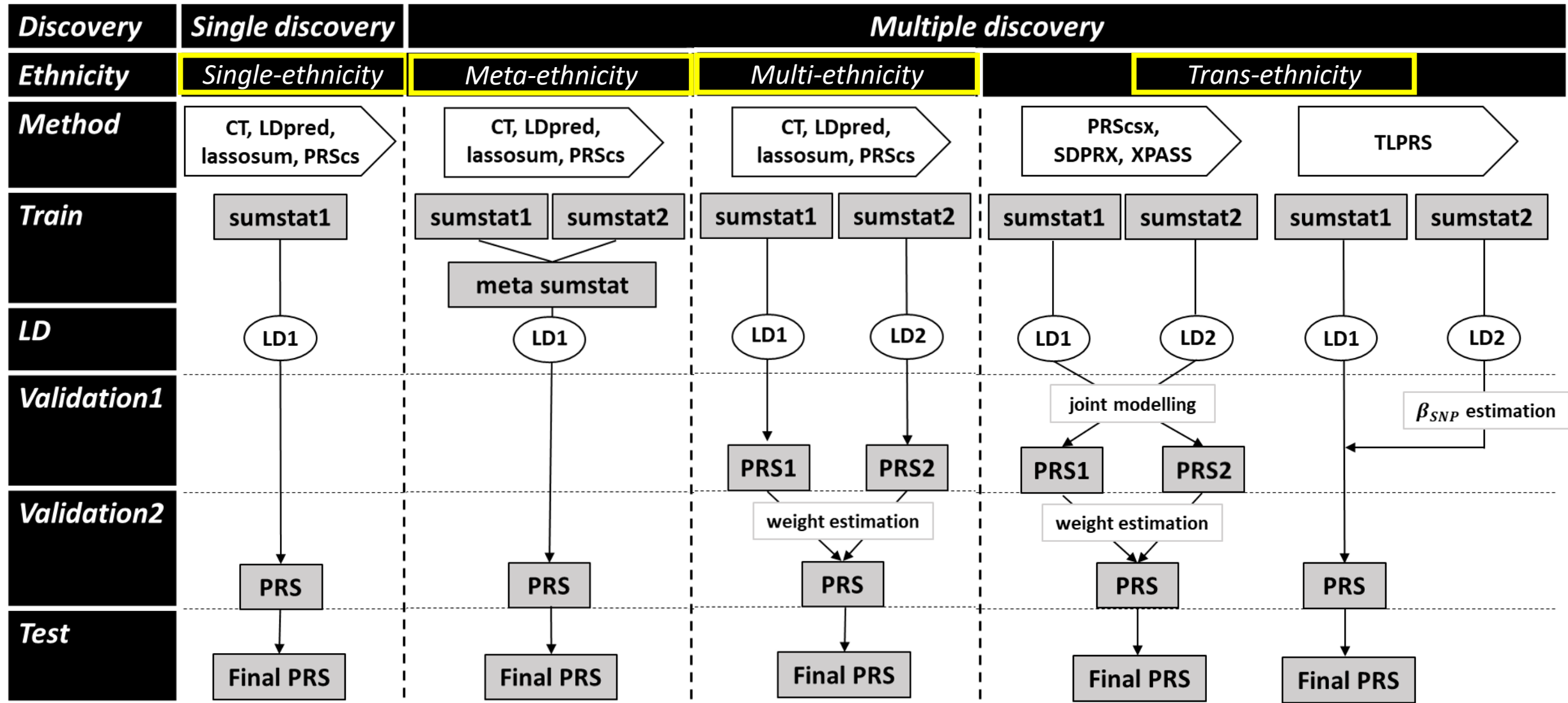
- **MT-GBLUP** (Maier et al. AJHG 2015) and **CTPR** (Chung et al. Nat Comm 2019) and
  - **MT-GBLUP** (Multi-Trait GBLUP) utilizes multiple traits to improve prediction accuracy using individual-level genotype data.
  - **CTPR** (Cross-Trait Penalized Regression) incorporate shared effects through penalty functions to account for correlations between traits.
  - $y_{ki} = \alpha_k + \sum_{j=1}^P x_{kij} \beta_{kj} + \epsilon_{ki}$ ,  $\epsilon_{ki} \sim N(0, \sigma_k^2)$
  - $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} RSS(\beta) + p_{\lambda_1}^{sp}(\beta) + p_{\lambda_2}^{ctp}(\beta)$ 
    - Lasso penalty:  $p_{\lambda_1}^{sp}(\beta) = \lambda_1 \sum_{k=1}^K \sum_{j=1}^P |\beta_{kj}|$
    - Minimax concave penalty (MCP):  $p_{\lambda_1, \gamma}^{sp}(\beta) = \lambda_1 \sum_{k=1}^K \sum_{j=1}^P \int_0^{|\beta_{kj}|} \left(1 - \frac{x}{\gamma \lambda_1}\right)^+ dx$
    - Cross-trait penalty:  $p_{\lambda_2}^{ctp}(\beta) = \frac{\lambda_2}{2} \sum_{k \neq k'}^K \sum_{j=1}^P a_{kk'j} (\beta_{kj} - \beta_{k'j})^2$
    - Summary statistics:  $p_{\lambda_2}^{ctp'}(\beta) = p_{\lambda_2}^{ctp}(\beta) + \frac{\lambda_2}{2} \sum_{k=1}^K \sum_{m=1}^M \sum_{j=1}^P a_{k(K+m)j} (\beta_{kj} - \hat{s}_{mj})^2$



# Multi-ancestry PRS

- Four Strategic Approaches for Multiple Populations
  - **Single-ethnic PRS:** Uses population-specific data to construct PRS optimized for a specific ancestry.
  - **Meta-ethnic PRS:** Uses meta-GWAS summary statistics that pool associations from both EUR and non-EUR ancestries to compute PRS.
  - **Multi-ethnic PRS:** Constructs individual PRSs from population-specific SNP effect sizes and combines them through a weighted average.
  - **Trans-ethnic PRS:** Jointly models EUR and non-EUR summary statistics using advanced PRS methods (e.g., PRS-CSx, XPASS, SDPRX, TL-PRS) to refine effect sizes.

# Multi-ancestry PRS



# Trans-ethnic PRS

- **PRS-CSx**: PRS with continuous shrinkage prior (Ruan et al. NatGenet, 2022)
  - PRS-CSx, an extension of PRS-CS with continuous shrinkage priors for cross-ethnicity.
  - For SNP effect size  $\beta_{jk}$  for SNP  $j$  in **population**  $k$ ,

$$\beta_{jk} \sim N\left(0, \frac{\sigma_k^2}{N_k} \Psi_j\right), \Psi_j \sim \text{Gamma}(a, \delta_j), \delta_j \sim \text{Gamma}(b, \phi),$$

- $\sigma_k^2$  is a variance of environmental effects  $\epsilon_k$ ;  $N_k$  is the sample size of population  $k$ ;  $\Psi_j$  represents the SNP-specific shrinkage parameter;  $\phi$  in  $\delta_j$  represents the global shrinkage parameter for across populations;  $a = 1$  and  $b = 1/2$  for Strawderman-Berger prior
- The posterior mean of  $\boldsymbol{\beta}_K$  for **population**  $k$ ,

$$E(\boldsymbol{\beta}_K | \hat{\boldsymbol{\beta}}_k) = (\mathbf{D}_k + \boldsymbol{\Psi}^{-1})^{-1} \hat{\boldsymbol{\beta}}_k$$

- $\hat{\boldsymbol{\beta}}_k = \mathbf{X}_k^T \mathbf{y}_k / N_k$ ;  $\mathbf{D}_k = \mathbf{X}_k^T \mathbf{X}_k / N_k$ ;  $\boldsymbol{\Psi} = \text{diag}\{\Psi_1, \dots, \Psi_M\}$
- The final PRS is  $PRS = w_{\hat{\phi},1} PRS_{\hat{\phi},1} + w_{\hat{\phi},2} PRS_{\hat{\phi},2} + \dots + w_{\hat{\phi},K} PRS_{\hat{\phi},K}$ .

# Trans-ethnic PRS

- **XPASS**: Cross-population analysis with summary statistics (Cai et al. AJHG 2021)

- For SNP effect size  $(\beta_{j1}, \beta_{j2})$  for SNP  $j$  in two populations,

$$\begin{pmatrix} \beta_{j1} \\ \beta_{j2} \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

- $\sigma_1^2$  and  $\sigma_2^2$ : variance components in two populations of 1 and 2
- $\rho$ : trans-ethnic genetic correlation of the same trait for populations

# Trans-ethnic PRS

- **SDPRX**: Summary-based Dirichlet process regression (Zhou et al. AJHG 2023)
  - SDPRX, an extension of SDPR with priors based on the genetic architectures of a trait in two populations by modelling a mixture distribution of four mutually exclusive scenarios
  - For SNP effect size  $(\beta_{j1}, \beta_{j2})$  for SNP  $j$  in two population,

$$\begin{pmatrix} \beta_{j1} \\ \beta_{j2} \end{pmatrix} \sim p_0 \begin{pmatrix} \delta_0 \\ \delta_0 \end{pmatrix} + p_1 \sum_{k=1}^{1000} \pi_{1k} \begin{pmatrix} N(0, \sigma_{1k}^2) \\ \delta_0 \end{pmatrix} + p_2 \sum_{k=1}^{1000} \pi_{2k} \begin{pmatrix} \delta_0 \\ N(0, \sigma_{2k}^2) \end{pmatrix} + p_3 \sum_{k=1}^{1000} \pi_{3k} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_{3k}^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

- $(p_0, p_1, p_2, p_3) \sim \text{Dir}(1)$  for which  $p_0, p_1, p_2$ , and  $p_3$  represents the cases of  $j$ th SNP is (1) null on both population, (2) population 1 specific, (3) population 2 specific, and (4) shared its effect between populations, considering the trans-ethnic genetic correlation  $\rho$ .

# Trans-ethnic PRS

- **TL-PRS:** Transfer Learning PRS for Fine-Tuning (Zhao et al. AJHG 2022)
  - TL-PRS, utilizes a gradient descent algorithm to fine tune a baseline PRS derived from GWAS summary statistics of a non-target ancestry group to a target ancestry group.
  - The model for the target ancestry group is as follows:

$$Y = \sum_{j=1}^M X_j \beta_j + C\gamma + \epsilon$$

- $Y$ : the target outcome of phenotype,  $X_j$ : genotypes for SNP  $j$ ;  $C$ : covariate matrix;  $\gamma$ : covariate coefficient vector;  $\epsilon$ : environmental error vector.
- Specifically, the true SNP effect size  $\beta_j$  is decomposed into two parts,  $\beta_j^S$  and  $\tau_j$ .
  - $\beta_j^S$  represents the estimated effect size from non-target population with large sample size
  - $\tau_j$  represents the difference between the  $\beta_j$  and  $\beta_j^S$ .
  - The starting value for this method is  $\beta_j^S$ .

# Trans-ethnic PRS

- **CTPR: Cross-eThnic Penalized Regression**

- It extends CTPR to multi-ethnic GWAS data by modeling population-specific LD structure

- $p_{\lambda_2}^{ctp}(\boldsymbol{\beta}) = \frac{\lambda_2}{2} \left\{ \sum_{k \neq k'}^K \sum_{j=1}^P LDD_{kk'j} (\beta_{kj} - \beta_{k'j})^2 + \sum_{k=1}^K \sum_{m=1}^M \sum_{j=1}^P LDD_{k(K+m)j} (\beta_{kj} - \hat{s}_{mj})^2 \right\}$

- SNP specific coefficient  $LDD_{k_1 k_2 j}$

- incorporates LD difference between population  $k_1$  and  $k_2$
- decreases when the LD structure around  $j$ th SNP is similar
- Scaled LDD (sLDD) for range of (0, 1)

$$sLDD_{k_1 k_2 j} = \frac{|LDSC_{k_1} - LDSC_{k_2}| - \min(|LDSC_{k_1} - LDSC_{k_2}|)}{\max(|LDSC_{k_1} - LDSC_{k_2}|) - \min(|LDSC_{k_1} - LDSC_{k_2}|)} \Rightarrow 0 \leq sLDD_{k_1 k_2 j} \leq 1$$

- Scaled  $sLDD^k$  for range of  $(1/\alpha, \alpha)$  ( $k$ : shape parameter,  $\alpha$ : scale parameter,  $\alpha > 1$ )

$$\alpha^{-2sLDD_{k_1 k_2 j}^k + 1} \Rightarrow \alpha^{-1} \leq \alpha^{-2sLDD_{k_1 k_2 j}^k + 1} \leq \alpha$$

# Methods Section

# Study Population

- Korean Cohorts

- The study utilizes the Korean Genome Epidemiology Study (**KoGES**, N = 211,725) and the Cohort for Reality and Evolution of Adult Asthma (**COREA**, N = 1,673).
- **KoGES** is composed of three sub-cohorts and genotyped using the specialized Koreanchip.
- **COREA** consists of asthma patients recruited from eleven university hospitals, stratified by Koreanchip ver 1.0 and 1.1.

- Sample QC and SNP QC

- Samples were excluded based on high missingness ( $> 5\%$ ), sex inconsistencies, or significant deviation from HWE ( $P < 10^{-5}$ ), resulting in **73,359** samples.
- Imputation filtering was applied to remove low-quality SNPs ( $\text{Imp } R^2 < 0.3$ ), rare variants ( $\text{MAF} < 0.5\%$ ), or those with high missing rates ( $> 5\%$ ), resulting in **4,734,789** SNPs.

# Target Phenotypes and Reference Panel

- Target Phenotypes for PRS
  - The study focuses on eleven complex traits, including eight continuous traits (e.g., BMI, SBP, DBP, GLU, TG, HDL, SMOKE, DRINK) and three binary traits (T2D, CAD, Asthma).
  - EAS summary statistics were sourced from BBJ, while EUR data were primarily obtained from the UKB and GWAS Catalog.
- Validation data and Reference Panel
  - The KoGES dataset (N=73,359) was split into two validation sets (N=3,000 each) and a large-scale test set (N=67,359).
  - The EAS and EUR 1000G ph3 ver5 served as LD reference panels.
  - The LD scores were computed based on genotype data from validation set 1 (N=3,000), 1000G EAS population (N=504), and 1000G EUR population (N=503).

# Genetic Correlation

- Cross-Trait Genetic Correlation (CTGC)
  - Genetic correlation ( $r_g$ ) quantifies the overlap between two traits within the same ancestry group (e.g., EAS-EAS or EUR-EUR).
  - For single-ethnic analysis, **LDSC** (Bulik-Sullivan, et al. Nat Genet 2015) is the standard for estimating genetic correlation when only GWAS summary statistics are available.
- Cross-Population Genetic Correlations (CPGC)
  - Cross-Population genetic correlation ( $r_{ge}$ ) measures the relationship between traits across different ancestry groups (e.g., EUR-EAS).
  - For trans-ethnic analysis, **POPCORN** (Cai et al. AJHG 2016) is utilized to estimate correlation coefficients by normalizing SNP effects with ancestry-specific allele variances.

# Selection Priority Strategy

- Selection Priority based on Genetic Correlation
  - Selection priority strategy focuses on traits with the **high genetic correlation** to the target outcome to achieve high predictive accuracy.
  - The framework involves calculating cross-trait correlations within each population using **LDSC** and cross-population correlations using **POPCORN**.
  - To identify the strongest genetic links to the target outcome, sub-matrices are created for all possible combinations of traits and populations.
  - By utilizing the most relevant traits and ethnic data, this systematic selection facilitates more accurate PRS modeling.

# PRS Estimation

- Single-Ethnic and Multi-Ancestry Modeling
  - Single-ethnic models (EAS/EUR) were built using CT, LDpred (infinitesimal/grid/auto), lassosum, and PRS-CS.
  - Meta-ethnic models utilized meta-analyzed summary statistics (EAS/EUR) generated from GWAMA and followed the same procedures for single PRS construction.
  - Multi-ethnic models combined the scaled single-ethnic PRS for EAS and EUR using weights estimated through linear/logistic regression on the validation 2 dataset.
  - Trans-ethnic models were built using PRS-CSx, XPASS, SDPRX, and TLPRS with regression coefficients estimated in the validation 2 dataset.

# PRS Evaluation

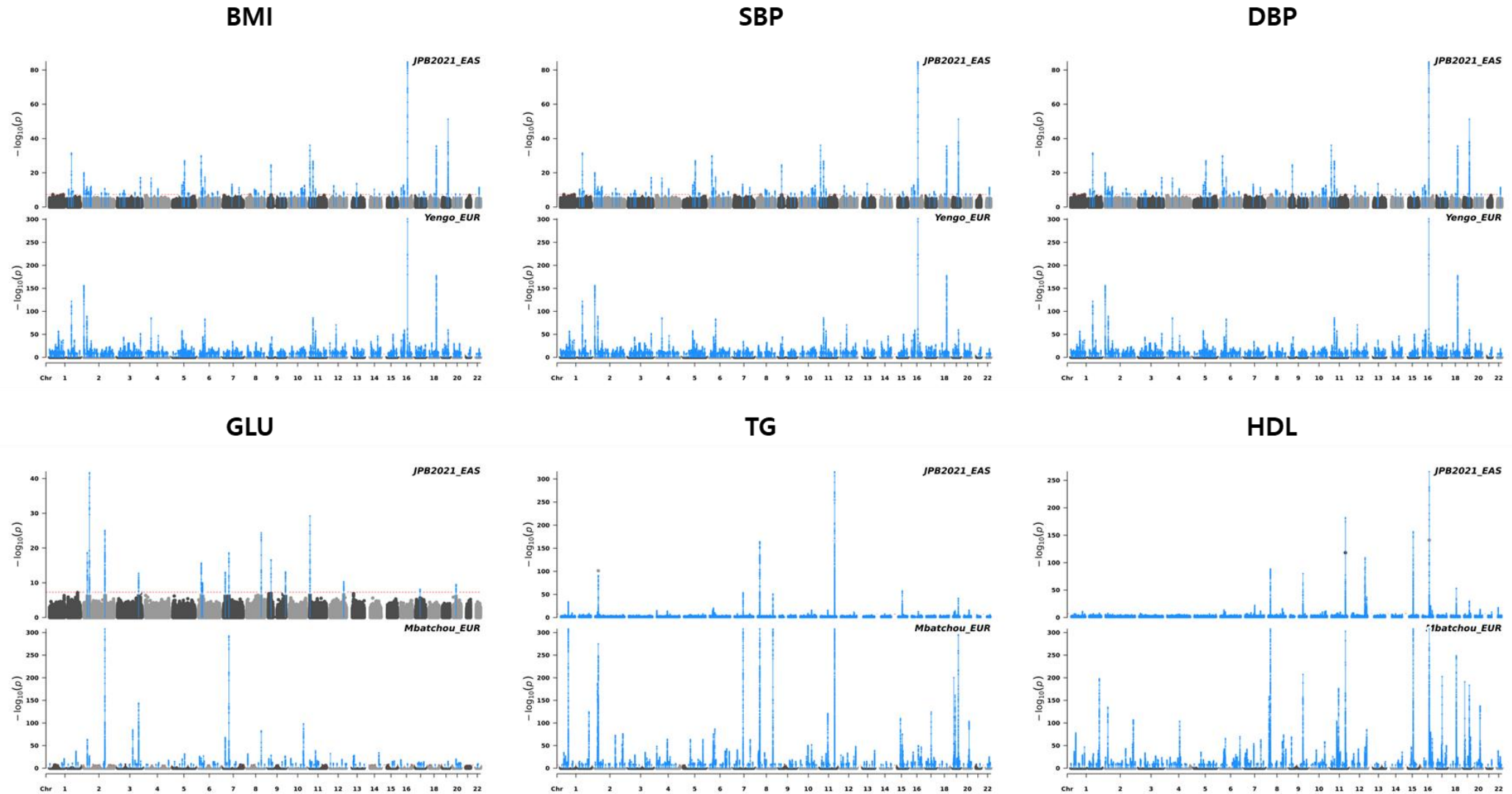
- Predictive Power Assessment
  - Model performance is evaluated using prediction  $R^2$  for continuous traits and AUC for binary traits with 10-fold CV.
  - Models are adjusted for covariates such as age, sex, and 5 genotype PCs.
  - The **DeLong test** is employed to statistically compare AUC improvements between the baseline model and the newly derived models.
  - We categorized the traits by genetic correlations into four groups: low, medium, high and suggested groups based on **DeLong P-values**.

# Results Section

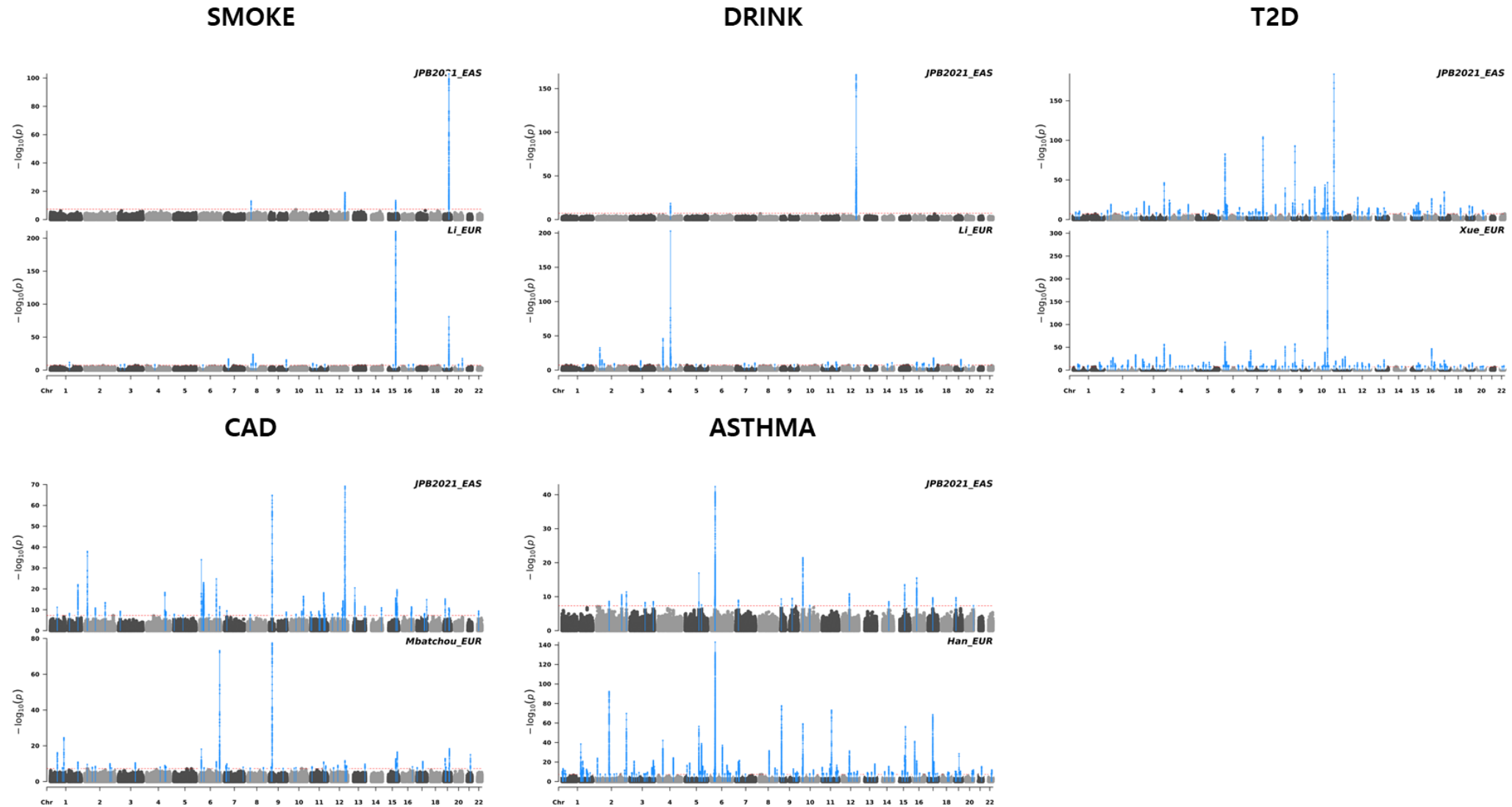
# Effect Size and P-value Comparisons (EAS vs EUR)

- Comparison of Summary Statistics for EAS vs EUR
  - EUR populations from UKB possess significantly larger sample sizes, ranging from 3 to 15 times those of EAS populations from BBJ across the 11 traits.
  - The larger EUR cohorts result in a higher number of detected SNPs and lower P-values while significant signals share similar genomic positions.
  - SNP effect sizes differ between the two populations, suggesting that raw effect estimates are not directly interchangeable.

# Manhattan Plots for EAS (Upper) and EUR (Lower)



# Manhattan Plots for EAS (Upper) and EUR (Lower)



# Cross-Trait Cross-Population Genetic Correlations

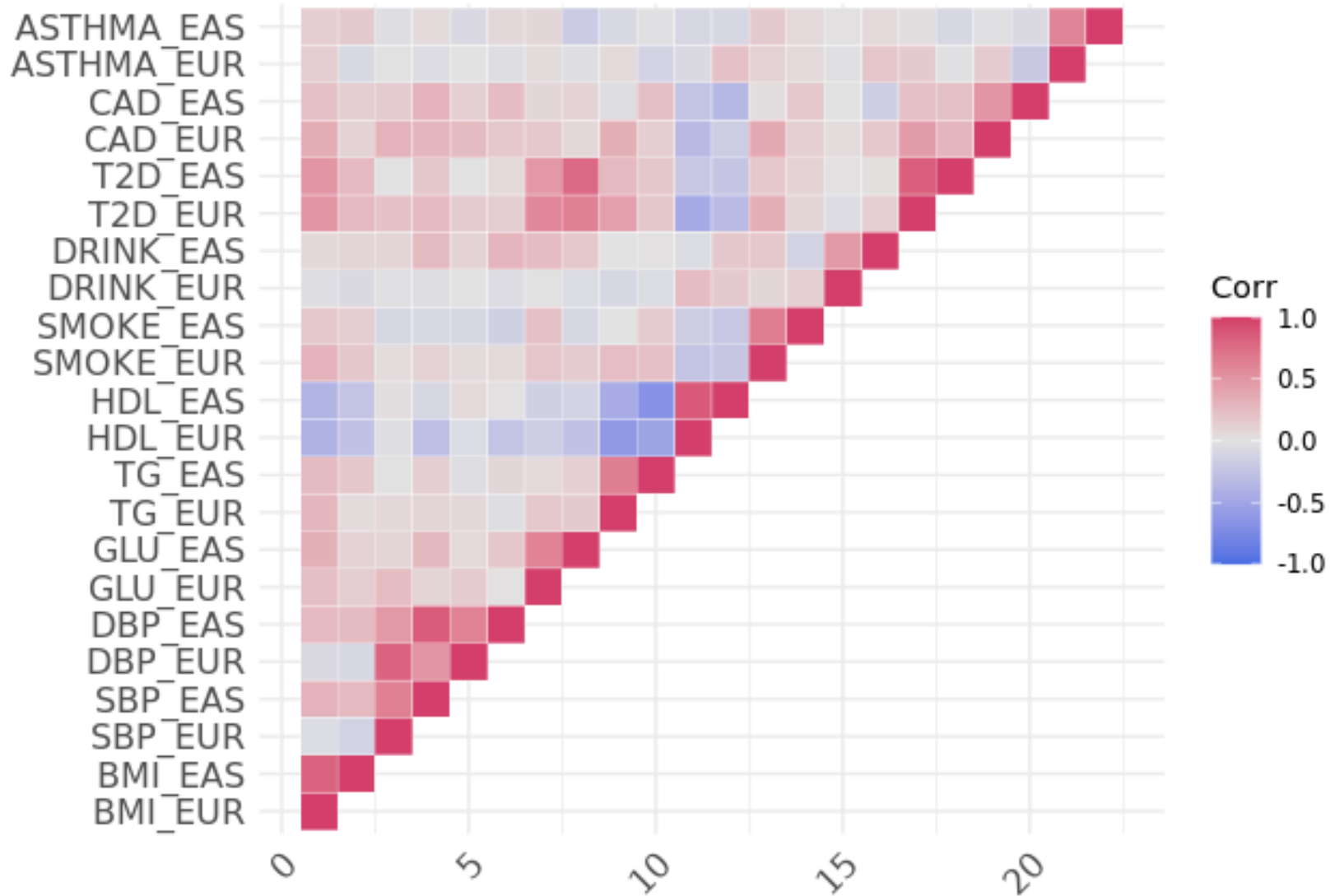
- Framework for Genetic Correlation

- Analysis of 11 traits in EAS and EUR populations was conducted using a  $2 \times 2$  sub-matrix to estimate within-ancestry and cross-population genetic correlations.

$$\begin{pmatrix} CTGC_{EUR-EUR}^{A-B} & CPGC_{EUR-EAS}^{A-B} \\ CPGC_{EAS-EUR}^{A-B} & CTGC_{EAS-EAS}^{A-B} \end{pmatrix}$$

- The matrix includes genetic correlations (diagonal elements) and cross-population genetic correlations (off-diagonal elements) for pairs of traits (A, B) and populations (EAS, EUR).
- The cross-trait correlations are generally higher.
- The cross-trait cross-population correlations often show decreased magnitude or even opposite directions.

# Cross-Trait Cross-Population Genetic Correlations



GLU

|     | T2D  |      |
|-----|------|------|
|     | EUR  | EAS  |
| EUR | 0.59 | 0.49 |
| EAS | 0.64 | 0.78 |

BMI

|     | T2D  |      |
|-----|------|------|
|     | EUR  | EAS  |
| EUR | 0.51 | 0.51 |
| EAS | 0.27 | 0.27 |

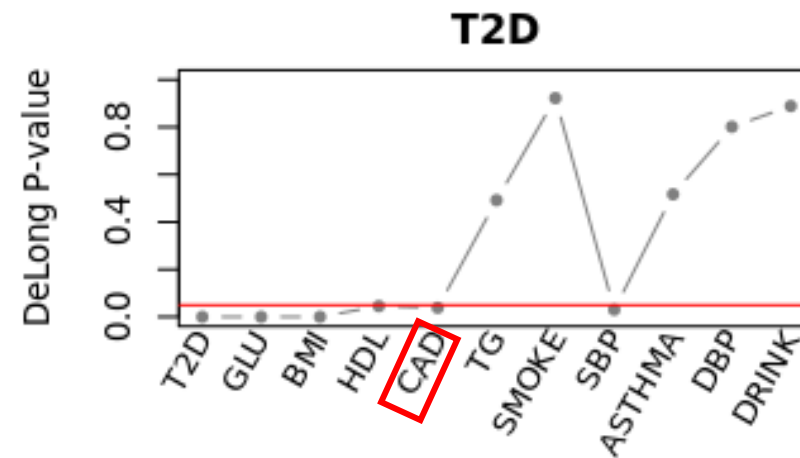
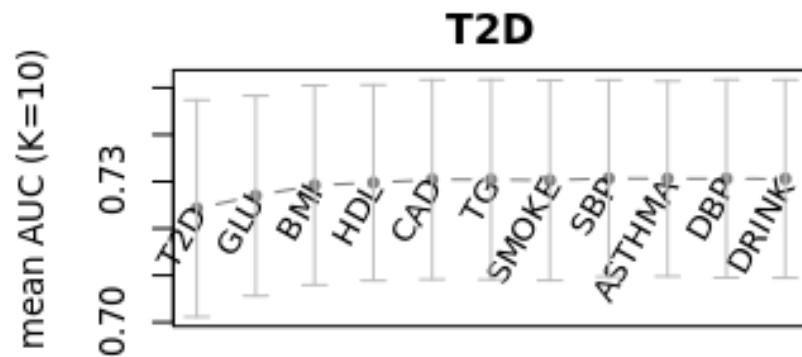
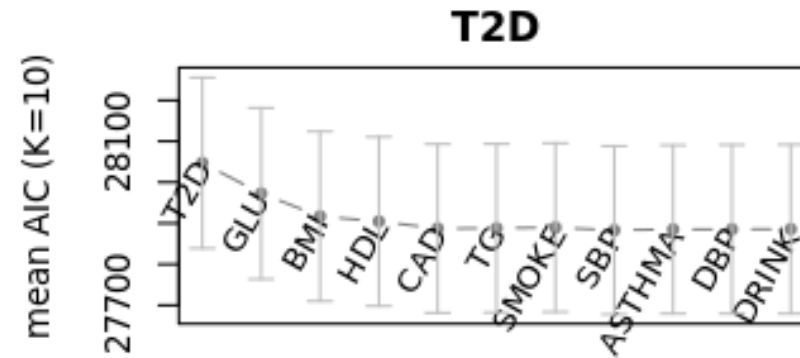
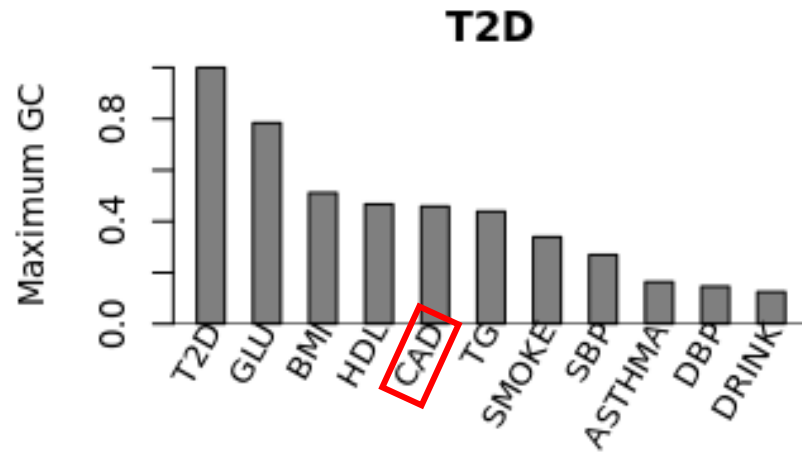
# Performance of Multi-trait and Trans-ethnic PRS

- Superiority of Multi-ancestry Approaches
  - Across all eleven traits, multi-ethnic or trans-ethnic approaches consistently outperformed single-ethnic models by leveraging both EAS and EUR populations.
  - In contrast, the meta-ethnic approach showed poor predictive performance, including negative correlations with continuous outcomes and lower AUC values.
- Top-Performing Models by Trait
  - Multi-ethnic methods (e.g., lassosum, PRS-CS) were most effective for TG, SMOKE, DRINK, and T2D, while trans-ethnic models (e.g., PRS-CSx, SDPRX) performed better for BMI, SBP, DBP, GLU, HDL, CAD, and asthma.
  - These results highlight the necessity of population-integrated modeling to capture a more comprehensive genetic architecture for both continuous and binary traits.

# Performance of Multi-trait and Trans-ethnic PRS

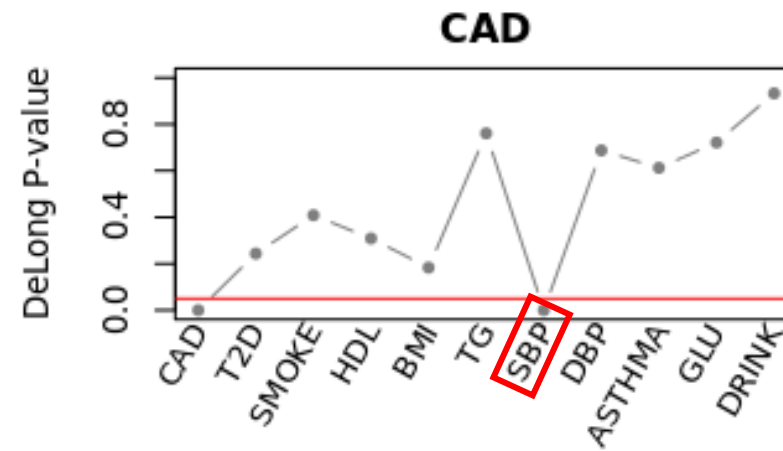
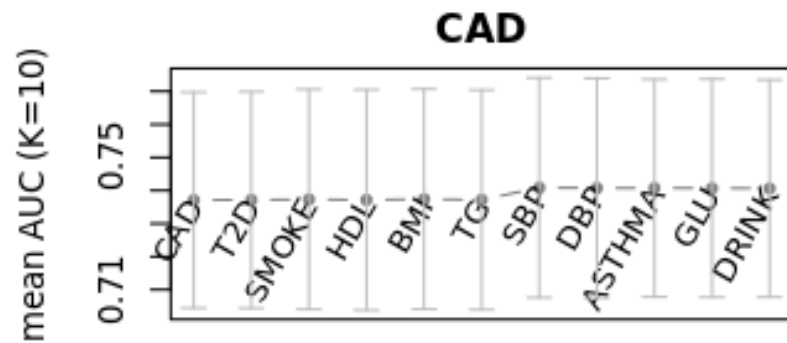
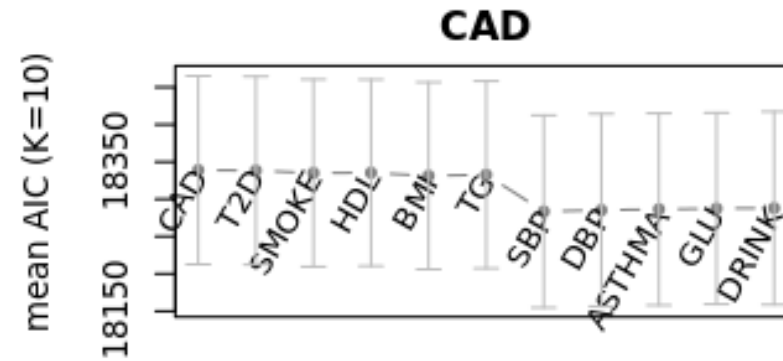
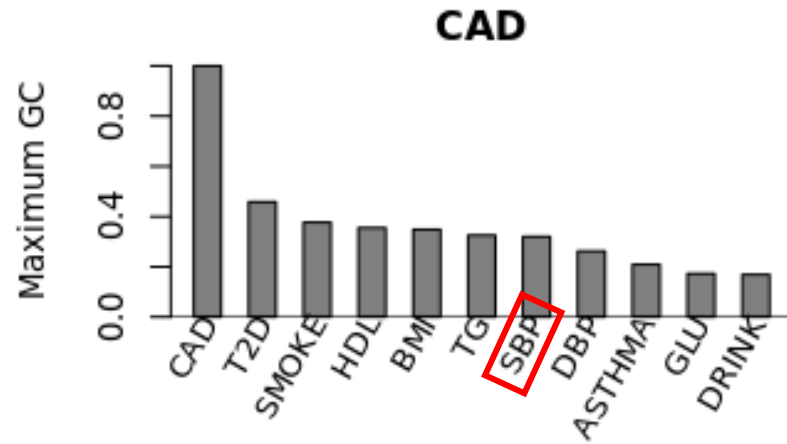
- Genetic Correlation-Based Selection Criteria
  - Trait selection was prioritized by genetic correlation; for example, **T2D** showed the highest correlation with **GLU** (Maximum GC=0.784), followed by **BMI** (0.511), **HDL** (0.402), **CAD** (0.348).
  - Comparing with single-trait PRS models, multi-trait PRS models showed improved prediction performance.
    - T2D prediction model, integrating PRSs for T2D, GLU, BMI, HDL, and CAD, exhibited enhanced performance (mean AIC = 27886.27, **mean AUC = 0.730**).
    - CAD prediction model, integrating PRSs for CAD, T2D, SMOKE, HDL, BMI, TG, and SBP, displayed improved performance (mean AIC = 18284.82, **mean AUC = 0.741**).
    - Asthma prediction model, integrating PRSs for asthma, HDL, CAD, GLU, DRINK, T2D, SMOKE, and BMI, also exhibited enhanced performance (mean AIC = 19553.22, **mean AUC = 0.607**).

# Contribution of Multi-trait and Trans-ethnic PRS



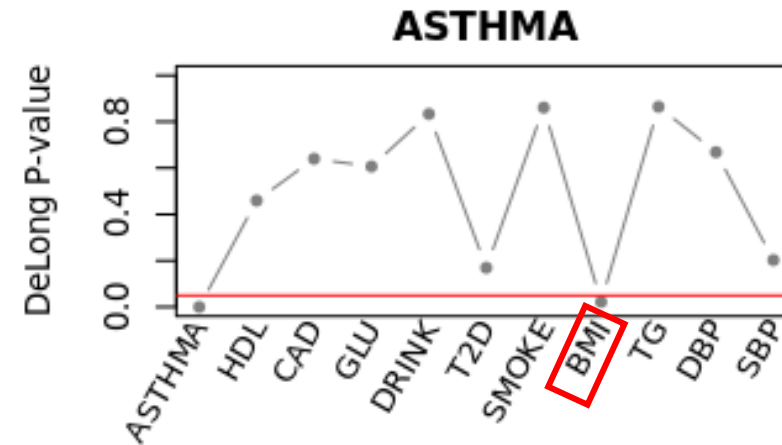
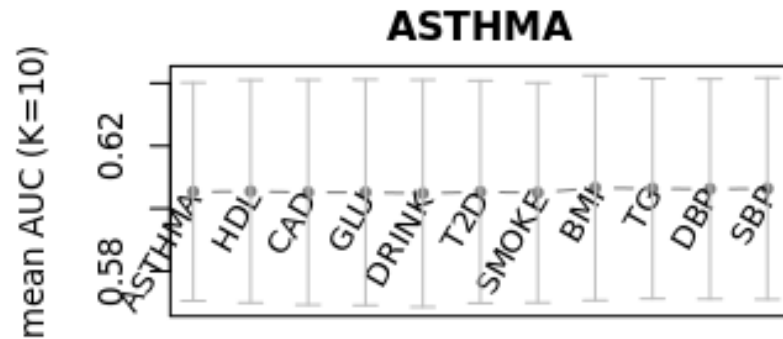
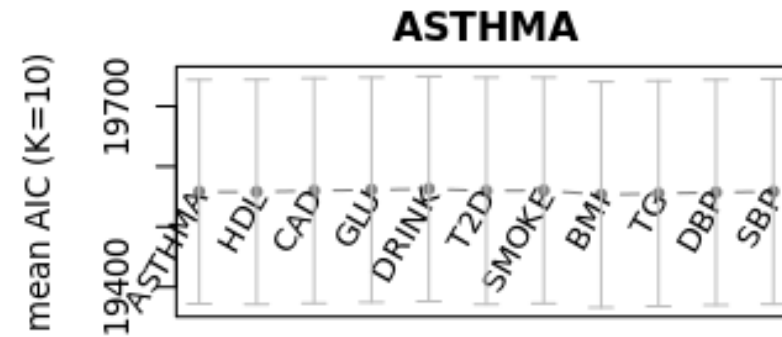
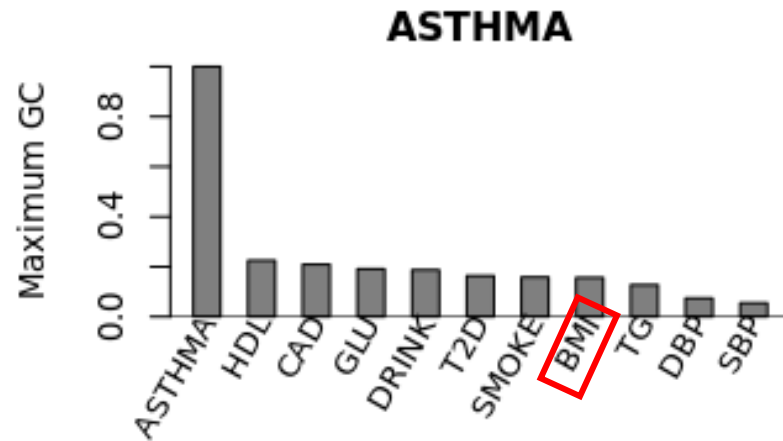
Red line represents the DeLong P-value=0.05

# Contribution of Multi-trait and Trans-ethnic PRS



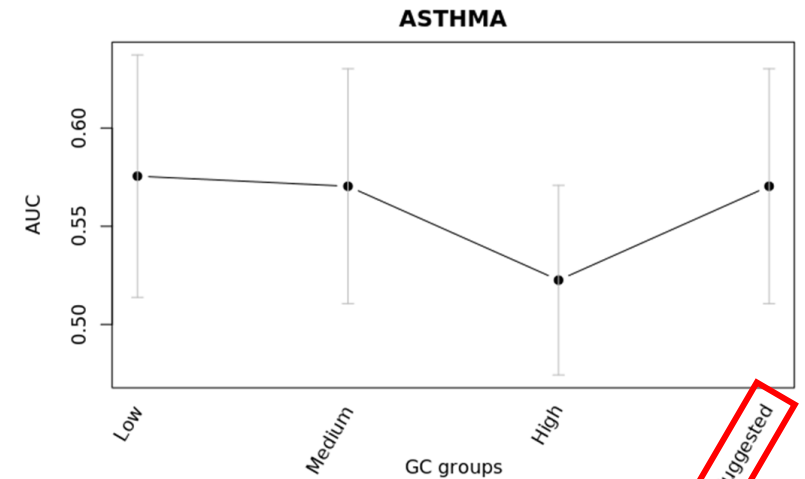
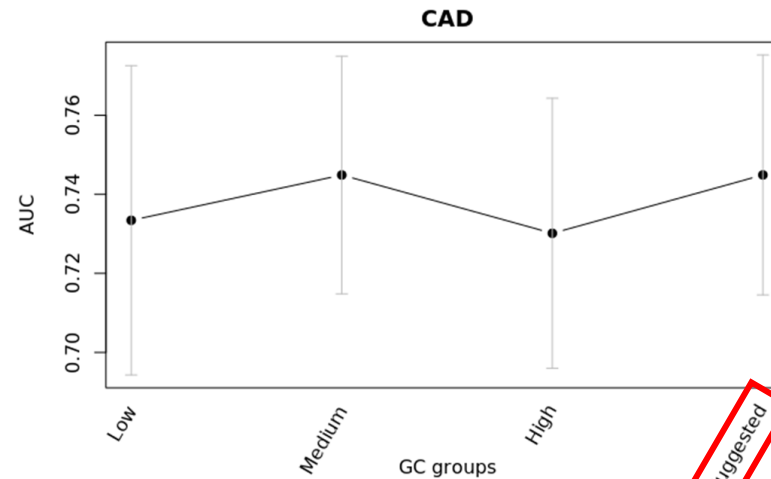
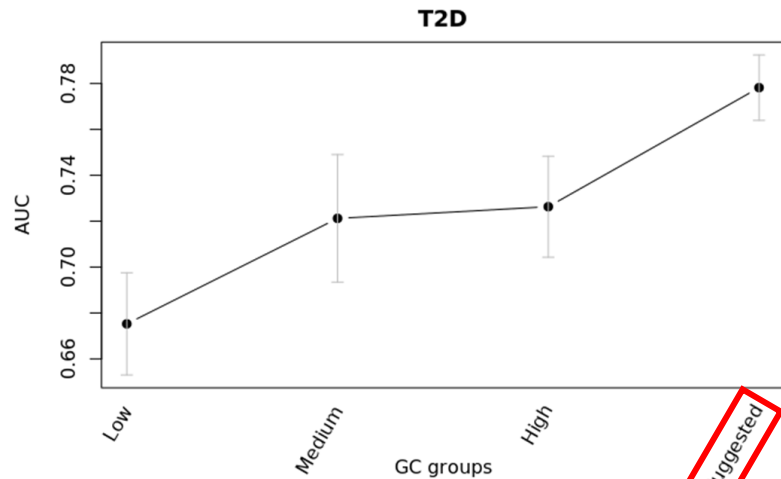
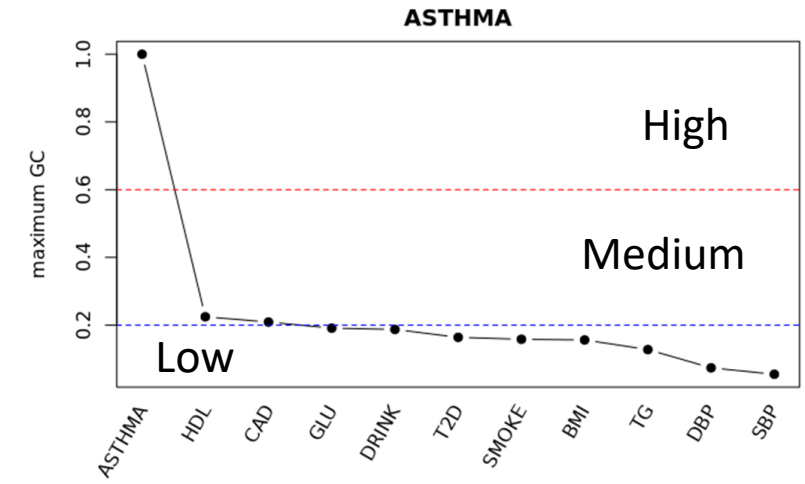
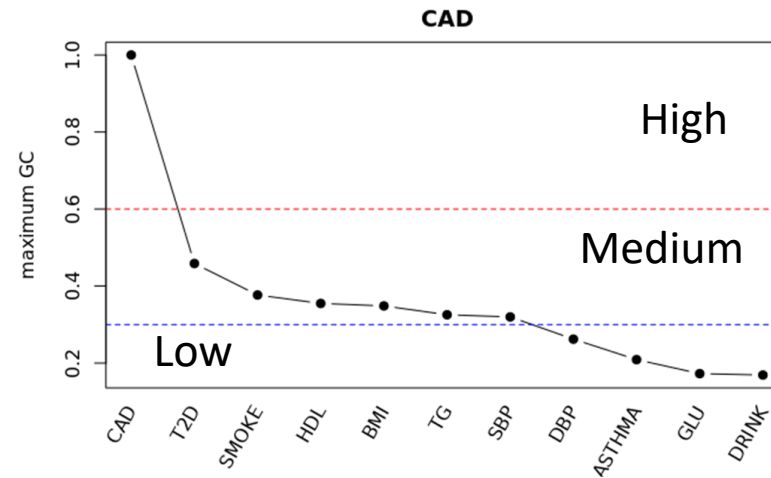
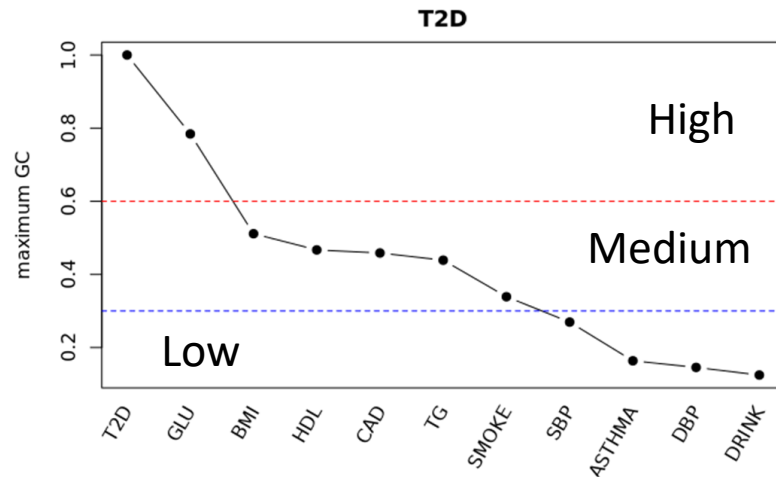
Red line represents the DeLong P-value=0.05

# Contribution of Multi-trait and Trans-ethnic PRS



Red line represents the DeLong P-value=0.05

# PRS Performance by Four Genetic Correlation Grouping



# Simulation Study

# Simulation Study

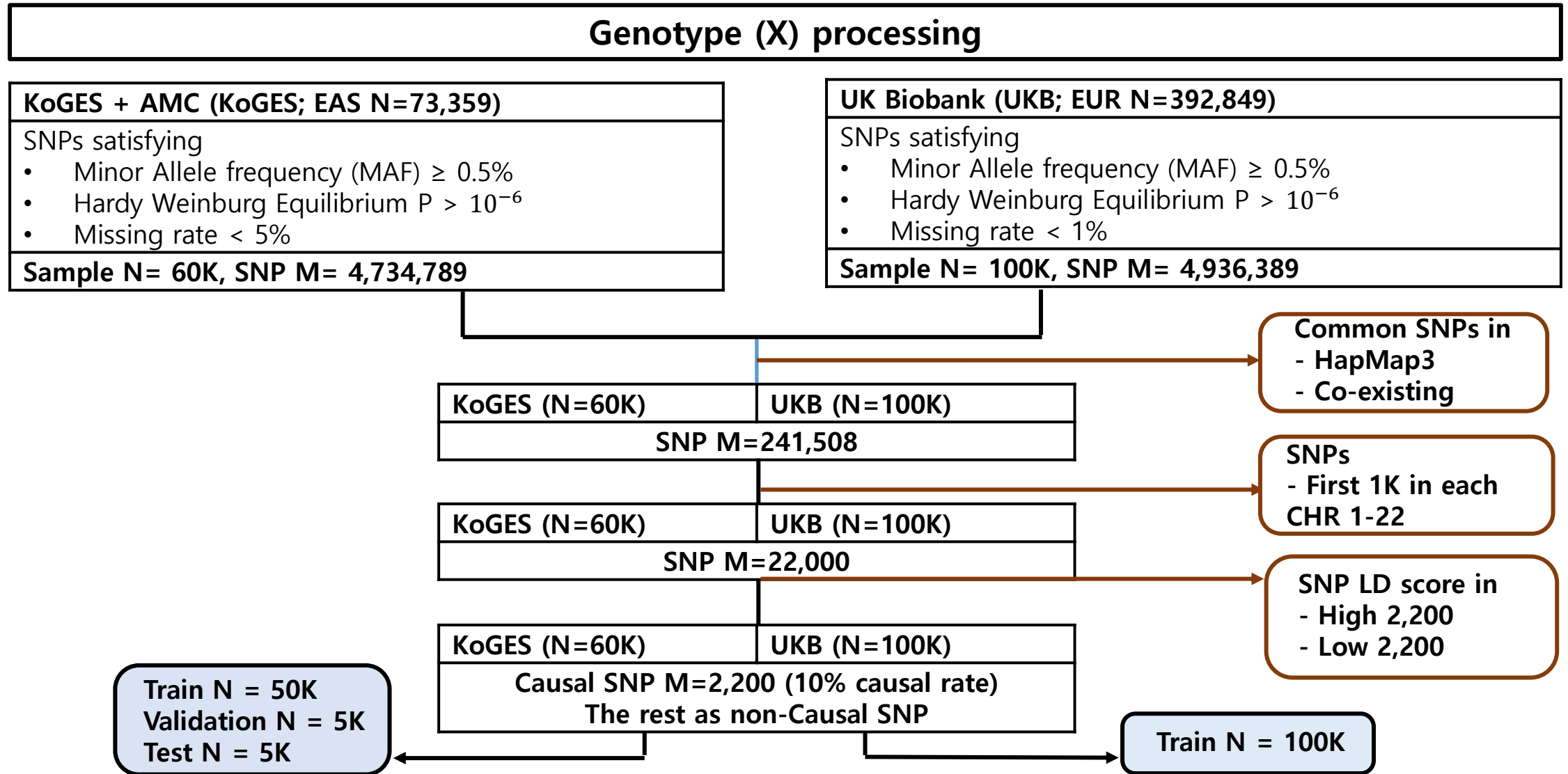
- Objectives of Simulation Study

- The study aims to evaluate how four key factors—(1) target sample size, (2) cross-trait genetic correlation (CTGC), (3) cross-population genetic correlation (CPGC), and (4) LD structure (cauLDGr)—impact PRS performance.

- Genotype Data and SNP Selection

- KoGES (Target/EAS) and UKB (Non-target/EUR) were processed to extract 1,398,733 common HM3 SNPs.
- To maintain the LD structure, the first 1K SNPs per chromosome were extracted and categorized into deciles based on the absolute difference in LD scores.
- The target KoGES population (N=60K) was split into training (50K), validation (5K), and test (5K) sets, while 100K UKB samples served as non-target training data.

# Simulation Study



# Simulation Study

- Phonotype Generation

- Phenotypes were generated for four traits with fixed heritability ( $h^2 = 0.3$ ) and a causal SNP proportion of 10% (2,200 SNPs).
- Genetic correlations were varied to simulate low/high scenarios: CTGC (0.4 to 0.8), CPGC (0.4 to 0.8), and cauLDGr (causal SNP LD difference group) (Low vs. High).
- For target trait (Trait 1), we converted it to binary trait with prevalence of 20%.

- PRS Computation

- PRS models were constructed using CT, PRS-CS, and PRS-CSx to compare single-ethnic, multi-ethnic, trans-ethnic, and the proposed multi-trait and trans-ethnic approaches.

# Simulation Settings

- Four Main Questions

1. Is the multi-ethnic or trans-ethnic PRS performance improved as **target sample size** increases in large number of non-target sample size?
2. Is the multi-ethnic or trans-ethnic PRS performance improved in high **CPGC** even though the causal SNPs have different LD structure?
3. Which methods are better, **multi-ethnic** or **trans-ethnic** PRS method?
4. How PRS for multi-trait and multi/trans ethnic PRS performs in different combination of **CTGC**, CPGC, and causal SNP LD score?

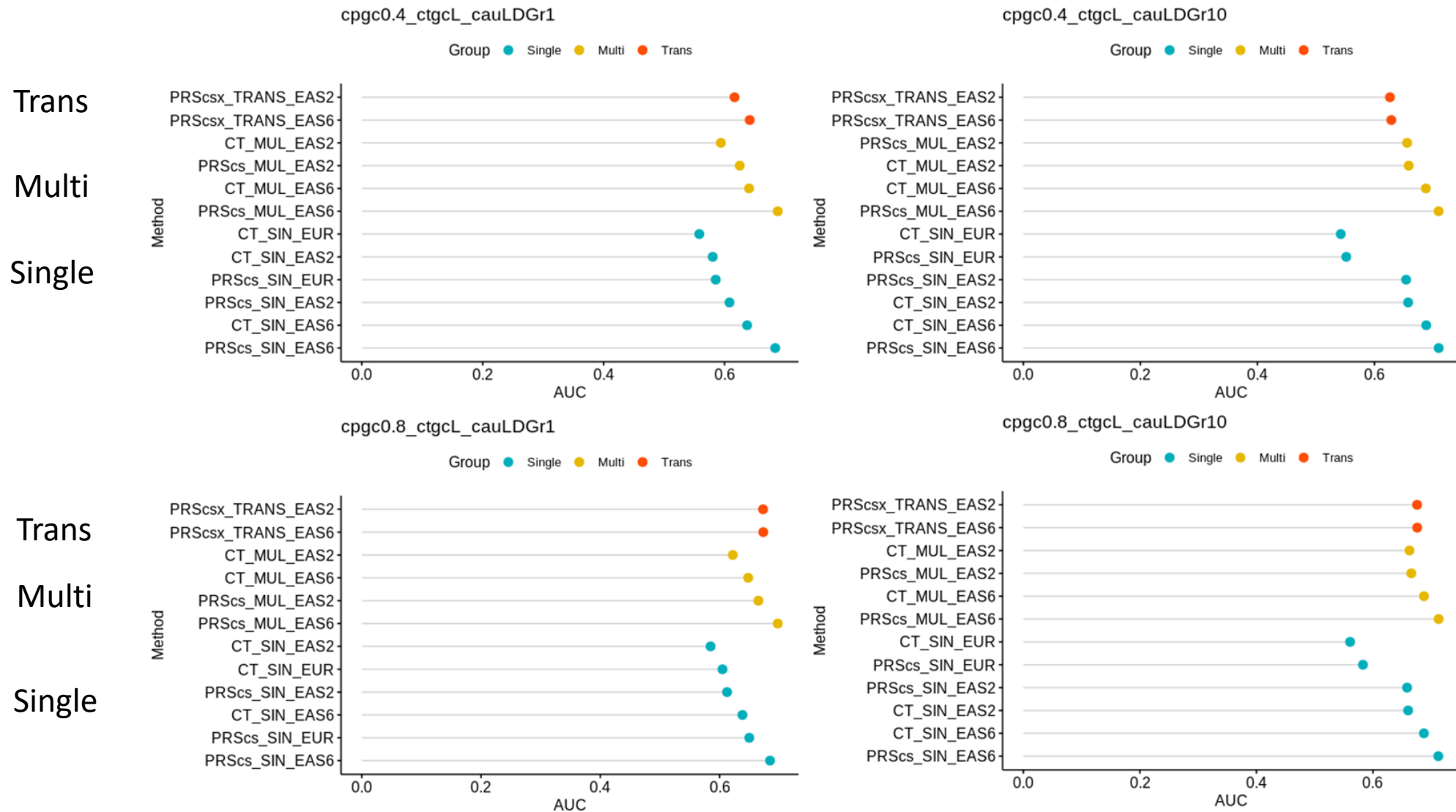
# Simulation Settings

| Question                                  | EAS Train N | CTGC            | CPGC            | cauLDGr | Trait            | Approach                                                                                |
|-------------------------------------------|-------------|-----------------|-----------------|---------|------------------|-----------------------------------------------------------------------------------------|
| 1. Target sample size                     | 10K, 50K    | 0.4             | 0.4, 0.8        | 1, 10   | Trait 1          | Single-ethnic PRS,<br>Multi/Trans-ethnic PRS                                            |
| 2. CPGC & cauLDGr                         | 50K         | 0.4             | <b>0.4, 0.8</b> | 1, 10   | Trait 1          | Single-ethnic PRS,<br>Multi/Trans-ethnic PRS                                            |
| 3. Multi-ethnic vs. Trans-ethnic PRS      | 50K         | 0.4             | 0.4, 0.8        | 1, 10   | Trait 1          | <b>Multi/Trans-ethnic PRS</b>                                                           |
| 4. Multi-trait and Multi/Trans ethnic PRS | 50K         | <b>0.4, 0.8</b> | 0.4, 0.8        | 1, 10   | <b>Trait 1-4</b> | Single-ethnic PRS,<br>Multi/Trans-ethnic PRS,<br>Multi-trait and Multi/trans ethnic PRS |

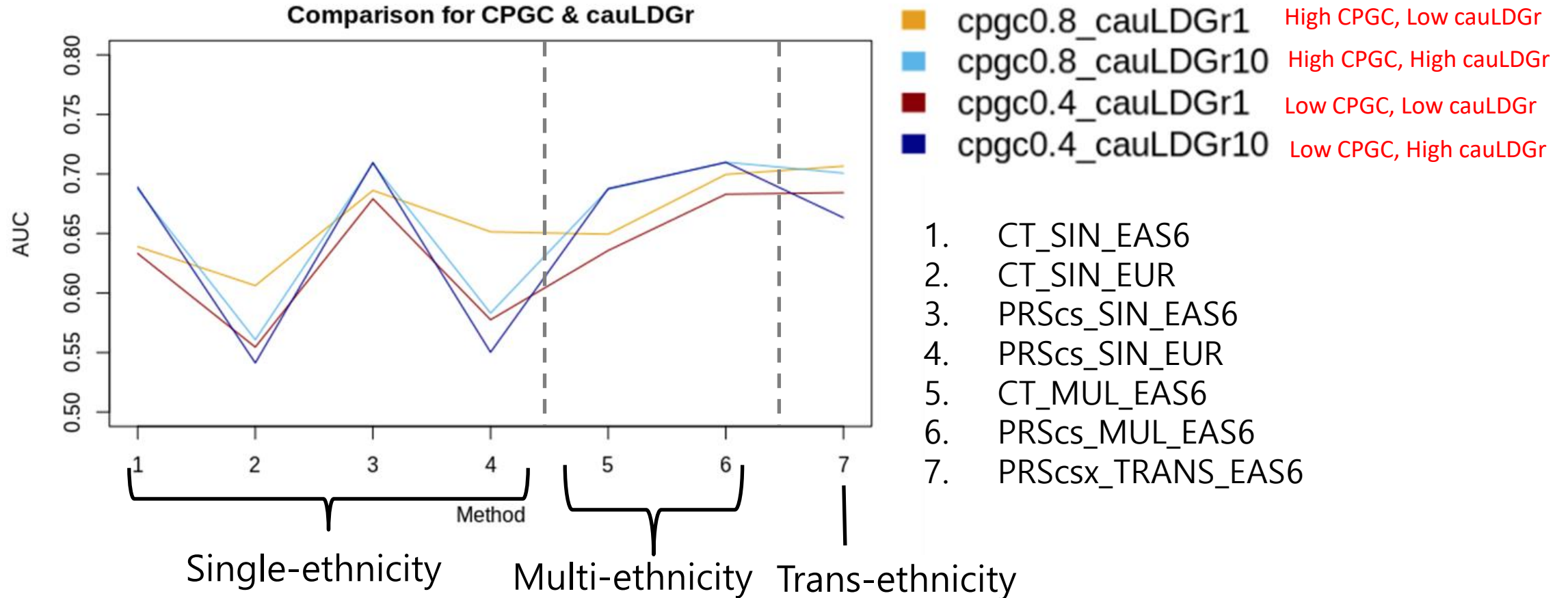
# Results of Simulation Study (Q1 & Q2)

- Target Sample Size (Q1)
  - PRS performance across all approaches (single, multi-ethnic, trans-ethnic) improves as the target population (EAS) sample size increases from 10K to 50K.
  - Performance of multi-ethnic PRS are robustly increasing depending on the size of target sample size no matter when the CPGC and causal SNP difference.
- Effect of CPGC and LD Differences (Q2)
  - Trans-ethnic PRS performance is primarily driven by high CPGC and low LD differences (cauLDGr) between populations.
  - Despite methodological adjustments, trans-ethnic models struggle under high LD differences, indicating a heavy reliance on shared genetic architecture.

# Target Sample Size (Q1)



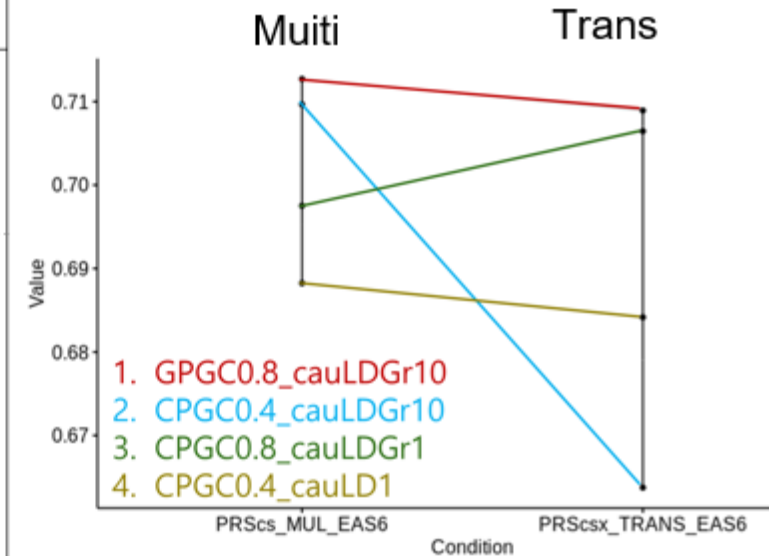
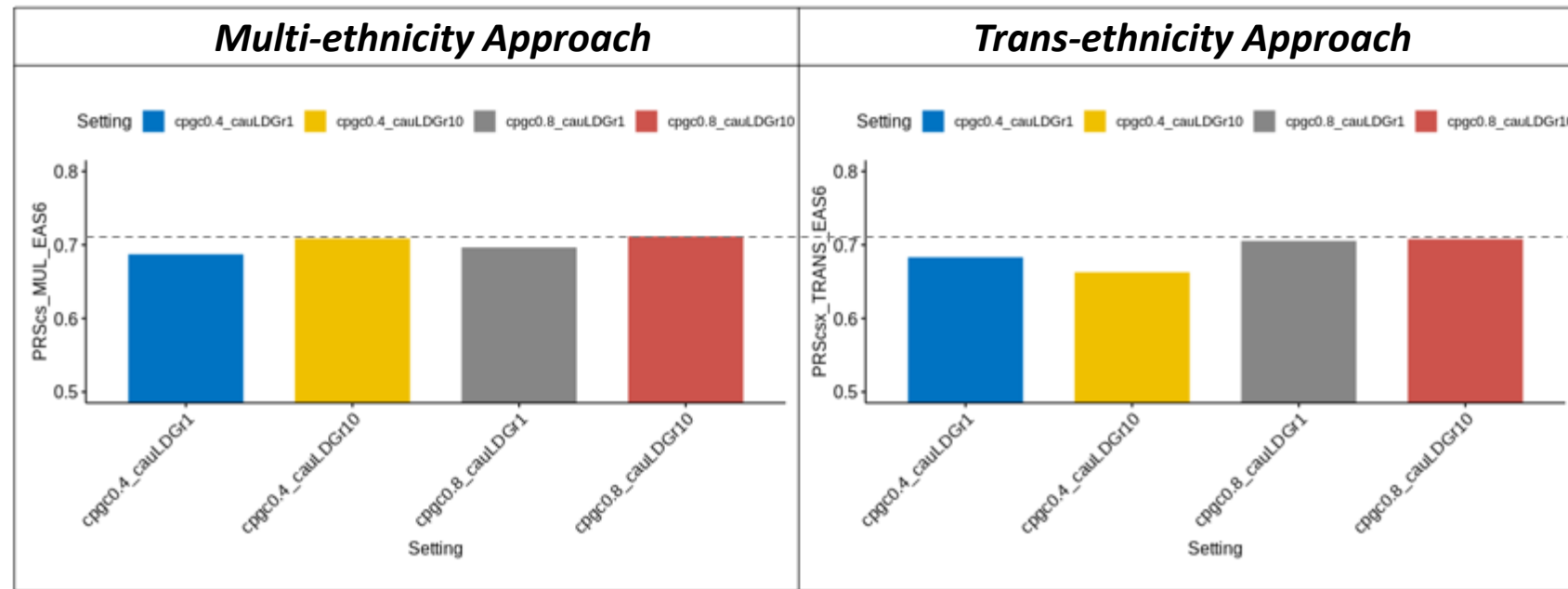
# Effect of CPGC and LD Differences (Q2)



# Results of Simulation Study (Q3 & Q4)

- Multi-ethnic vs. Trans-ethnic PRS (Q3)
  - Multi-ethnic PRS approach demonstrates superior robustness, providing consistently stable and reliable performance across most simulation scenarios.
  - Trans-ethnic PRS achieves peak performance under ideal conditions (high CPGC/low LD difference) but shows high sensitivity to setting changes.
- Optimal Strategies for Complex Scenarios (Q4)
  - Low CPGC/High LD Difference: A combination of multi-trait and multi-ethnic approaches is recommended, especially when non-target traits are highly correlated (CTGC).
  - High CPGC/Low LD Difference: Trans-ethnic PRS is the best option; adding multiple traits provides no significant gain.

# Multi-ethnic vs. Trans-ethnic PRS (Q3)



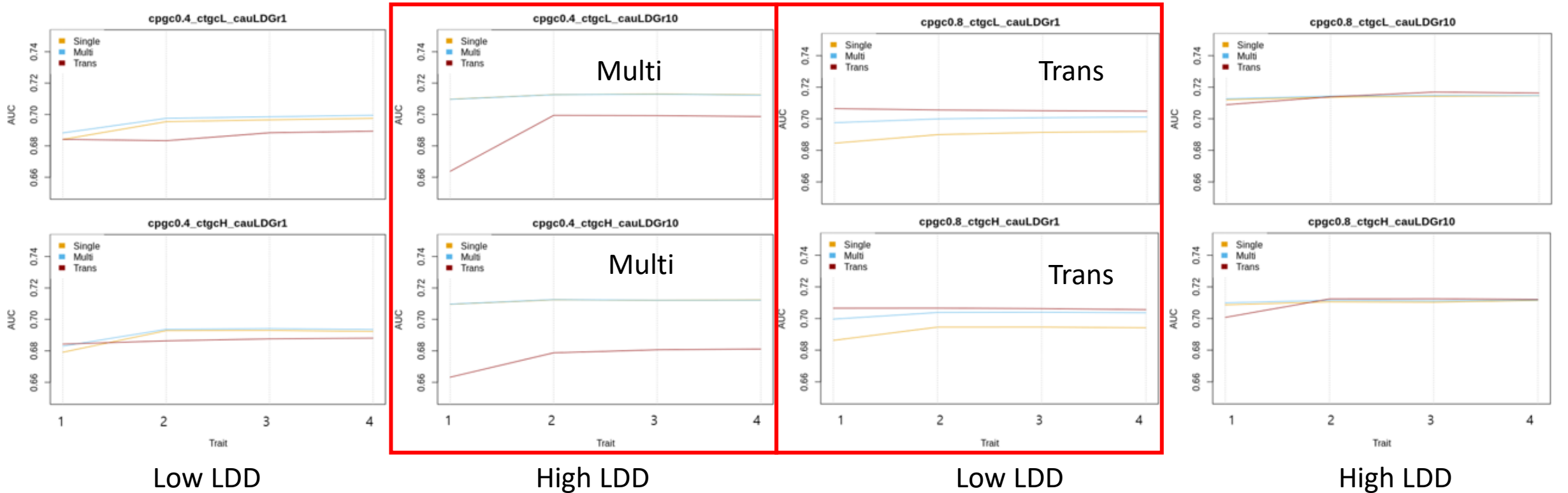
# Optimal Strategies for Complex Scenarios (Q4)

Low CPGC

High CPGC

CPGC = 0.4

CPGC = 0.8



# Discussion and Conclusion

# Key Findings on Trans-ethnic PRS

- Superiority of Integrated Ancestry Approaches
  - Leveraging diverse ethnic backgrounds (**multi-ethnic** or **trans-ethnic**) yielded substantial improvements in prediction accuracy compared to single-ethnic GWAS.
  - The **multi-ethnic** PRS approach shows consistently robust performance and **trans-ethnic PRS** achieves the best results under ideal conditions such as high CPGC/low LD difference.
- Limitations of Simple Meta-Analysis
  - **Meta-ethnic PRS** approaches that fail to account for LD showed limited strength in prediction power.
  - This underscores that incorporating LD structure is crucial for capturing the true genetic susceptibility across different populations.

# Trait Selection Strategy and Future Works

- Genetic Correlation based Selection Priority Strategy
  - Significant performance gains were observed when traits with cross-population genetic correlations of 0.3–0.4 were selected, based on mean AIC, mean AUC, and DeLong p-values.
- Future Works
  - Development of an advanced **Cross-ethnic CTPR** model that integrates pleiotropic effects across both multiple traits and multiple populations.
  - This extension will leverage Cross-ethnic CTPR with a selection priority strategy to jointly incorporate multiple traits and populations, aiming to improve PRS prediction accuracy.

# Part II: Survival MR Methods and Real Data Analysis

# Outline

- Background and Introduction
- Causal Effect Estimation via CoxPH and Additive Models
- Survival MR for Gene Expression on Prognosis in Colorectal Cancer

# Background and Introduction

# Survival Models in MR

## Statistical Hazard Models for Mendelian Randomization in Large-scale Survival Data

Jungeun Kim<sup>1</sup>, Seunghwan Park<sup>1</sup> and Wonil Chung<sup>1,2#</sup>



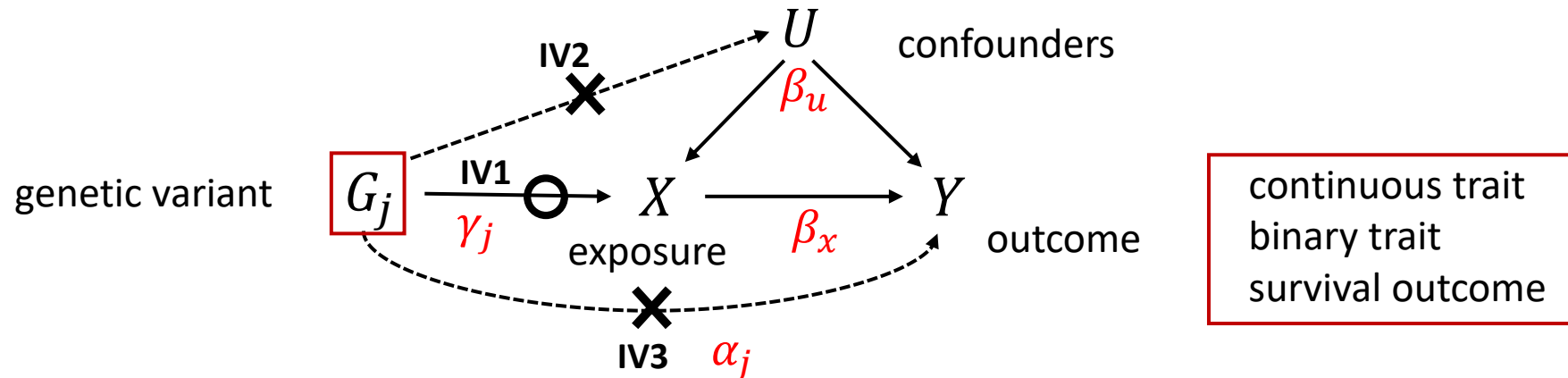
### Affiliations

<sup>1</sup> Department of Statistics and Actuarial Science, Soongsil University, Seoul, 06978, Korea

<sup>2</sup> Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

# What is Mendelian Randomization?

- Mendelian Randomization (MR)
  - MR examines the causal effect of an **exposure** on an **outcome** in observational studies
- Assumptions for a valid instrumental variable (IV) in MR
  - The genetic variants meet the following assumptions to be a valid IV (Angrist, Imbens et al. 1996)



IV1) The genetic variant is directly associated with the exposure ( $\gamma_j$ )

IV2) The genetic variant is not related to factors known to obscure the connection between the exposure and the effect ( $\beta_u$ )

IV3) The genetic variant has no effect other than through the exposure ( $\alpha_j = 0$ )

# MR Methods

- Individual-based MR Methods

- TSLS (Two-Stage Least Squares)

- Uses a two-stage process to estimate causal effects while addressing endogeneity between exposure and outcome variables.

- Summary-based MR Methods

- IVW (Inverse Variance Weighted) (Johnson and Uk 2012)

- Combines causal effect estimates from genetic variants using inverse variance weighting.

- WME (Bowden, Davey Smith et al. 2015)

- Combines causal effect estimates from genetic variants using weighted median.

# Why MR Needs to be extended to Survival Data?

- Survival outcomes
  - Many real outcomes, such as cancer prognosis and mortality, are **time-to-event outcomes** rather than continuous or binary outcomes.
  - Extending MR to survival data requires researchers to choose between (1) **Cox proportional hazards (CoxPH) model** and (2) **additive hazards model**.

# Hazard Model Choice Matters

- Two Survival Models
  - **CoxPH model** estimates a hazard ratio  $\exp(\beta_x)$ , and it is **non-collapsible** because this is a ratio-scale measure.
  - **Additive model** estimates a hazard difference  $\beta_x$ , and it is **collapsible** because this is a linear, absolute-scale measure.
- Consequence
  - CoxPH and additive model can yield **different causal conclusions in MR**, purely because of **model structure** rather than biology.
  - This issue raises two open questions: (1) which model is more appropriate, and (2) how can we compare them fairly?

# Cox Proportional Hazards Model (CoxPH)

- CoxPH Model

- The hazard function  $h^m(t|X, U)$  represents a **multiplicative hazard**.
- This model is the most **widely used** survival model, making it a natural starting point.

$$\Rightarrow h^m(t|X, U) = h_0^m \cdot \exp(\beta_x^m X + \beta_u^m U)$$

$$S^m(t|X, U) = \exp(-h_0^m t \cdot \exp(\beta_x^m X + \beta_u^m U))$$

$$\because H(t) = \int h(t)dt, S(t) = \exp(-H(t))$$

- Key limitations

- The hazard ratio  $\exp(\beta_x^m)$  is **non-collapsible**, meaning it changes with the adjustment set even when there is no confounding.
- This property complicates fair comparison across models or studies.

# Additive Hazards Model (Aalen)

- Additive Model

- The hazard function  $h^a(t|X, U)$  represents an **additive hazard**, absolute-scale structure.
- The effect measure  $\beta_x^a$  is **collapsible**, meaning it stays stable under adjustment.

$$\begin{aligned} \Rightarrow h^a(t|X, U) &= h_0^a + \beta_x^a X + \beta_u^a U \\ S^a(t|X, U) &= \exp(-(h_0^a + \beta_x^a X + \beta_u^a U)t) \end{aligned}$$

- Practical challenge

- This model is **less commonly used** in applied research.
- The baseline hazard  $h_0^a$  is estimated nonparametrically, which complicates direct comparison to CoxPH.

# Research Objectives

- Methodological objectives
  - This study builds a framework that fairly compares **CoxPH** vs. **additive models** despite their incomparable effect scales.
  - This study also establishes **model diagnostics (e.g. Martingale residuals and Arjas residuals)**, along with a validating simulation, to decide which model fits better.
- Applied objectives
  - This study applies the framework to real **two-sample MR data** from TCGA and UK Biobank.
  - This study evaluates the causal effects of **gene expression** on **prognosis in colorectal cancer** under both hazard models.

# Causal Effect Estimation via CoxPH and Additive Models

# Restricted Mean Survival Time (RMST)

- RMST

- RMST is defined as area under the survival curve up to  $\tau$ , and this measure is model-free.
- Researchers set  $\tau$  as the minimum of the maximum observed event time under each model, giving a common time window. (i.e.  $\tau = \min(\max t_i^m, \max t_i^a)$ )


$$\begin{aligned}\mu^m(h_0^m, \beta_x^m, X) &= \int_0^\tau S^m(t|X) dt \\ \mu^a(h_0^a, \beta_x^a, X) &= \int_0^\tau S^a(t|X) dt\end{aligned}$$

- Why it works

- Researchers can compute RMST from either model's fitted survival curve.
- RMST provides a shared metric that is independent of the hazard formulation.

# Calibration via RMST

- Objective function

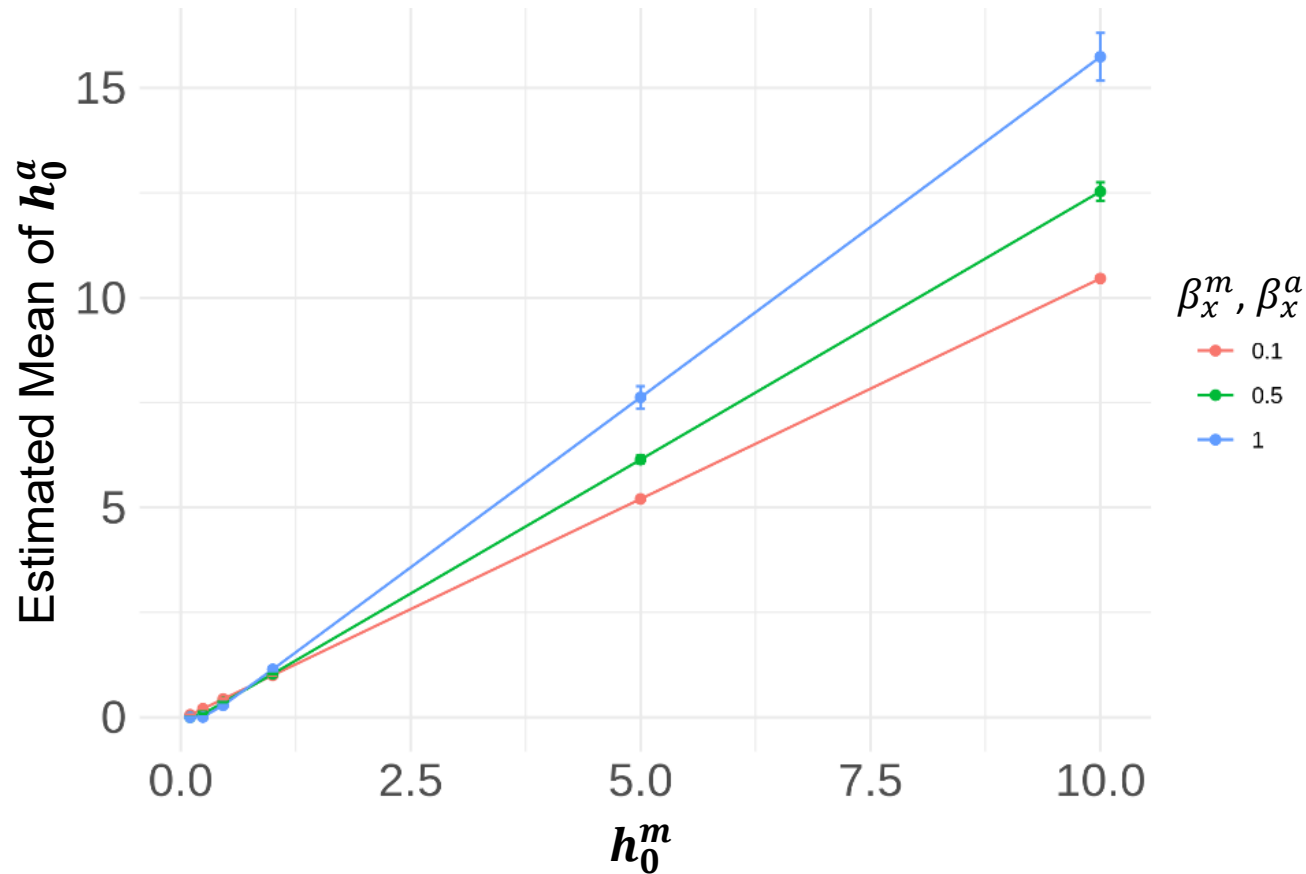
- The objective function adjusts the additive model's baseline hazard  $h_0^a$  to match CoxPH's RMST with a fixed  $h_0^m$ .
- Researchers solve this objective via a grid search over candidate  $h_0^a$  values.

➡ Estimated  $h_0^a = \underset{h_0^a}{\operatorname{argmin}} |\mu^m(h_0^m, \beta_x^m, X) - \mu^a(h_0^a, \beta_x^a, X)|$

- What this achieves

- This calibration produces survival data on a comparable scale across both models.
- This calibration also isolates estimator-performance differences from model-structure differences, which is the methodological core of the study.

# Estimated Parameters under Varying Baseline Hazards



Additive-model parameter  $h_0^a$  estimated using the objective function for CoxPH data generated with  $h_0^m$  parameter

As  $\beta_x^a$  increases, requiring a larger  $h_0^a$  to maintain comparable mean survival times.

Relationship Between the Estimated Mean of  $h_0^m$  and  $h_0^a$

# Simulation Study

- Genotype Data

- Genotype data were simulated for **1000** individuals and **100** SNPs, each with MAF of **0.5**.
- For each SNP  $j$ , the genotype  $G_{ij}$  for individual  $i$  was drawn from a discrete distribution taking values 0, 1 or 2, with corresponding probabilities (0.25, 0.5, 0.25), assuming HWE.
- This process was repeated across **100** independent simulation iterations.

# Simulation Settings

| Parameter                                         | Setting                         |
|---------------------------------------------------|---------------------------------|
| Hazard model                                      | CoxPH vs. Additive hazard       |
| Causal estimation method                          | TSLS, IVW, WME                  |
| Instrument strength ( $\gamma_j$ )                | 0.01, 0.03, 0.05, 0.1, 0.3, 0.5 |
| Causal effect size ( $\beta_x^m, \beta_x^a$ )     | 0, 0.05, 0.1, 0.3, 0.5          |
| Confounder effect size ( $\beta_u^m, \beta_u^a$ ) | 0, 0.05, 0.1, 0.5, 1, 3         |
| Baseline hazard function( $H_0$ )                 | Exponential                     |
| Pleiotropic effect ( $\alpha_j$ )                 | 0                               |
| Sample overlap                                    | 0%                              |

# Instrument Strength

- Instrumental Strength

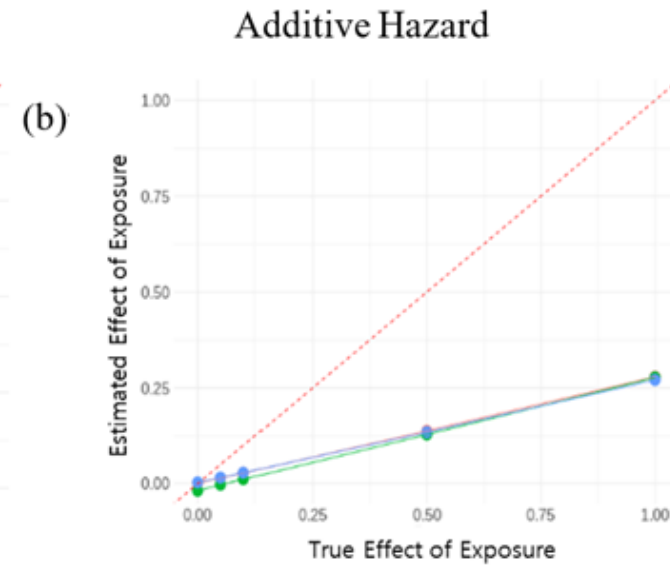
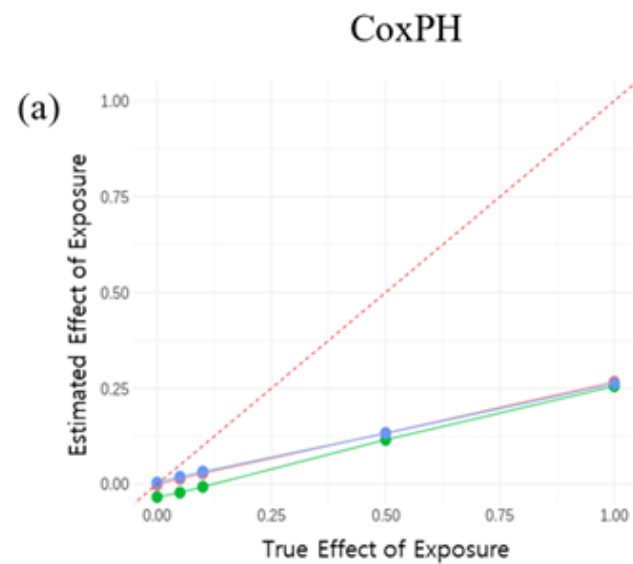
- Values of  $\gamma_j$  ranging from 0.01 to 0.5 were evaluated using a linear model with 100 independent instruments, where the exposure was generated as

$$X_i = \sum_{j=1}^{100} G_{i,j} \gamma_j + 0.5U_i + 0.5\epsilon_i, U_i \sim N(0,0.5), \epsilon_i \sim N(0,0.5)$$

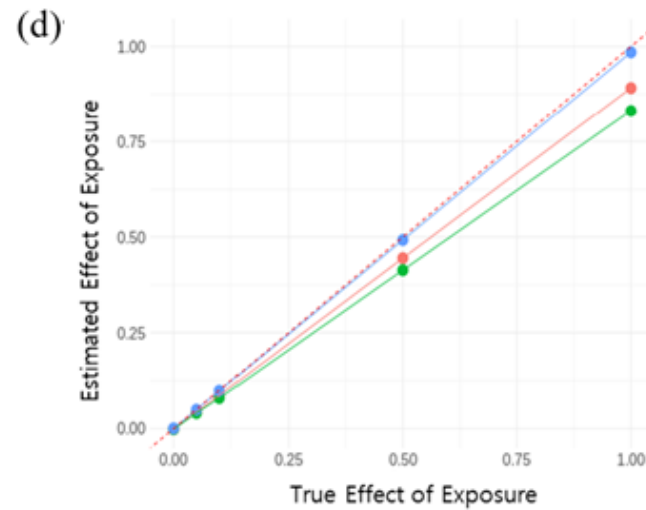
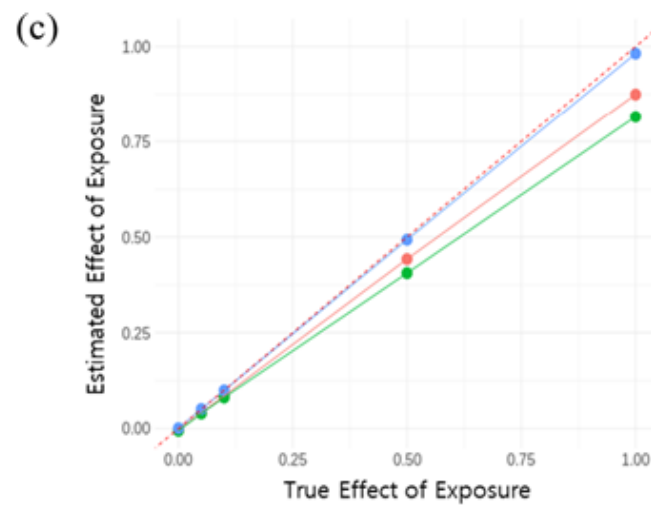
- When  $\gamma_j = \mathbf{0.01}$  or  $\mathbf{0.03}$ , the  $F$ -statistic was 1.40 or 4.62, falling well below the conventional threshold of 10 and thus categorized as a **weak** instrument strength.
- For  $\gamma_j = \mathbf{0.05}$  or  $\mathbf{0.1}$ , the  $F$ -statistic increased to 11.09 or 41.44, indicating a **moderate** instrument strength.
- For  $\gamma_j = \mathbf{0.3}$  or  $\mathbf{0.5}$ , the  $F$ -statistic increased to 365.68 or 1014.42, indicating a **strong** instrument strength.
- The performance of three causal effect estimation methods—TSLS, IVW, and WME were evaluated.

# Instrument Strength

$\gamma_j = 0.01$   
weak instrument



$\gamma_j = 0.3$   
moderate instrument

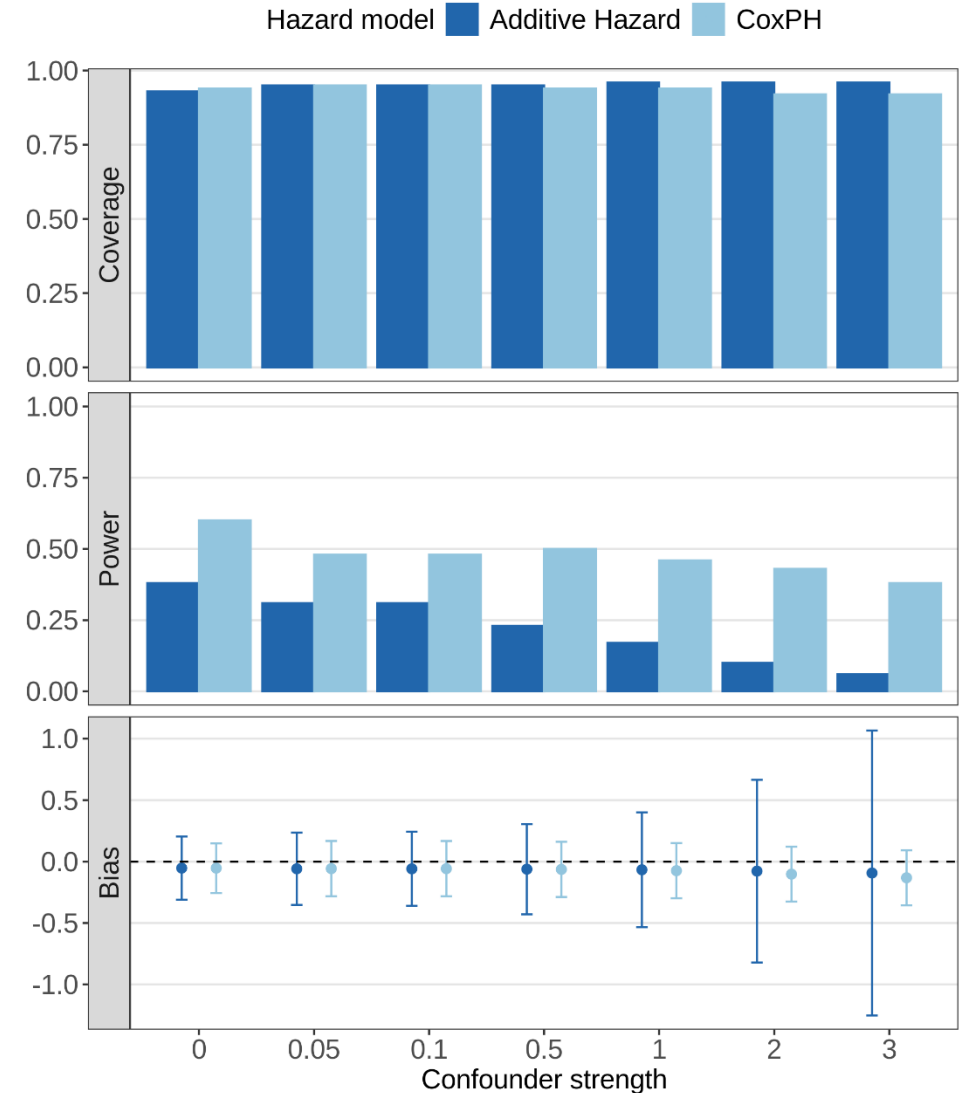


● IVW  
● Weighted Median  
● TSLS

# Confounder Strength

## • Confounder Strength

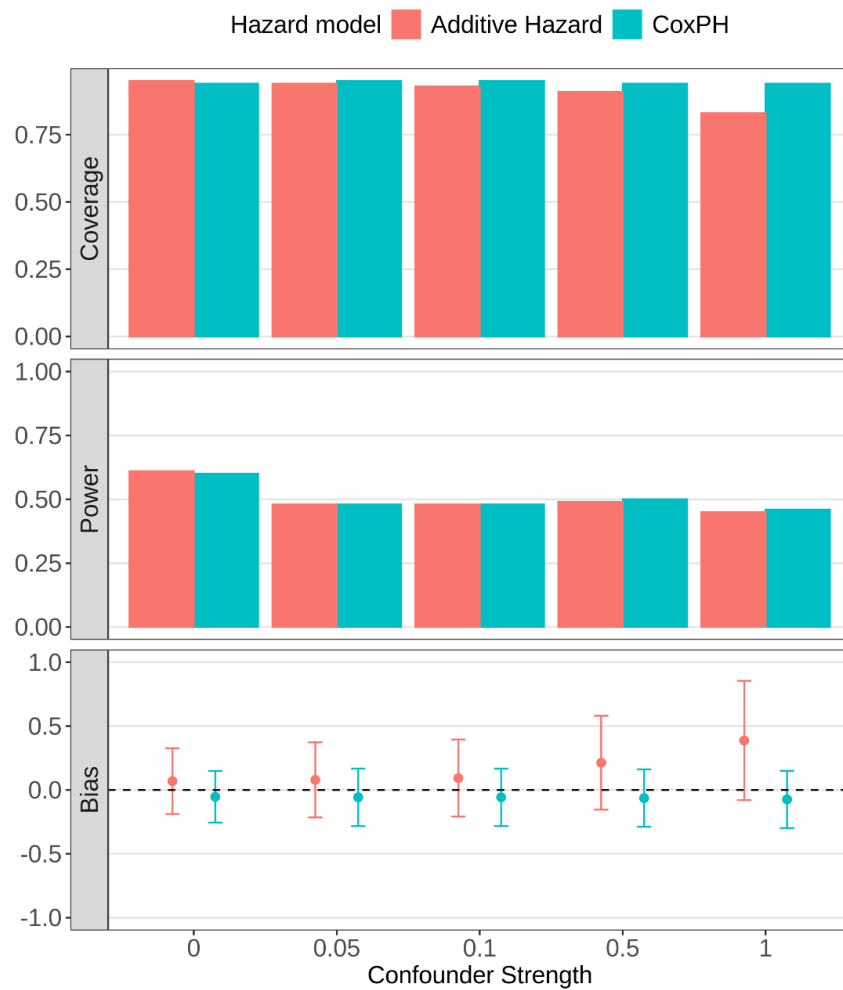
- The simulation was conducted under the setting  $\gamma_j = 0.1$ ,  $\beta_x^a = \beta_x^m = 0.5$ , and  $h_0^m = 1$ , with  $h_0^a$  determined adaptively according to the confounding effect.
- Across all levels of confounding, both models maintained nominal coverage near 95%, indicating valid inference.
- However, power decreased steadily with increasing confounder strength, with the additive model showing a more pronounced reduction compared to the CoxPH model.



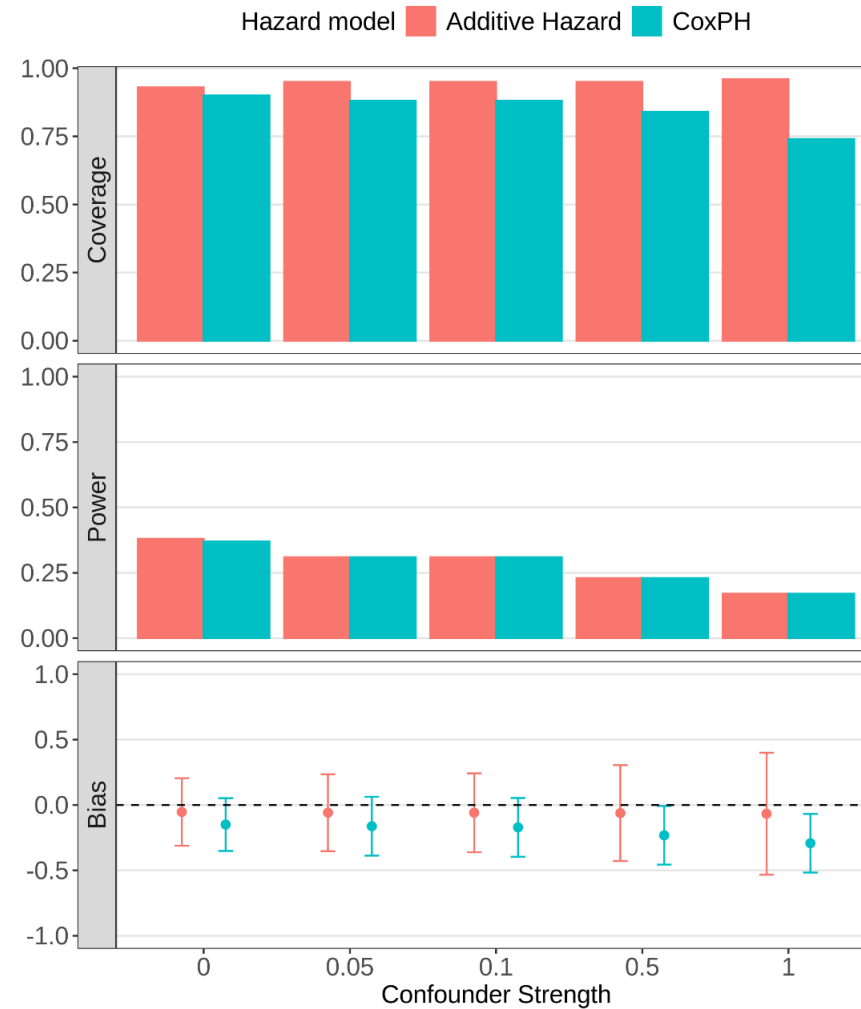
# Model Misspecification

- Data were generated from CoxPH model and analyzed with both models.
  - The CoxPH estimator achieves superior power and exhibits more stable bias profiles compared to the additive hazard model, particularly as the degree of confounding increases.
- Data were generated from Additive model and analyzed with both models.
  - The additive estimator achieves superior power and exhibits more stable bias profiles compared to the CoxPH model, particularly as the degree of confounding increases.

# Model Misspecification



Performance of causal effect estimators with data generated from the **CoxPH model**



Performance of causal effect estimators with data generated from the **Additive model**

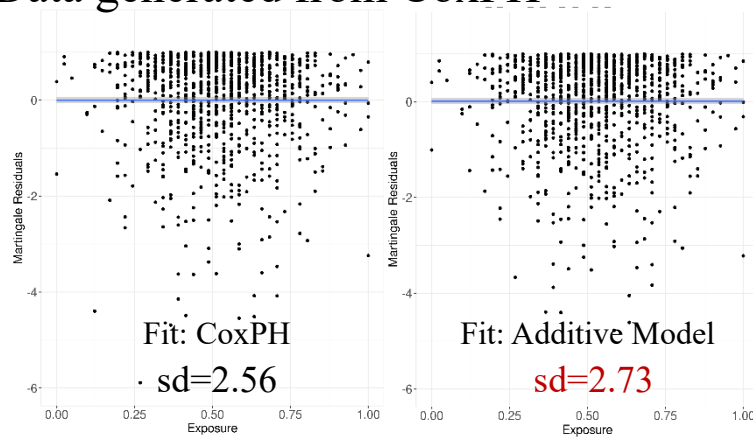
# Goodness of Fit

- Martingale Residuals

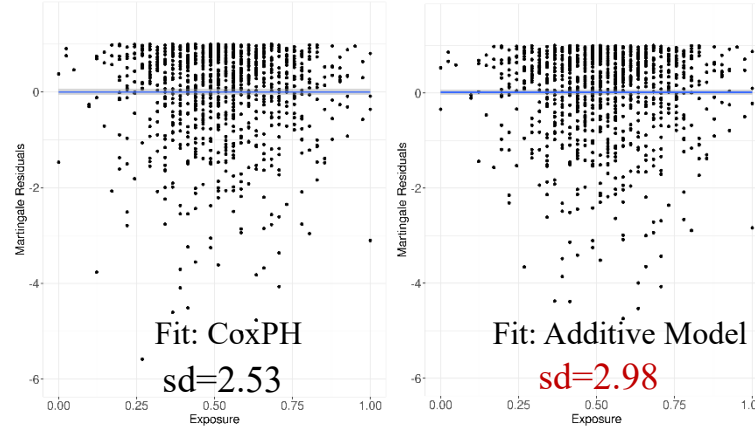
- The additive hazard parameter  $h_0^a$  was set to the value estimated from the objective function under each condition.
- The simulation design was based on a CoxPH model with a  $h_0^m$  of 0.241, no confounding, and varying exposure effect sizes ( $\beta_x = 0.5, 1, 2, 3$ ).
- When data were generated from CoxPH model, additive hazard model produced larger variable residuals. The **standard deviation of the residuals** increased with larger exposure effects ( $sd = 2.73, 2.98, 5.63$ ), indicating growing model misspecification.
- When data were generated from additive model, CoxPH model exhibited larger residual dispersion. The **standard deviation of the residuals** from the CoxPH fits increased with larger exposure effects ( $sd = 2.84, 4.28, 4.45$ ), indicating growing model misspecification.

# Martingale Residuals

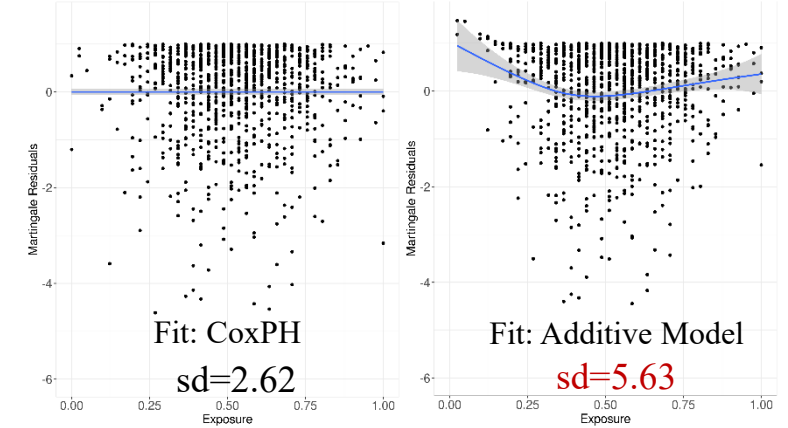
Data generated from CoxPH



$\beta$  of exposure = 0.5

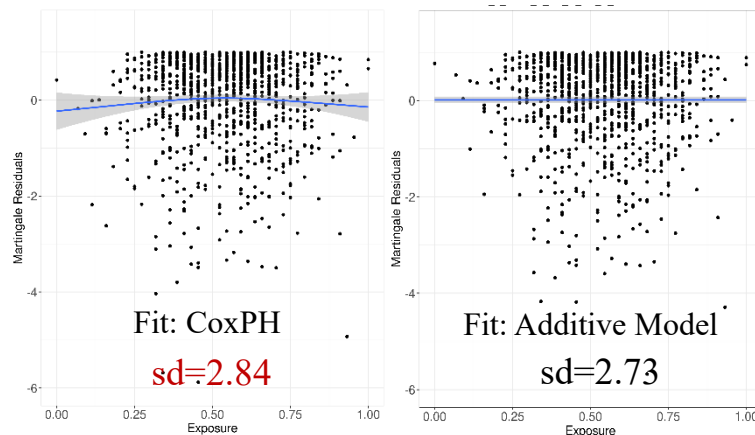


$\beta$  of exposure = 1

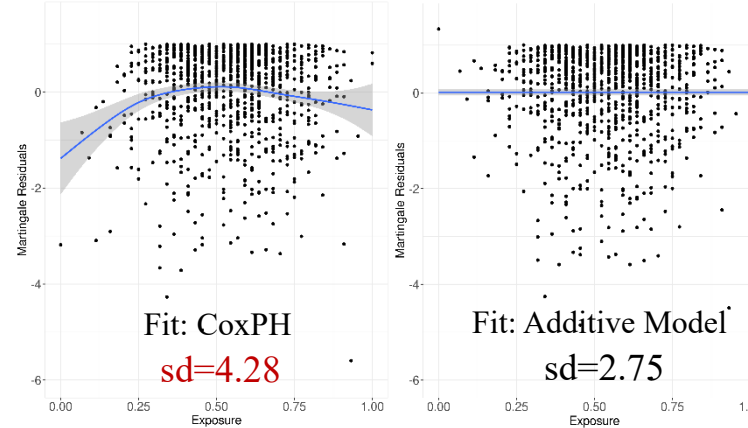


$\beta$  of exposure = 3

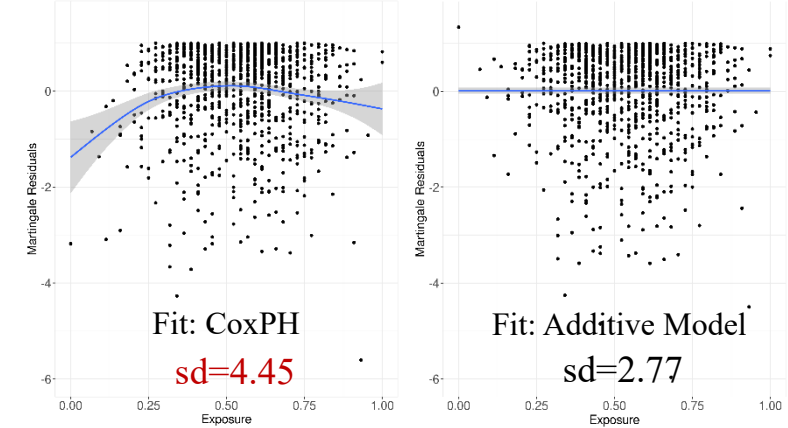
Data generated from Additive Models



$\beta$  of exposure = 0.5



$\beta$  of exposure = 1

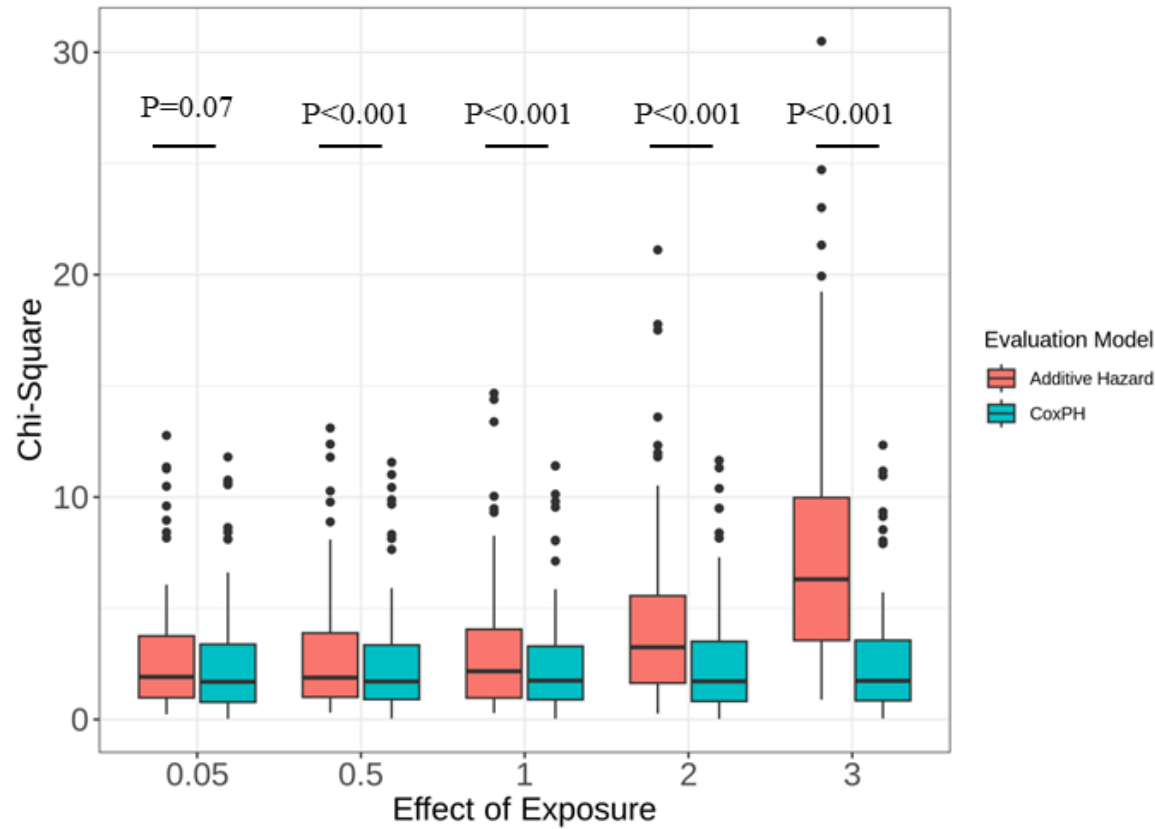


$\beta$  of exposure = 3

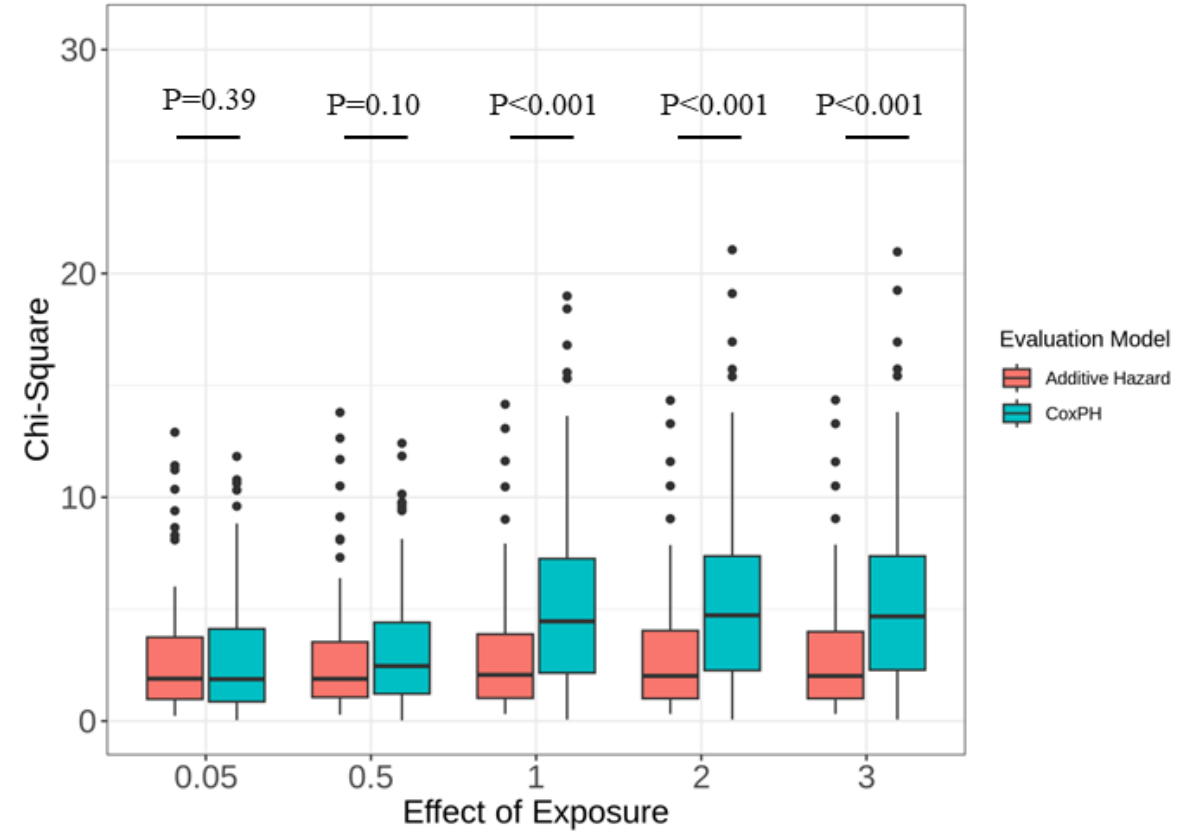
# Goodness of Fit

- Martingale Chi-square statistics
  - When the data were generated from CoxPH model, additive hazard model produced larger **martingale chi-square statistics**.
  - Under data generated from additive model, CoxPH model produced larger **martingale chi-square statistics** as the exposure effect increased, indicating inadequate model fit.
  - **Paired t-tests** across  $\beta_x$  values showed significantly smaller chi-square statistics for the correctly specified model.
- Sum of Arjas Residuals
  - When data were generated from CoxPH model, additive model produced larger **sum of Arjas residuals**.
  - When data were generated from additive model, CoxPH model exhibited larger **sum of Arjas residuals**.

# Martingale Chi-square Statistics

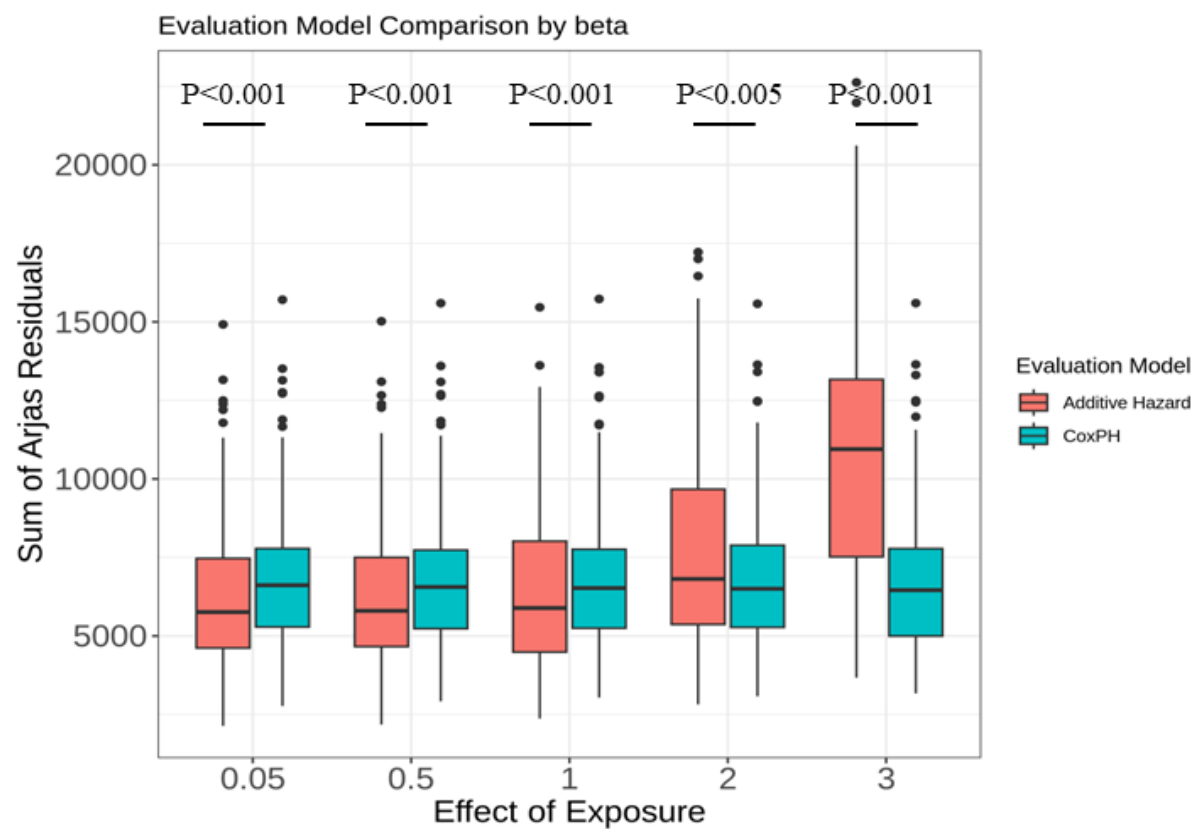


Data generated from CoxPH

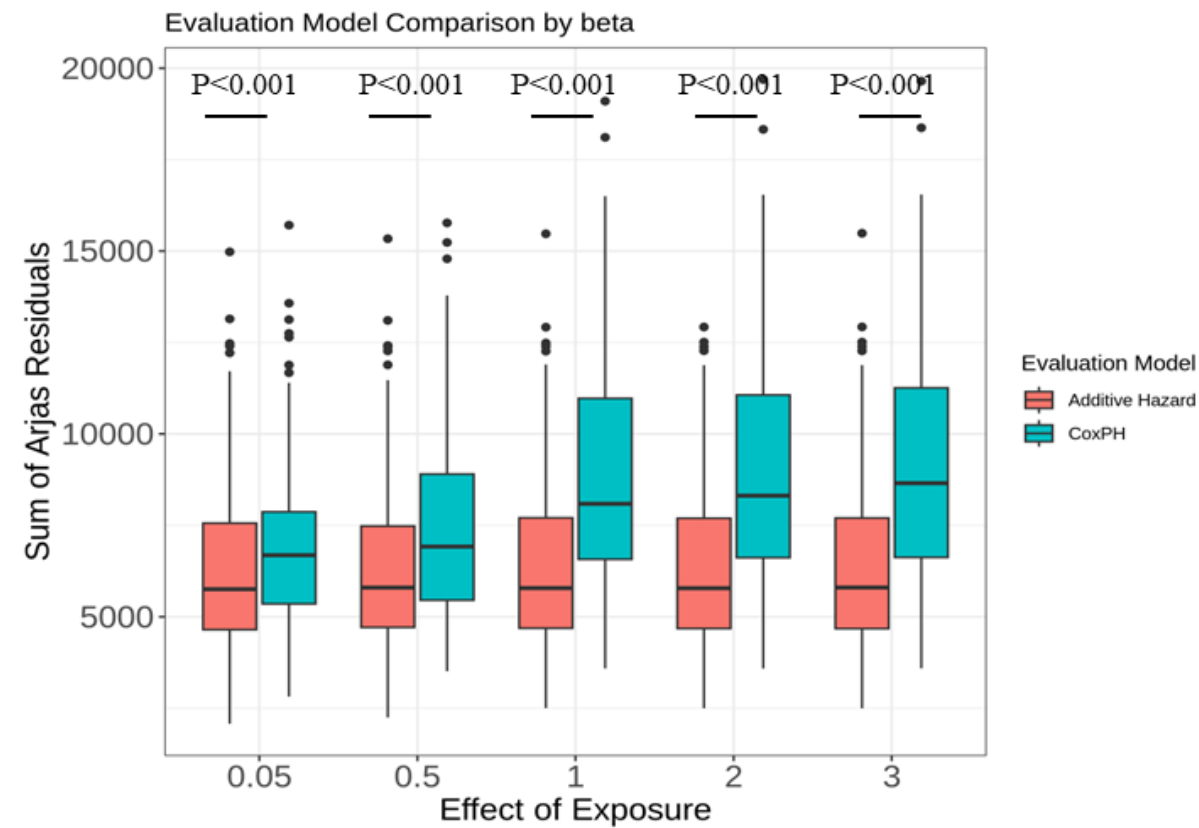


Data generated from Additive Models

# Sum of Arjas Residuals



Data generated from CoxPH



Data generated from Additive Models

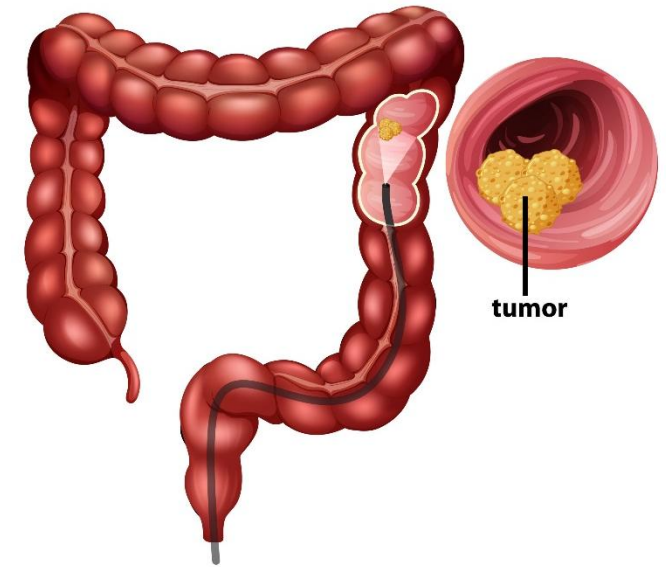
# Discussion

- Causal inference in survival-based MR critically depends on selecting a model whose structural assumptions align with the **true data-generating mechanism**.
- Both hazard models showed reduced bias and improved precision with **stronger instruments** and **weaker confounding**.
- **Model misspecification** introduced non-negligible bias in causal estimates under both hazard models, underscoring the necessity of principled model selection and diagnostic verification.
- Each hazard model achieved optimal goodness of fit when its structural form matched the data-generating mechanism, as confirmed by **smaller martingale chi-square statistics** and **Arjas residuals** under correct specification.

# Survival MR for Gene Expression on Prognosis in Colorectal Cancer

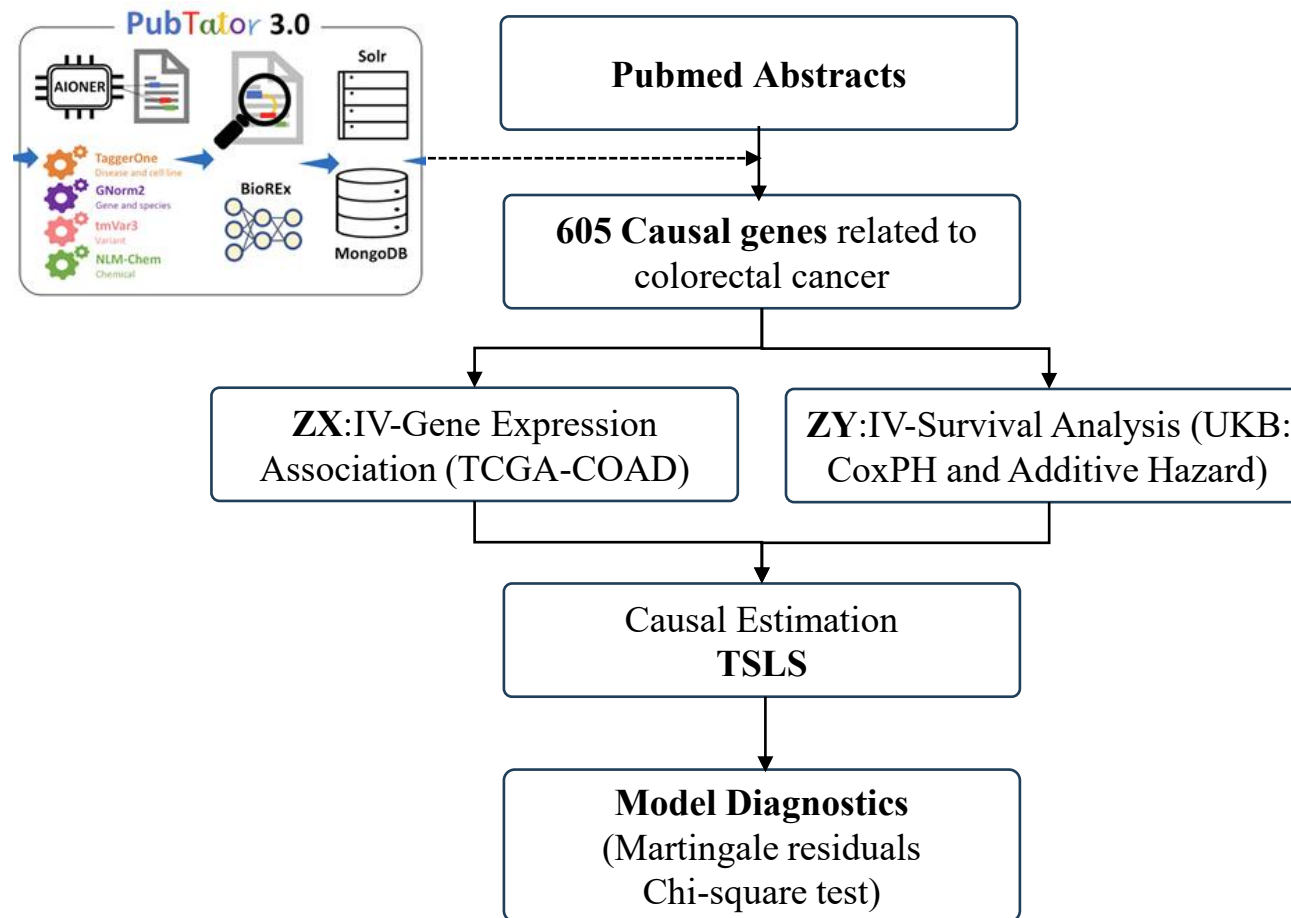
# Colorectal Cancer (CRC)

- **CRC** remains one of the leading causes of cancer-related mortality worldwide.
- No curative or causally targeted therapy yet exists and treatment relies on **chemotherapy regimens**.
- Therefore, identifying **causal molecular determinants** of prognosis remains a critical need for developing mechanism-based and personalized therapeutics.



# Overview of Study

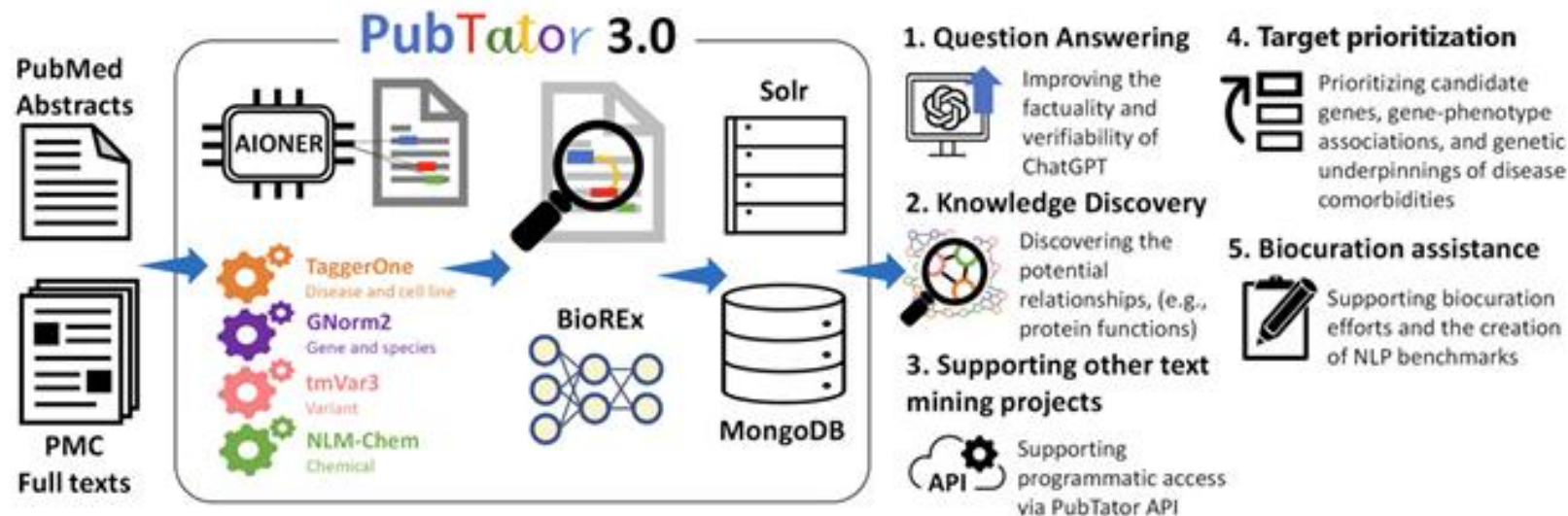
## Overall Analytic Flow for Causal Gene Identification and Model Evaluation



# Gene Candidates in Colorectal Cancer via Text Mining

- Large-scale **text mining** of PubMed abstracts was conducted using PubTator3.0.
- A total of **605 causal gene–disease associations** related to colorectal cancer were extracted from the published literature, defined by causal indicators such as “increase”, “decrease” and “caused by” (Pubtator3.0 tutorial).

Graphical abstract

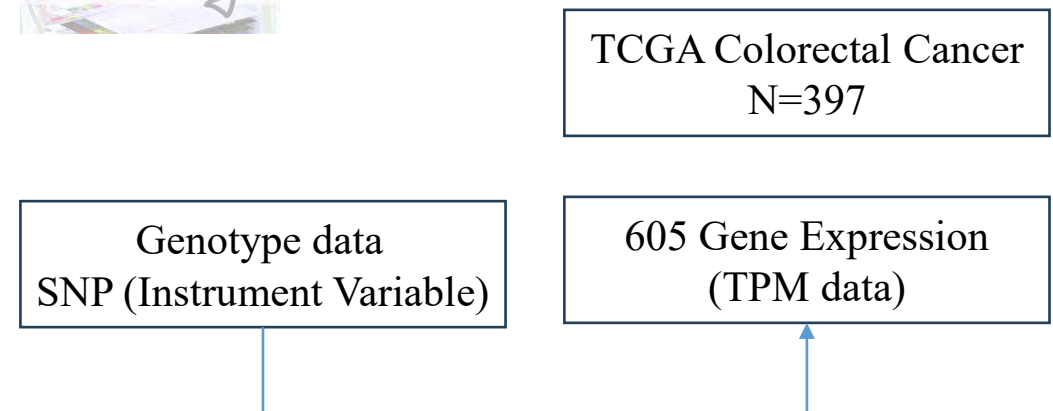


# Exposure Data

- TCGA Data
  - The exposure data were obtained from The Cancer Genome Atlas Colon Adenocarcinoma (**TCGA-AOAD**), containing a total of **397 colorectal cancer patients** with available genotype and transcriptomic data.
  - Genotype data were provided in PLINK binary format and contained **11,521,695 SNPs**. After excluding variants with **MAF < 0.01**, a total of **6,534,214 SNPs** were retained.
  - RNA-seq gene expression data (TPM) contains **20,600** transcripts for 605 genes.



## TCGA Colorectal Cancer



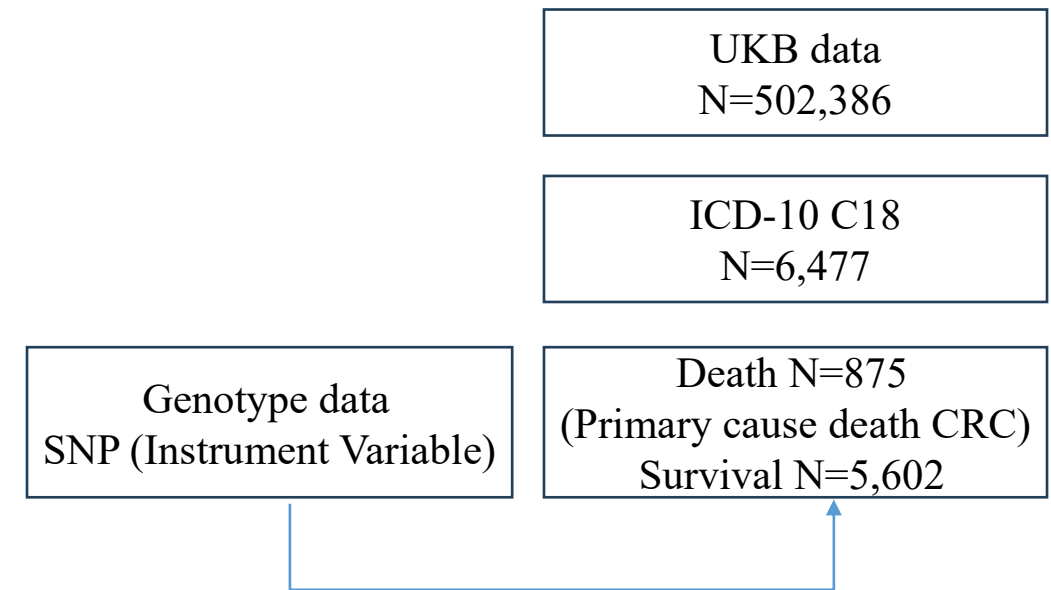
# Outcome Data

- UK Biobank Data

- The outcome data were derived from the **UK Biobank**, comprising **502,386** participants.
- The same **SNP sets** as TCGA were considered.
- CRC cases were identified using the ICD-10 C18, resulting in **6,477 patients**.
- 875 deaths** were attributed primarily to CRC. The event indicator was defined as death due to CRC, with all others including deaths due to other diseases treated as censored.
- Survival time** was calculated as the interval between the date of CRC diagnosis and the date of death, or the date of last follow up on January 20th, 2023.



## UKB Colorectal Cancer



# Association Studies

- eQTL in TCGA
  - eQTLs were identified using PLINK 2.0 **linear regression**, adjusting for age, sex, and PC1–PC5, across SNP–gene expression tests with a significance threshold of  $p < 1 \times 10^{-5}$ .
  - LD clumping (500 kb,  $p < 0.0001$ ,  $r^2 < 0.3$ ) was performed using the 1000G Ph3 reference panel to obtain independent index SNPs.
- Survival Analysis in UK Biobank
  - Based on instrumental SNPs identified from the eQTL analysis, associations of gene-specific PRS with colorectal cancer survival were evaluated in the UK Biobank.
  - **Survival analyses** were conducted under both **CoxPH** and **additive** models. In each model, gene-specific PRS as well as age, sex, and PC1–PC5 were included.

# Two-Sample MR Analysis

- Two Sample MR
  - **Exposure:** Genetic associations with gene expression ( $Z \rightarrow X$ ) in TCGA
  - **Outcome:** Genetic associations with survival outcomes ( $Z \rightarrow Y$ ) in UK Biobank
- TSLS Approach for Survival Data
  - **Step 1:** Let  $\hat{\gamma}_j$  denote the estimated effect of SNP  $j$  on the exposure obtained from the TCGA, and  $G_{ij}$  denote the genotype of individual  $i$  at SNP  $j$  in UK Biobank. The gene-specific PRS for individual  $i$  is defined as  $\tilde{X}_i = \sum_{j=1}^p \hat{\gamma}_j G_{ij}$  with  $p$  significant eQTL SNPs .
  - **Step 2:** The causal effect of the PRS on survival was then estimated by fitting one of the following hazard models:  
CoxPH:  $h(t | \tilde{X}_i) = h_0(t) \exp(\beta \tilde{X}_i)$ , Additive:  $h(t | \tilde{X}_i) = h_0(t) + \beta \tilde{X}_i$

# Survival Model Adequacy of Gene-Specific PRSs

- **Martingale residual chi-square tests** were used to assess the goodness of fit for PRSs of **605 genes**.
- A **chi-square test p-value < 0.05** was considered evidence of **inadequate model fit**.
- For PRSs fitted by both models, the model with the **smaller** chi-square statistic was considered **superior**.

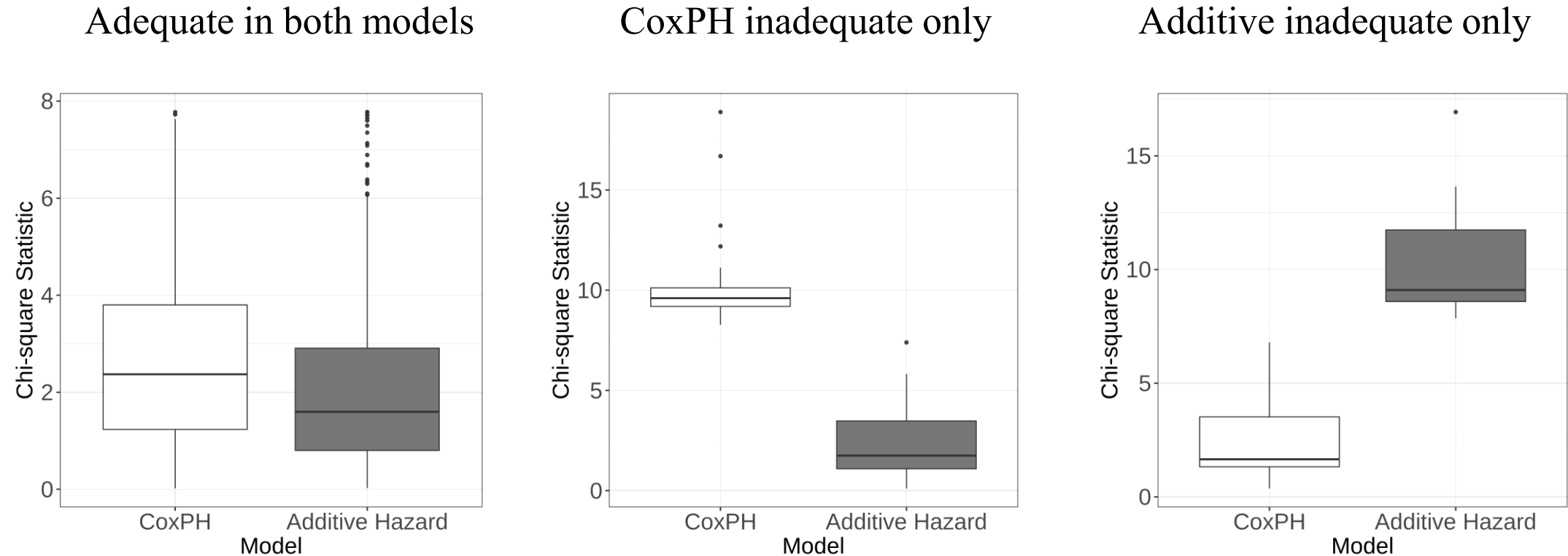
| Fit Category                   | Number of genes |
|--------------------------------|-----------------|
| <b>Adequate in both models</b> | <b>558</b>      |
| CoxPH inadequate only          | 32              |
| Additive inadequate only       | 15              |
| Inadequate in both             | 0               |
| Total                          | 605             |



| Superior Adequate Fit | Number of genes (% in 605 genes)  |
|-----------------------|-----------------------------------|
| CoxPH                 | 15+(220 in 558)= <b>235</b> (39%) |
| Additive Model        | 32+(338 in 558)= <b>370</b> (61%) |

**Additive model** provided a better overall fit more frequently than CoxPH model

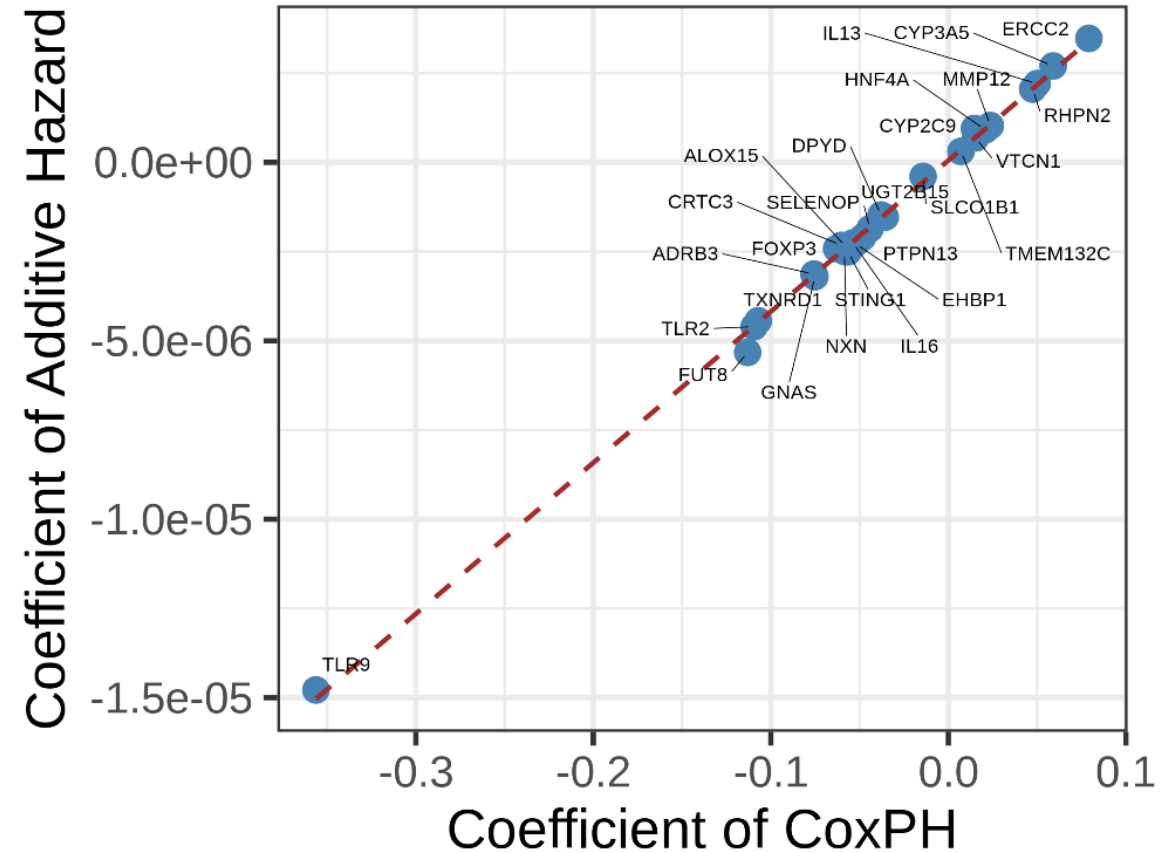
# Survival Model Adequacy of Gene Specific PRSs



Boxplot of Statistics of Martingale Residuals Chi-Square

# Causal Estimation

- Causality for 605 Gene-specific PRSs
  - Both CoxPH model and additive model were fitted with age, sex, PC1-5 and gene-specific PRSs.
  - The CoxPH model identified **31** gene-specific PRSs as significant ( $p < 0.05$ ).
  - The additive model identified **28** gene-specific PRSs as significant ( $p < 0.05$ ).
  - **27** gene PRSs were significant in both models, indicating substantial overlap in identified genetic signals.



Causal Effects and PRSs of 27 Commonly Significant Genes

# Causality and Model Adequacy

| Category                               | Adequate<br>in both models | CoxPH<br>inadequate only | Additive<br>inadequate only | Number of genes |
|----------------------------------------|----------------------------|--------------------------|-----------------------------|-----------------|
| Significance<br>in both models         | 27                         | 0                        | 0                           | 27              |
| Significance<br>in CoxPH model only    | 4                          | 0                        | 0                           | 4               |
| Significance<br>in Additive model only | 1                          | 0                        | 0                           | 1               |
| Non significance<br>in both models     | <b>526</b>                 | <b>15</b>                | <b>32</b>                   | 573             |
| Total                                  |                            |                          |                             | 605             |

Model adequacy and statistical significance of 605 gene-specific PRSs

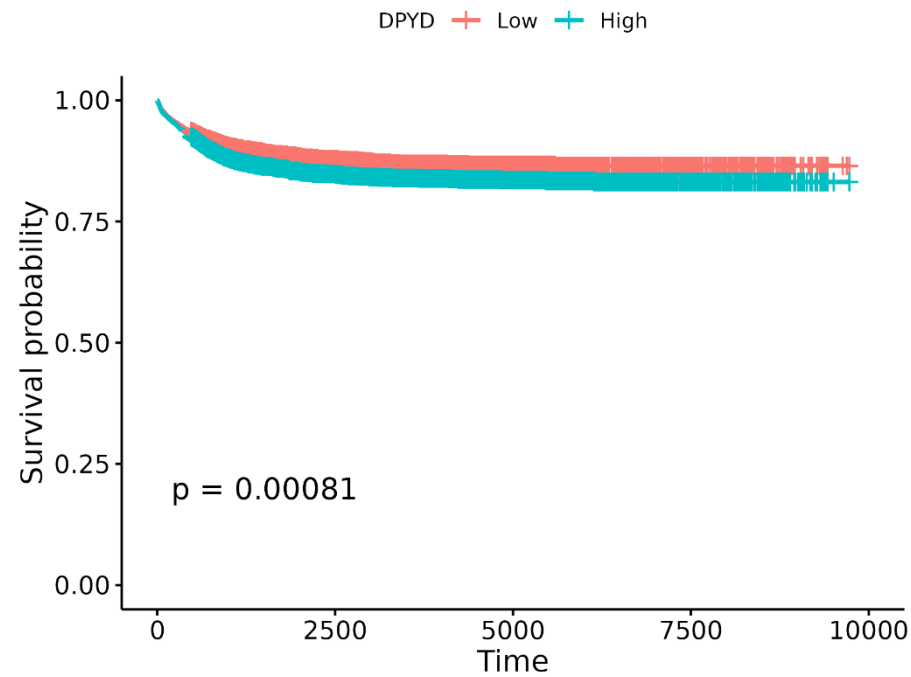
# PRS-Based Survival Differences

- Evaluation of PRS-Based Survival Differences in CRC Patients in UK Biobank
  - Survival data from 6,477 CRC patients in UK Biobank (5,602 censored, 875 deaths) were analyzed.
  - PRSs were divided into High vs. Low groups based on the median PRS values.
  - Among **235** genes for which the CoxPH model showed better adequacy, **15** genes exhibited significant (log-rank  $p < 0.05$ ).
  - Among **370** genes adequately fitted by the additive hazard model, **15** genes were significant.

| Adequate Fit    | Number of genes<br>(% in 605 genes) | Number of genes<br>(log rank $p < 0.05$ in CRC patients) |
|-----------------|-------------------------------------|----------------------------------------------------------|
| CoxPH           | 15+220 in 558=235(39%)              | 15                                                       |
| Additive Hazard | 15+338 in 558=370(61%)              | 15                                                       |

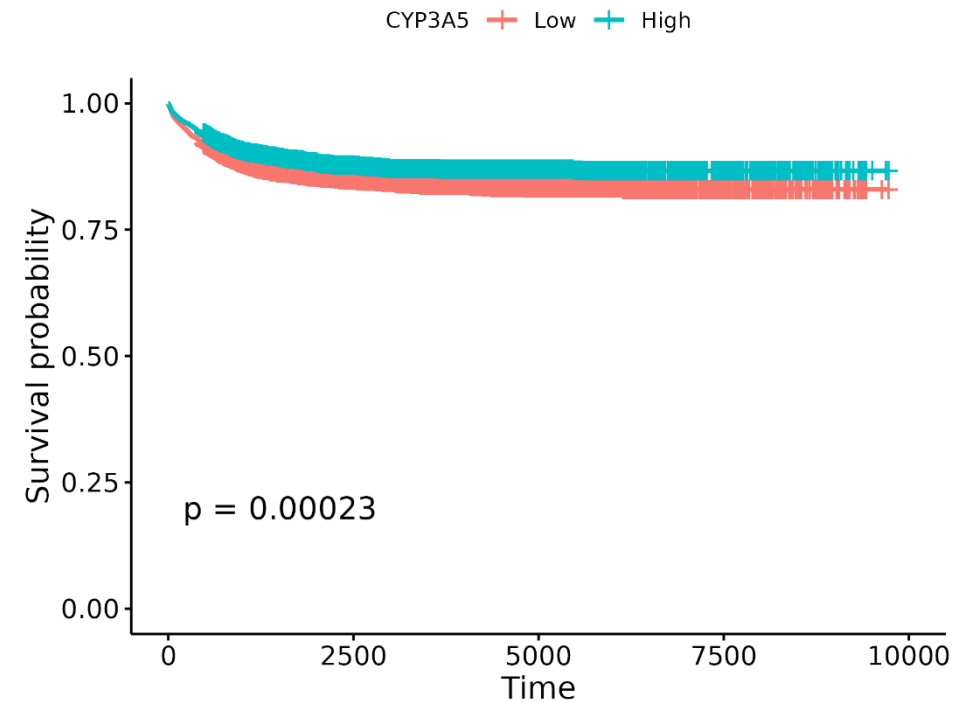
# PRS-Based Survival Differences

Kaplan-Meier Curve for DPYD



CoxPH is adequate

Kaplan-Meier Curve for CYP3A5



Additive model is adequate

# Discussion

- Across 605 gene-specific PRSs, **additive model** provided adequate fit more frequently than CoxPH model, suggesting that an additive structure may better capture the underlying risk patterns in colorectal cancer survival data.
- **Causal genetic signals** showed **substantial consistency** across models: **27** of 31 CoxPH-significant and **27** of 28 additive-significant gene PRSs were significant under both models, indicating robust identification of key causal genes.
- **Model adequacy varied substantially** across genes, underscoring that gene-specific model selection — rather than a single universal hazard model — is essential for reliable causal inference in large-scale survival MR.
- **PRS-based survival analyses** revealed distinct survival trajectories between high and low PRS groups in adequately fitted genes, providing further evidence for the biological relevance of identified causal gene effects.

# Future Work

- Future simulation studies should extend the current framework by varying **censoring rates**, various **baseline hazard functions**, and incorporating **additional covariates** to assess robustness across a broader range of data-generating scenarios.
- Model adequacy should be further evaluated through **sensitivity analyses** and external validation using independent cohorts beyond TCGA and UK Biobank.
- Beyond gene expression, integrating **additional clinical and molecular biomarkers** may provide more comprehensive and robust evidence for causal model selection and inference in colorectal cancer.

# Acknowledgement

## Soongsil University

- Dr. Jungeun Kim
- Dr. Fatemeh Yavartanoo
- Seunghwan Park
- Youngkwang Cho

## Yonsei University

- Prof. Sun Ha Jee
- Hyungwoo Seo

## Seoul National University

- Prof. Taesung Park
- Prof. Sungho Won
- Dr. Jinyeon Jo

## Harvard School of Public Health

- Prof. Liming Liang

Thank You!