Accounting for winner's curse and pleiotropy in two-sample Mendelian randomization

Chong Wu

Department of Statistics

Florida State University

Department of Biostatistics and Bioinformatics

Duke University School of Medicine

March 9, 2022

Outline

- **My research goal and motivation**
- Breaking winner's curse in two-sample MR
- Correcting pleiotropy in two-sample MR
- Other works and future directions

Research goal

My long-term research goal is to develop new methods, theories, and software to:

- identify likely causal risk factors and biomarkers for a complex disease (prostate cancer, Alzheimer's disease, etc.)
- enhance risk prediction to advance precision medicine

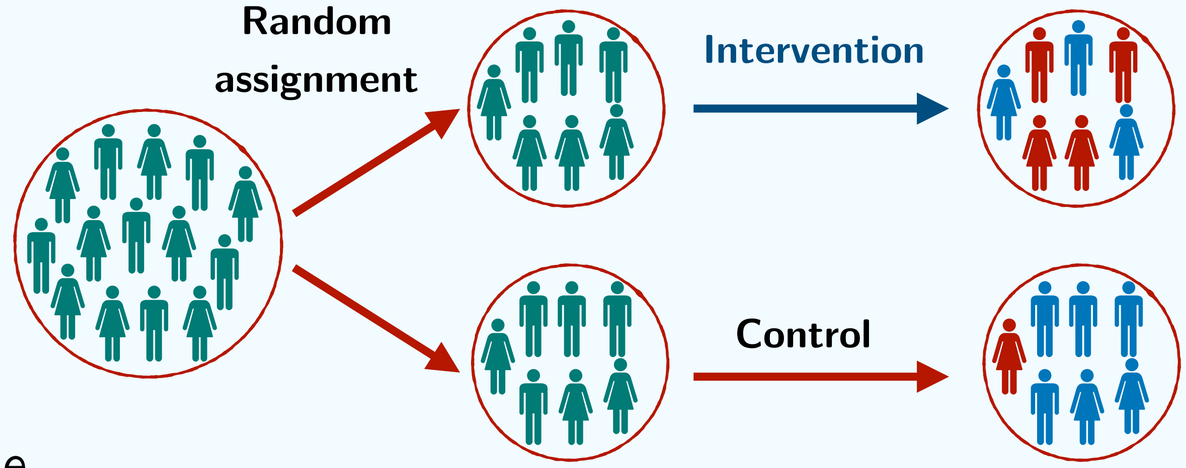
Research Interests: causal inference (Mendelian randomization), machine learning, statistical genetics (polygenic risk score, integrative analysis, TWAS, PWAS)

Data we work on: UK Biobank (genotype, risk factors, & disease status), GTEx (splicing, gene expression, & genotype), ROS/MAP (protein & genotype), GWAS summary data, functional annotations, DNA methylation

Causal inference in observational data

Does X (risk factor) cause Y (complex disease)?

- **Example:** Does smoking cause lung cancer?
- Randomized clinical trial
  - Gold standard
  - Randomization balances participant characteristics between the groups
- **Challenges:** randomized clinical trial would be both not feasible and unethical



Causal inference in observational data

Example: identify causal biomarkers for a complex disease

Why:

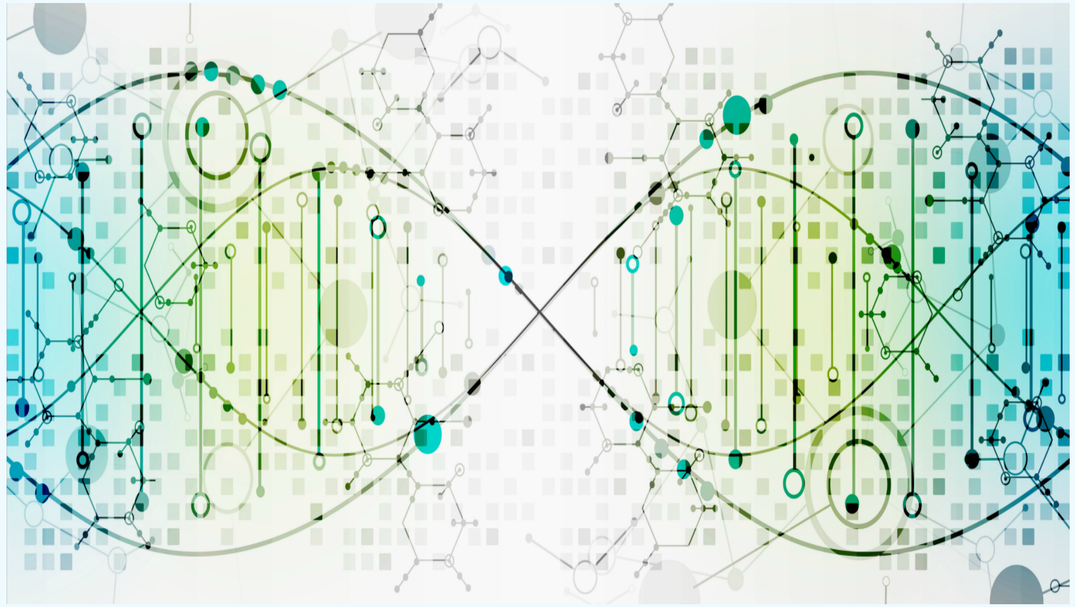
- understand the etiology
- drug development

Challenges:

- the number of biomarkers is large
- biomarkers are correlated

Goal:

identify likely causal biomarkers by using observational data

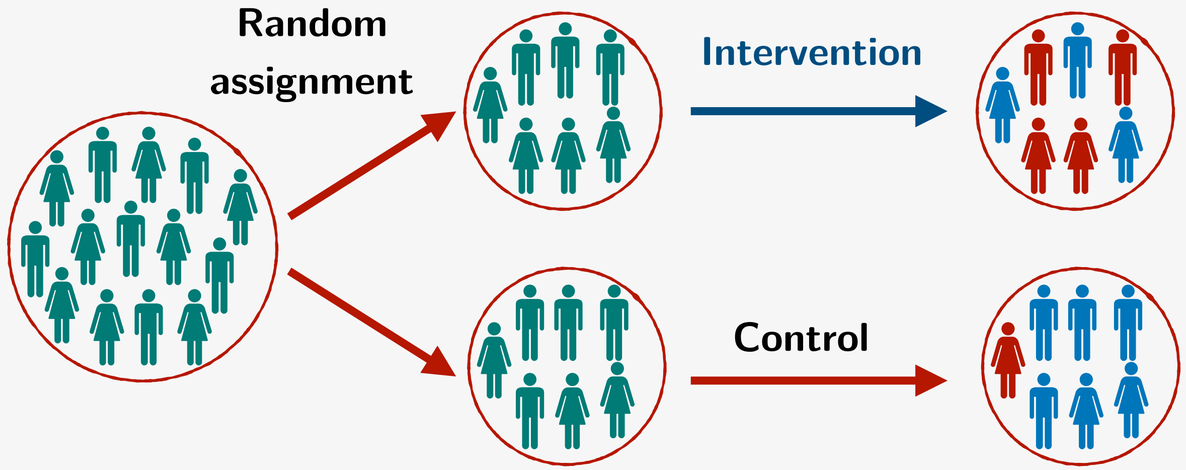


This figure is downloaded from Google Image

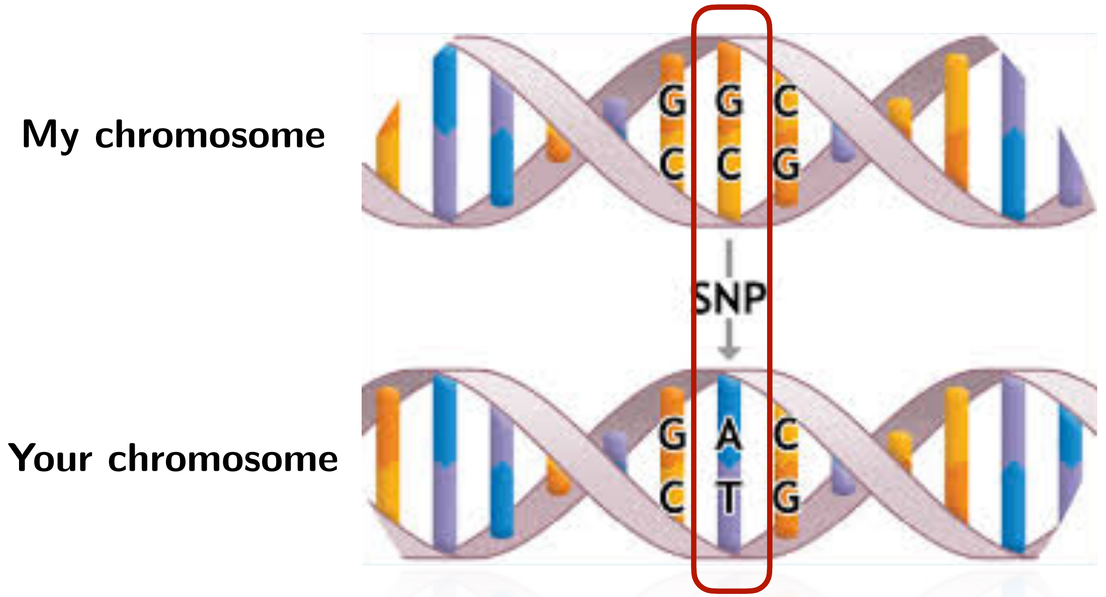
Mendelian randomization

Randomized clinical trial

- Gold standard
- Randomization balances participant characteristics between the groups



- Genome: genetic information encoded in 23 chromosome pairs
- SNP
  - variation in a single base pair
  - inherited randomly and fixed at conception

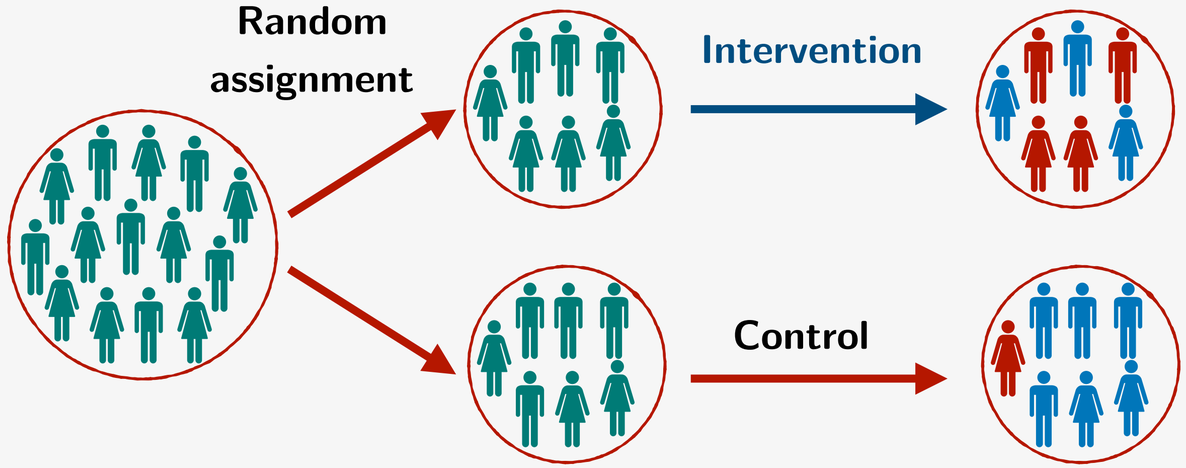


This figure is downloaded from Google Image

Mendelian randomization

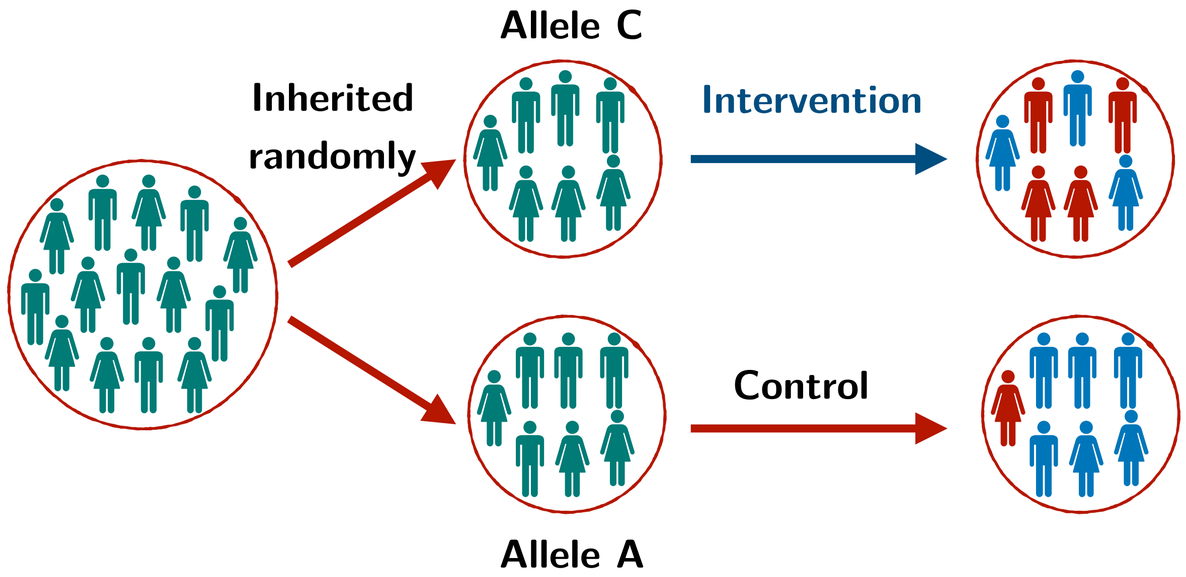
Randomized clinical trial

- Gold standard
- Randomization balances participant characteristics between the groups

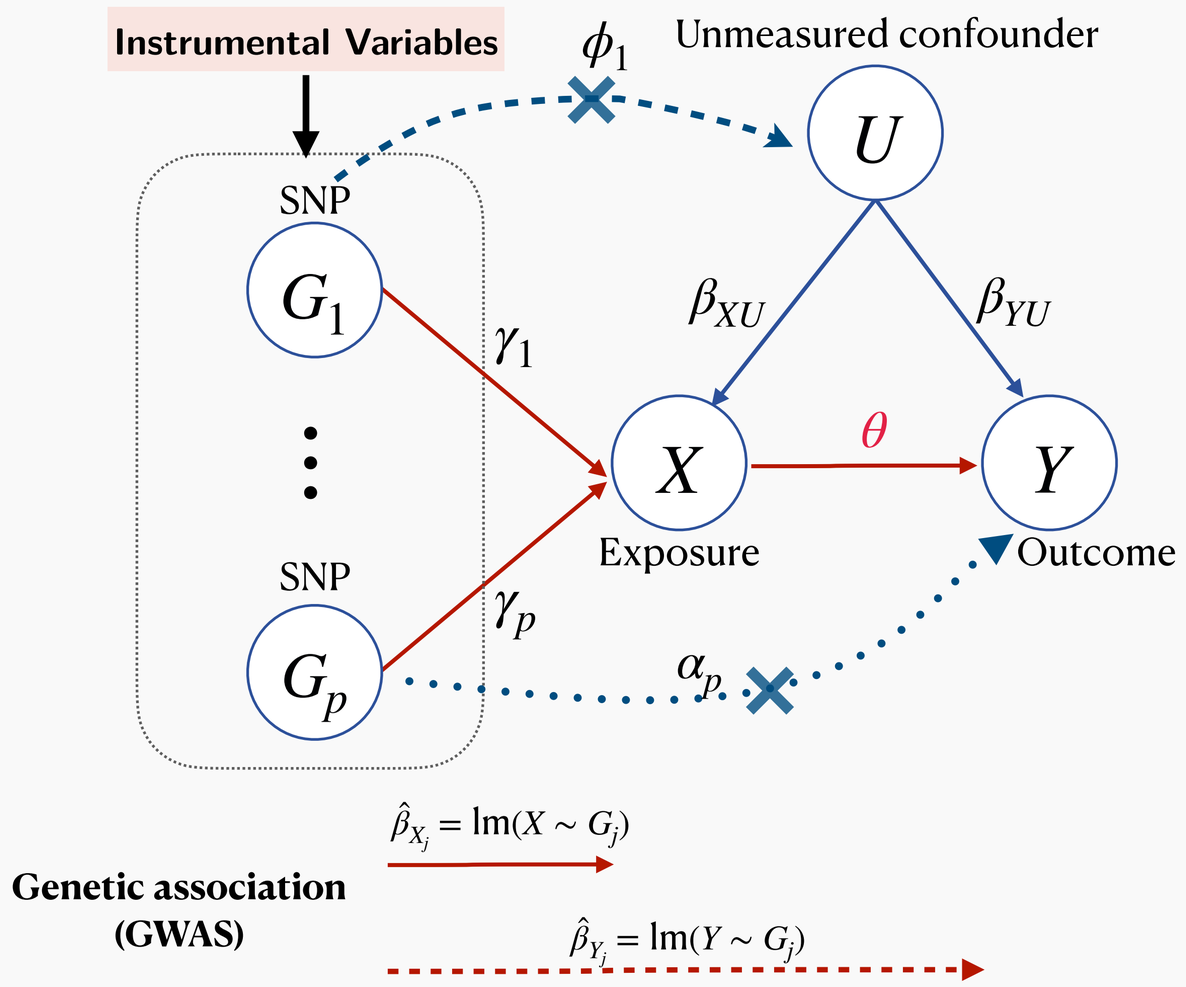


Hypothetical example

- Allele A: not smoking
- Allele C: smoking
- Not associated with unmeasured confounding factors (e.g., drinking)
- No direct effect on the outcome (e.g., lung cancer)



Mendelian randomization



Structure equation model:

$$\beta_{X_j} = \gamma_j + \phi_j \cdot \beta_{XU}$$

$$\beta_{Y_j} = \beta_{Y_j,\mathrm{M}} + \beta_{Y_j,\mathrm{D}} = \theta \cdot \beta_{X_j} + (\alpha_j + \phi_j \cdot \beta_{YU}) \triangleq \theta \cdot \beta_{X_j} + r_j$$

SNP j is a valid instrumental variable (IV) if

- **Relevance:** $\gamma_j \neq 0$
- **Independence:** $\phi_j = 0$
- **Exclusion restriction:** $\alpha_j = 0$

For a valid IV SNP j:

$$\beta_{X_j} = \gamma_j$$

$$\beta_{Y_j} = \theta \cdot \beta_{X_j}$$

Two-sample summary-data MR

Two-sample MR setup:

	Original data	Summary data
Exposure GWAS	$\left\{(X_i^*, G_{ij}^*)\right\}_{i=1}^{n_X}$	$\left\{(\hat{\beta}_{X_j}, \sigma_{X_j})\right\}_{j=1}^{p}$
Outcome GWAS	$\left\{(Y_i, G_{ij})\right\}_{i=1}^{n_Y}$	$\left\{(\hat{\beta}_{Y_j}, \sigma_{Y_j})\right\}_{j=1}^{p}$

Strengths of two-sample MR:

- Increase the power
- Expand the scope of MR studies

Inverse variance weighted (IVW) estimator:

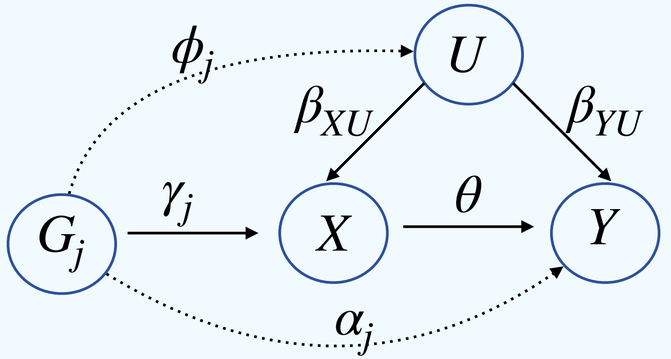
- Assume all IVs are valid
- Assume no measurement error: $\hat{\beta}_{X_j} = \beta_{X_j}$
- $\hat{\beta}_{Y_j} = \theta \cdot \hat{\beta}_{X_j} + \epsilon_j$
- The IVW estimator:

$$\hat{\theta}_{\texttt{IVW}} = \frac{\sum_{j=1}^{p} \hat{\beta}_{X_j}\hat{\beta}_{Y_j}/\sigma_{Y_j}^2}{\sum_{j=1}^{p} \hat{\beta}_{X_j}^2/\sigma_{Y_j}^2}$$

Motivation

Assumptions: SNP j is a valid IV if

- **Relevance:** $\gamma_j \neq 0$
- **Independence:** $\phi_j = 0$
- **Exclusion restriction:** $\alpha_j = 0$



Motivation: break the "winner's curse" bias induced by the relevance assumption

Motivation: build robust and powerful estimators when valid IV assumptions are violated

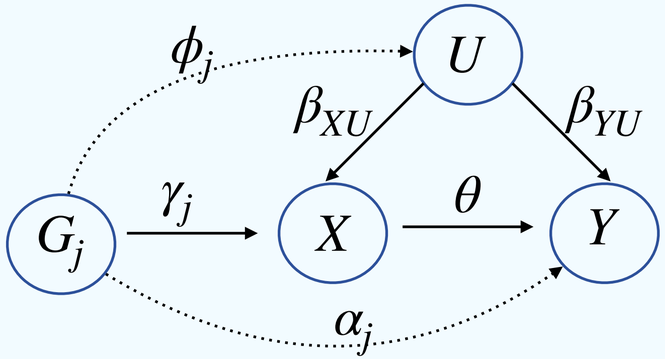
Outline

- My research goal and motivation
- **Breaking winner's curse in two-sample MR**
- Correcting pleiotropy in two-sample MR
- Other works and future directions

Winner's curse

- To meet relevance assumption $\gamma_j \neq 0$, we select SNP j if

$$\left| \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} \right| > \lambda, \quad \lambda = \Phi^{-1}(1-\alpha), \quad j = 1, \ldots, p$$



- MR analysis relies on the assumption:

$$\begin{bmatrix} \hat{\beta}_{Y_j} \\ \hat{\beta}_{X_j} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} \beta_{Y_j} \\ \beta_{X_j} \end{bmatrix}, \begin{bmatrix} \sigma^2_{Y_j} & 0 \\ 0 & \sigma^2_{X_j} \end{bmatrix} \right), \quad j = 1, \ldots, p$$

- In two sample MR, one often uses the **same** exposure GWAS to select IVs and estimate β_{X_j}, thus $\hat{\beta}_{X_j}$ follows a **truncated** normal distribution, leading to a downward bias in $\hat{\theta}_{\text{IVW}}$
- Ideally, we hope to use a third independent GWAS data to select IVs (often impractical)

Bias characteristics in simulation studies

Simulation Design

- True causal effect: $\theta = 0.2$
- Dimension, sample size: $p = 200{,}000$, $n_X = n_Y = 100{,}000$
- SNP-exposure effect:

$$\beta_{X_j} \sim \pi \cdot \text{Turncated Normal}\big(0,\ \varepsilon_x^2; (-\infty, -a], [a, +\infty)\big) + (1-\pi)\cdot\delta_0$$

$$\varepsilon_x^2 = 1 \times 10^{-5}, \quad \boxed{a \in \{0, 0.001, 0.002, \cdots, 0.011\}}$$

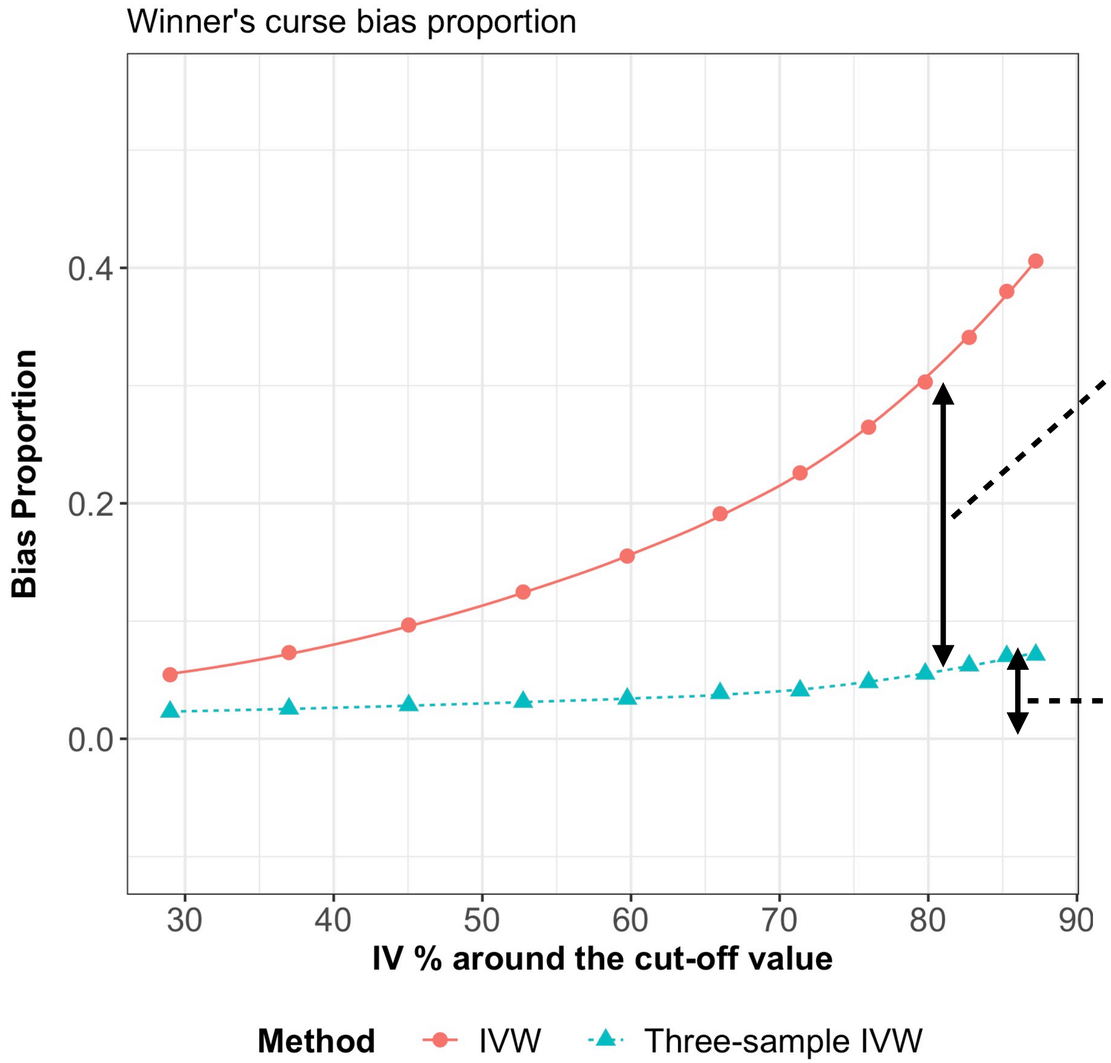
When a increases, more SNPs are selected

- Measurement errors: $\sigma_{X_j} = 1/\sqrt{n_X}$, $\sigma_{Y_j} = 1/\sqrt{n_Y}$
- Selection cutoff: $\lambda = \Phi^{-1}\big(1 - \frac{\alpha}{2}\big) = 5.45$, $\alpha = 5 \times 10^{-8}$

- The proportion of SNPs (IVs) near the cutoff:

$$\frac{\#\ \text{of SNPs with p-value lies between } 5\times10^{-8} \text{ and } 5\times10^{-10}}{\#\ \text{of SNPs selected}}$$

Bias characteristics in simulation studies



Winner's curse bias: same exposure GWAS to select IVs and estimate β_{X_j}

Measurement error bias: ignore estimation error on β_{X_j}: $\hat{\beta}_{X_j} = \beta_{X_j}$

Key idea to correct 'winner's curse' bias

- In three-sample MR:

$$\hat{\beta}_{X_j} \perp \underbrace{\left|\frac{\hat{\beta}'_{X_j}}{\sigma'_{X_j}}\right| > \lambda}_{\text{on a third GWAS}} \quad \Longrightarrow \quad \mathbb{E}\left[\hat{\beta}_{X_j} \middle| \left|\frac{\hat{\beta}'_{X_j}}{\sigma'_{X_j}}\right| > \lambda\right] = \mathbb{E}[\hat{\beta}_{X_j}] = \beta_{X_j}$$

- **Q:** How to make $\hat{\beta}_{X_j}$ independent with the selection criterion in two-sample MR

Benefits of this idea

Rerandomized Inverse Variance Weighted (RIVW) estimator

Step 1. Randomized SNP selection

- Create a pseudo SNP-risk association for each SNP:

$$\left|\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j\right| > \lambda \quad \Longrightarrow \quad \text{Select SNP } j \quad \Longleftrightarrow \quad S_j = \left|\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j\right| - \lambda > 0$$

where $Z_j \sim N(0,\eta^2)$

Step 2. Rao-Blackwellization

- Construct an unbiased initial estimator:

$$\hat{\beta}^{\text{init}}_{X_j} = \hat{\beta}_{X_j} - \frac{Z_j \sigma_{X_j}}{\eta^2} \quad \text{satisfies} \quad \mathbb{E}\left[\hat{\beta}^{\text{init}}_{X_j} \,|\, \text{SNP } j \text{ is selected}\right] = \mathbb{E}\left[\hat{\beta}^{\text{init}}_{X_j}\right] = \beta_{X_j}$$

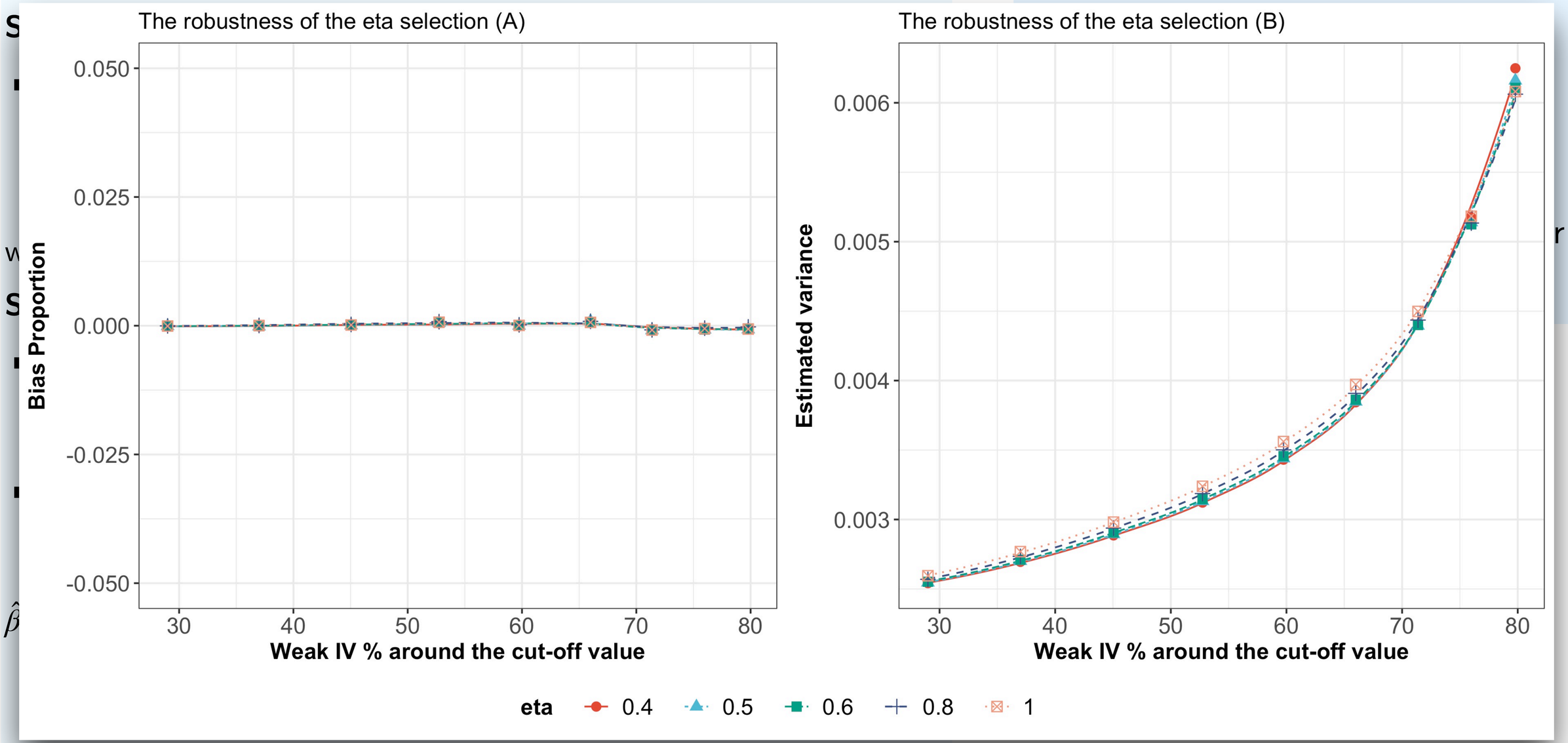
- Improve the initial estimator by Rao-Blackwellization

$$\hat{\beta}_{X_j,\texttt{RB}} = \mathbb{E}\left[\hat{\beta}^{\text{init}}_{X_j} \,|\, S_j > 0, \hat{\beta}_{X_j}\right] = \hat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta} \frac{\phi\left(\frac{\lambda}{\eta} - \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right) - \phi\left(-\frac{\lambda}{\eta} - \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right)}{\Phi\left(-\frac{\lambda}{\eta} + \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right) + \Phi\left(-\frac{\lambda}{\eta} - \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right)}$$

Step 1 comments:

- SNPs with large $\hat{\beta}_{X_j}/\sigma_{X_j}$ are indifferent to the Step 1
- $\eta = 0.5$ is a tuning parameter (our estimator is not sensitive to η)

Rerandomized Inverse Variance Weighted (RIVW) estimator



Rerandomized Inverse Variance Weighted (RIVW) estimator

Step 1. Randomized SNP selection

- Create a pseudo SNP-risk association for each SNP:

$$\left|\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j\right| > \lambda \quad \Longrightarrow \quad \text{Select SNP } j \quad \Longleftrightarrow \quad S_j = \left|\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j\right| - \lambda > 0$$

where $Z_j \sim N(0, \eta^2)$

Step 2. Rao-Blackwellization

- Construct an unbiased initial estimator:

$$\hat{\beta}^{\text{init}}_{X_j} = \hat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta^2} Z_j \text{ satisfies } \mathbb{E}\left[\hat{\beta}^{\text{init}}_{X_j} \middle| \text{SNP } j \text{ is selected}\right] = \mathbb{E}\left[\hat{\beta}^{\text{init}}_{X_j}\right] = \beta_{X_j}$$

- Improve the initial estimator by Rao-Blackwellization

$$\hat{\beta}_{X_j,\text{RB}} = \mathbb{E}\left[\hat{\beta}^{\text{init}}_{X_j} \middle| S_j > 0, \hat{\beta}_{X_j}\right] = \hat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta} \frac{\phi\left(\frac{\lambda}{\eta} - \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right) - \phi\left(-\frac{\lambda}{\eta} - \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right)}{\Phi\left(-\frac{\lambda}{\eta} + \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right) + \Phi\left(-\frac{\lambda}{\eta} - \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta}\right)}$$

Step 1 comments:

- SNPs with large $\hat{\beta}_{X_j}/\sigma_{X_j}$ are indifferent to the Step 1
- $\eta = 0.5$ is a tuning parameter (our estimator is not sensitive to η)

Step 2 comments:

- $\hat{\beta}^{\text{init}}_{X_j} = \hat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta^2} Z_j \perp \left|\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j\right| > \lambda$
- $\hat{\beta}_{X_j,\text{RB}}$ is also an unbiased estimator

Rerandomized Inverse Variance Weighted (RIVW) estimator

Restore the correct center after SNP selection

- After randomized selection + Rao-Blackwellization:

$$\hat{\beta}_{X_j,\mathrm{RB}} - \beta_{X_j} \,\Big|\, \text{SNP } j \text{ is selected} \overset{\text{i.i.d.}}{\sim} \mathcal{G}(0,\, \sigma^2_{X_j,\mathrm{RB}}), \quad j = 1,\ldots,p$$

- Classical two-sample MR:

$$\hat{\beta}_{X_j} - \beta_{X_j} \,\Big|\, \text{SNP } j \text{ is selected} \overset{\text{i.i.d.}}{\sim} \mathcal{TN}(\text{Bias},\, \sigma^2_{X_j}), \quad j = 1,\ldots,p$$

Details on $\hat{\sigma}^2_{X_j,\mathrm{RB}}$

RIVW Estimator

$$\hat{\theta}_{\mathtt{RIVW}} = \frac{\sum_{j\in\mathcal{S}_\lambda} \hat{\beta}_{Y_j}\hat{\beta}_{X_j,\mathrm{RB}}/\sigma^2_{Y_j}}{\sum_{j\in\mathcal{S}_\lambda} (\hat{\beta}^2_{X_j,\mathrm{RB}} - \hat{\sigma}^2_{X_j,\mathrm{RB}})/\sigma^2_{Y_j}}$$

To consider β_{X_j}'s are measured with error[1]

[1] Ye, T., Shao, J., & Kang, H. (2021). Debiased inverse-variance weighted estimator in two-sample summary-data mendelian randomization. *The Annals of Statistics*, *49*(4), 2079-2100.

Theoretical guarantee

Asymptotic Normality: Under certain regularity conditions, the RIVW estimator converges to a standard normal distribution after appropriate scaling

$$\mathsf{V}_{\texttt{RIVW}}^{-\frac{1}{2}}\left(\hat{\theta}_{\texttt{RIVW}}-\theta\right)\to\mathcal{N}(0,1)\,.$$

Consistent variance estimation: Under certain regularity conditions, we have

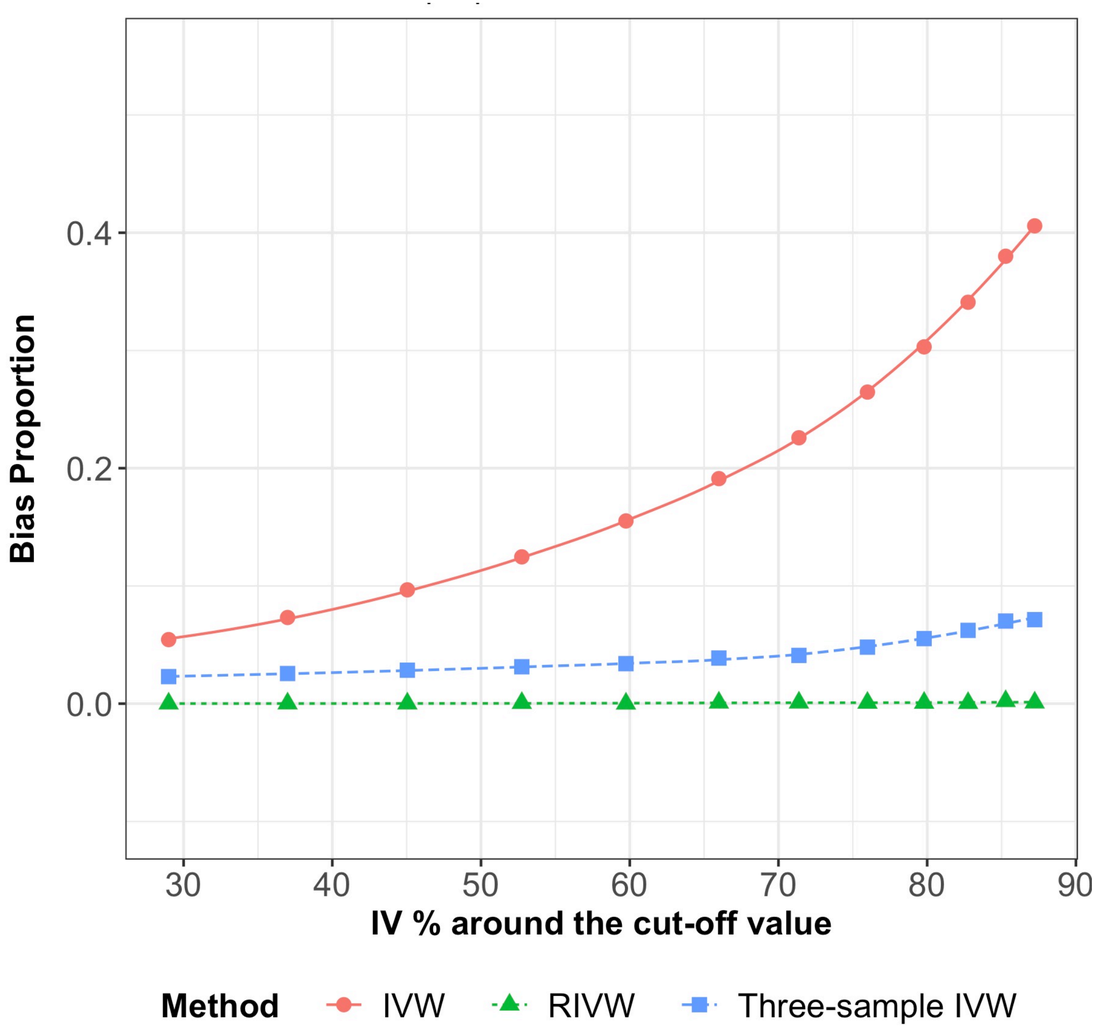
$$\frac{\hat{\mathsf{V}}_{\texttt{RIVW}}}{\mathsf{V}_{\texttt{RIVW}}}\xrightarrow{p}1.$$

Assumption Details

- A level $1-\alpha$ confidence interval can be constructed as:

$$\left[\ \hat{\theta}_{\texttt{RIVW}}-z_{\alpha/2}\sqrt{\hat{\mathsf{V}}_{\texttt{RIVW}}}\ ,\ \hat{\theta}_{\texttt{RIVW}}+z_{\alpha/2}\sqrt{\hat{\mathsf{V}}_{\texttt{RIVW}}}\ \right],\quad\text{where }\hat{\mathsf{V}}_{\texttt{RIVW}}=\frac{\sum_{j\in S_\lambda}\left(\hat{\beta}_{Y_j}\hat{\beta}_{X_j,\texttt{RB}}-\hat{\theta}_{\texttt{RIVW}}(\hat{\beta}^2_{X_j,\texttt{RB}}-\hat{\sigma}^2_{X_j,\texttt{RB}})\right)^2/\sigma^4_{Y_j}}{\left(\sum_{j\in S_\lambda}\left(\hat{\beta}^2_{X_j,\texttt{RB}}-\hat{\sigma}^2_{X_j,\texttt{RB}}\right)/\sigma^2_{Y_j}\right)^2}$$

Simulation results



Simulation results

Simulation Settings (when proportion of valid IV is low):

$\pi = 0.002,\ \varepsilon_x^2 = 5 \times 10^{-5}, a = 0.002$

$\gamma \sim \pi \cdot \text{Turncated Normal}\left(0,\ \varepsilon_x^2; (-\infty, -a], [a, +\infty)\right) + (1 - \pi) \cdot \delta_0$

	Cut-off	$\hat{\beta}$	Monte SD	SD	Coverage	Power	# SNPs
Two-sample	5.45	0.167	0.044	0.044	0.884	0.95	10
Three-sample	5.45	0.195	0.053	0.052	0.953	0.96	10
dIVW	0	0.222	0.217	0.213	0.971	0.12	ALL
RIVW	4.06	0.201	0.039	0.040	0.951	1.00	99

More Simulation results

Real data results: same-trait analysis

BMI-BMI analysis

- The causal effect size equals **1**
- Exposure GWAS: BMI GWAS data from the UK Biobank (N = 461,460, ID: ukb-b-19953)
- Outcome GWAS: BMI GWAS data from the GIANT consortium (N = 234,069, ID: ieu-a-2)[1]
- All data are downloaded from the IEU GWAS website[2]

	Threshold	Effect size	SE	95% CI	# IVs
RIVW	5×10^{-5}	**1.005**	0.022	[0.962, 1.048]	920
IVW	5×10^{-8}	0.833	0.014	[0.806, 0.860]	404
IVW	5×10^{-10}	0.857	0.016	[0.826, 0.888]	277
IVW	5×10^{-30}	1.014	0.034	[0.947, 1.081]	25

Details

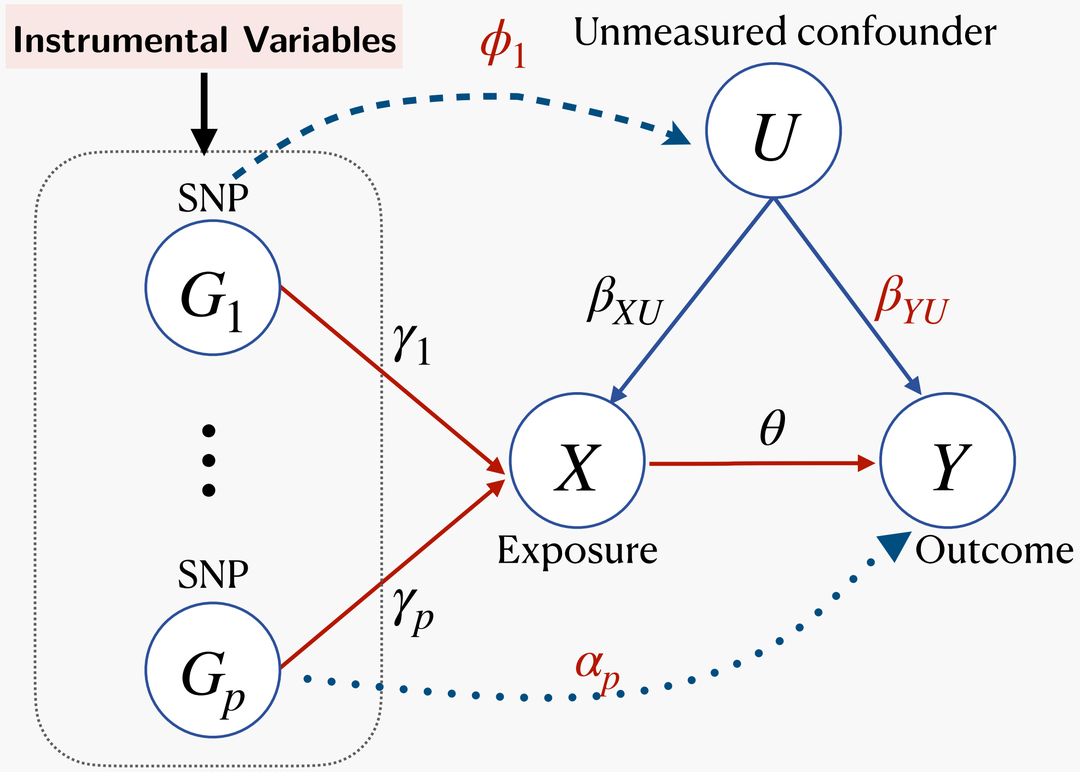
1 Locke, Adam E., et al. "Genetic studies of body mass index yield new insights for obesity biology." *Nature* 518.7538 (2015): 197-206.

2 Hemani, Gibran, et al. "The MR-Base platform supports systematic causal inference across the human phenome." elife 7 (2018): e34408.

Outline

- My research goal and motivation
- Breaking winner's curse in two-sample MR
- **Correcting pleiotropy in two-sample MR**
- Other works and future directions

Two Sample MR



$$r_j = \alpha_j + \phi_j \cdot \beta_{YU}$$

$$\beta_{Y_j} = \theta \cdot \beta_{X_j} + r_j$$

Structure equation model:

$$\beta_{X_j} = \gamma_j + \beta_{XU} \cdot \phi_j$$

$$\beta_{Y_j} = \beta_{Y_j,\mathrm{M}} + \beta_{Y_j,\mathrm{D}} = \theta \cdot \beta_{X_j} + (\alpha_j + \beta_{YU} \cdot \phi_j) \triangleq \theta \cdot \beta_{X_j} + r_j$$

For a valid IV j:

- $\phi_j = 0$ **and** $\alpha_j = 0 \longrightarrow r_j = 0$

For an invalid IV j:

- $\phi_j \neq 0$ **and (or)** $\alpha_j \neq 0 \longrightarrow r_j \neq 0$

Q: How to model r_j?

Causal Analysis with Rerandomization Estimator (CARE)

$$\underbrace{\hat{\beta}_{Y_j}}_{\text{response}} = \underbrace{\theta}_{\substack{\text{target}\\ \text{parameter}}} \cdot \underbrace{\beta_{X_j}}_{\substack{\text{true}\\ \text{covariate}}} + \underbrace{r_j}_{\substack{\text{unknown}\\ \text{parameter}}} + \underbrace{\nu_j}_{\text{noise}}, \qquad \underbrace{\hat{\beta}_{X_j,\mathrm{RB}} = \beta_{X_j} + u_j,}_{\substack{\text{covariates are}\\ \text{measured with error}}} \qquad j \in \mathcal{S}_\lambda$$

- In IVW, we assume all IVs are valid ($r_j = 0$) and ignore the measurement error (β_{X_j} is known)

$$\hat{\beta}_{Y_j} = \theta \cdot \beta_{X_j} + \nu_j, \qquad l(\theta) = \frac{1}{2} \sum_{j \in \mathcal{S}_\lambda} \frac{\left(\hat{\beta}_{Y_j} - \theta \cdot \beta_{X_j}\right)^2}{\sigma_{Y_j}^2}$$

Bias correction term for measurement error

- Bias-corrected least squares function:

$$l\left(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}\right) = \sum_{j \in \mathcal{S}_\lambda} l_j\left(\theta, \{r_j\}_{j \in \mathcal{S}_\lambda}\right) \triangleq \frac{1}{2} \sum_{j \in \mathcal{S}_\lambda} \frac{\left(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j,\mathrm{RB}} - r_j\right)^2}{\sigma_{Y_j}^2} - \frac{1}{2} \sum_{j \in \mathcal{S}_\lambda} \frac{\theta^2 \cdot \hat{\sigma}_{X_j,\mathrm{RB}}^2}{\sigma_{Y_j}^2}$$

Derivation details

Causal Analysis with Rerandomization Estimator (CARE)

- Bias-corrected least squares function:

$$l\left(\theta, \{r_j\}_{j\in\mathcal{S}_\lambda}\right) = \sum_{j\in\mathcal{S}_\lambda} l_j\left(\theta, \{r_j\}_{j\in\mathcal{S}_\lambda}\right) \triangleq \frac{1}{2}\sum_{j\in\mathcal{S}_\lambda} \frac{\left(\hat{\beta}_{Y_j} - \theta\cdot\hat{\beta}_{X_j,\mathrm{RB}} - r_j\right)^2}{\sigma^2_{\hat{Y}_j}} - \frac{1}{2}\sum_{j\in\mathcal{S}_\lambda} \frac{\theta^2\cdot\hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma^2_{\hat{Y}_j}}$$

- When all IVs are valid, the solution equals that of the RIVW estimator
- Invalid IVs may be selected due to widespread pleiotropic effects
- As invalid IVs provide biased estimates, we only use valid IVs to estimate θ:

$$\min_{\theta\in\mathbb{R}, r_j\in\mathbb{R}} \hat{l}\left(\theta, \{r_j\}_{j\in\mathcal{S}_\lambda}\right) \triangleq \frac{1}{2}\sum_{j\in\mathcal{S}_\lambda} \frac{\left(\hat{\beta}_{Y_j} - \theta\cdot\hat{\beta}_{X_j,\mathrm{RB}} - r_j\right)^2 - \theta^2\cdot\hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma^2_{\hat{Y}_j}} I(r_j = 0) \quad \text{subject to} \quad \sum_{j\in\mathcal{S}_\lambda} I(r_j \neq 0) = m$$

Causal Analysis with Rerandomization Estimator (CARE)

- The objective function is

$$\min_{\theta\in\mathbb{R},r_j\in\mathbb{R}} \hat{l}\left(\theta,\{r_j\}_{j\in\mathcal{S}_\lambda}\right) \triangleq \frac{1}{2}\sum_{j\in\mathcal{S}_\lambda} \frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\hat{\beta}_{X_j,\mathrm{RB}}-r_j\right)^2-\theta^2\cdot\hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma^2_{Y_j}} I(r_j=0) \quad \text{subject to} \quad \sum_{j\in\mathcal{S}_\lambda} I(r_j\neq 0)=m$$

- For a fixed m, obtain an estimated valid IV set $\hat{M}$ by a revised coordinate descent algorithm[1]
- Select the optimal m by the BIC: $-2\hat{l}\left(\theta,\{r_j\}_{j\in\mathcal{S}_\lambda}\right)+\log\left(\min(n_X,n_Y)\right)\cdot m$
- Obtain the estimator $\hat{\theta}$ and its standard deviation by the RIVW

Algorithm Details

Challenges: How to conduct inference **after model selection**?

- Perfect model selection is hard to achieve due to weak signals

[1] Xue, H., Shen, X., & Pan, W. (2021). Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal of Human Genetics*, *108*(7), 1251-1269.

Causal Analysis with Rerandomization Estimator (CARE)

- Bagging[1] (bootstrap smoothing) to reduce variability and eliminate discontinuities in model selection: **treat each IV as a subject in bagging**

$$\min_{\theta\in\mathbb{R},r_j\in\mathbb{R}} \hat{l}_b^*\left(\theta,\{r_j\}_{j\in\mathcal{S}_\lambda}\right) \triangleq \frac{1}{2}\sum_{j\in\mathcal{S}_\lambda} w_{bj}^* \frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\hat{\beta}_{X_j,\mathrm{RB}}-r_j\right)^2-\theta^2\cdot\hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma^2_{Y_j}} I(r_j=0) \quad \text{subject to} \quad \sum_{j\in\mathcal{S}_\lambda} I(r_j\neq 0)=m,$$

where $\{w_{bj}^*\}_{j\in\mathcal{S}_\lambda}$ follows a multinomial distribution with equal event probabilities

- Smoothing the estimator by averaging over the B (say, 2,000) bootstrap replications:

$$\widetilde{\theta}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b, \text{ where } \hat{\theta}_b \text{ is calculated by the previous procedure}$$

- The corresponding variance can be estimated by a conservative estimator:

$$\frac{\sum_{b=1}^{B}(\hat{\theta}_b-\tilde{\theta})^2}{B-1}$$

1 Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Two-sample MR with overlapped samples

- Previously, we assume no overlapped samples between exposure GWAS and outcome GWAS:

$$\begin{bmatrix} \hat{\beta}_{Y_j} \\ \hat{\beta}_{X_j} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} \beta_{Y_j} \\ \beta_{X_j} \end{bmatrix}, \begin{bmatrix} \sigma^2_{Y_j} & 0 \\ 0 & \sigma^2_{X_j} \end{bmatrix} \right), \quad j = 1, \ldots, p$$

- Samples may be overlapped due to the current trend of collecting biobank data:

$$\begin{bmatrix} \hat{\beta}_{Y_j} \\ \hat{\beta}_{X_j} \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} \beta_{Y_j} \\ \beta_{X_j} \end{bmatrix}, \begin{bmatrix} \sigma^2_{Y_j} & \rho\sigma_{X_j}\sigma_{Y_j} \\ \rho\sigma_{X_j}\sigma_{Y_j} & \sigma^2_{X_j} \end{bmatrix} \right), \quad j = 1, \ldots, p$$

- ρ can be estimated by LD score regression[1]

[1] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., ... & Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, *47*(11), 1236-1241.

CARE with overlapped samples

- Using the same idea, we only need to revise the objective function into:

$$\tilde{l}\left(\theta, \{r_j\}_{j\in\mathcal{S}_\lambda}\right) \triangleq \frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j} - \theta\cdot\hat{\beta}_{X_j,\mathrm{RB}} - r_j\right)^2}{\sigma^2_{Y_j}} - \underbrace{\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\theta^2\cdot\hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma^2_{Y_j}}}_{\substack{\text{bias correction}\\ \text{for the measurement error}}} + \underbrace{\theta\sum_{j\in\mathcal{S}_\lambda}\frac{\rho\cdot\hat{\sigma}_{X_j,\mathrm{RB}}}{\sigma_{Y_j}}}_{\substack{\text{bias correcction}\\ \text{for sample overlap}}}$$

- The other steps follow and remain the same

Computational time in CARE

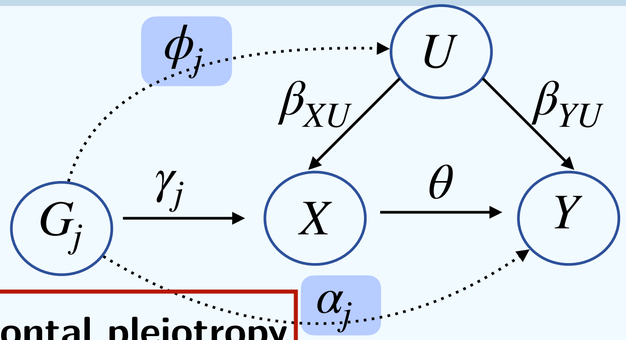
Computational time

- The algorithm is written in C++ through RcppArmadillo and is highly optimized
- For a simulation (over 12,000 replications) with an average of 328 IVs, the computational time for CARE (with 2,000 bootstraps) is 13.3 seconds by a single core in FSU Research Computing Center

Simulations

Simulation Design

- Dimension, sample size: $p = 200{,}000$, $n_X = n_Y = 500{,}000$
- Setups:
  - Proportion of IVs: $\pi_1 + \pi_2 = 0.02$; π_1—valid IVs; π_2—invalid IVs;
  - $\beta_{XU} = \beta_{YU} = 0.3$; δ_0: point mass at zero
  - $\sigma_x^2 = \sigma_y^2 = \sigma_u^2 = 1 \times 10^{-5}$

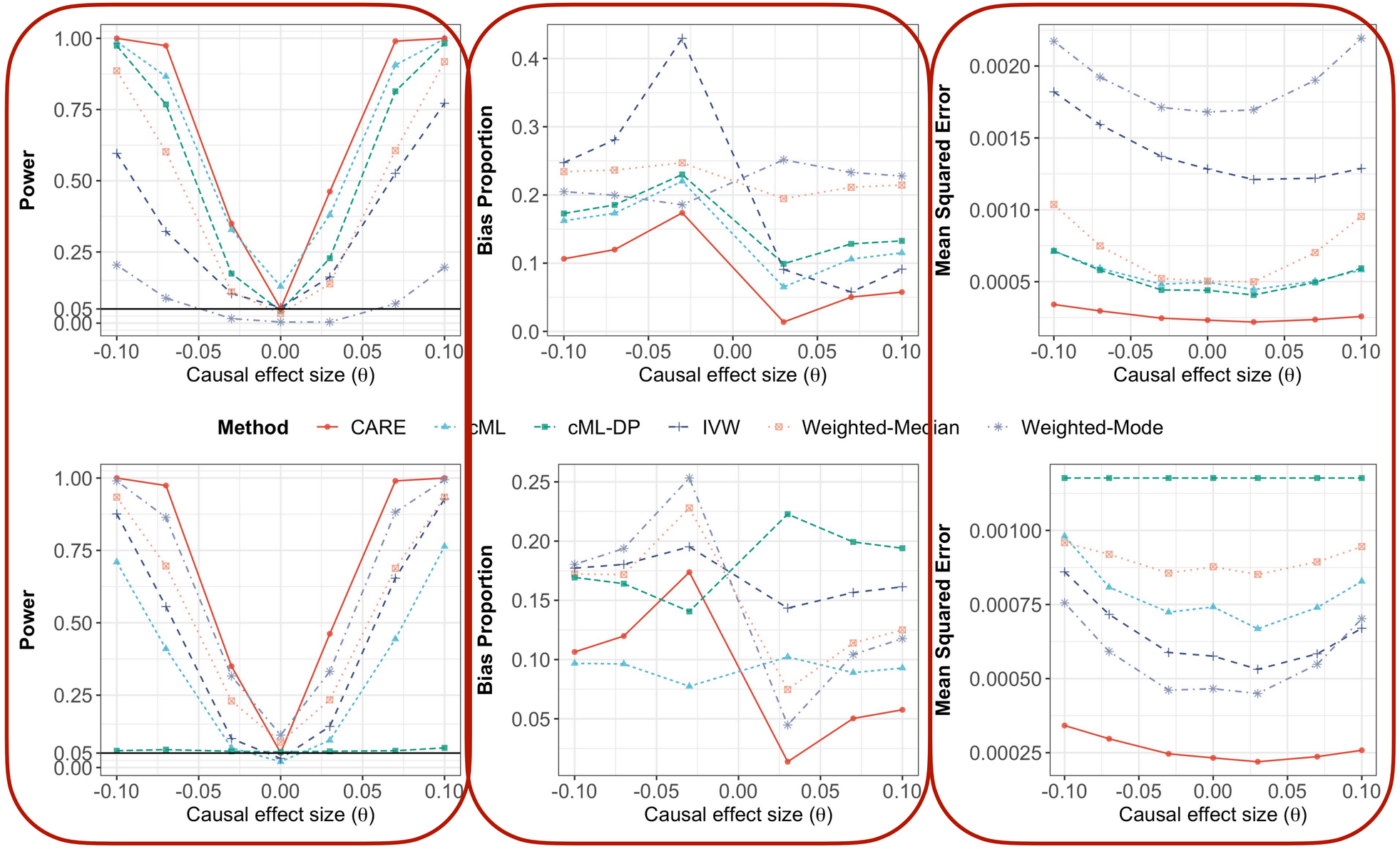


Valid IVs | **Directional pleiotropy** | **Balanced horizontal pleiotropy**

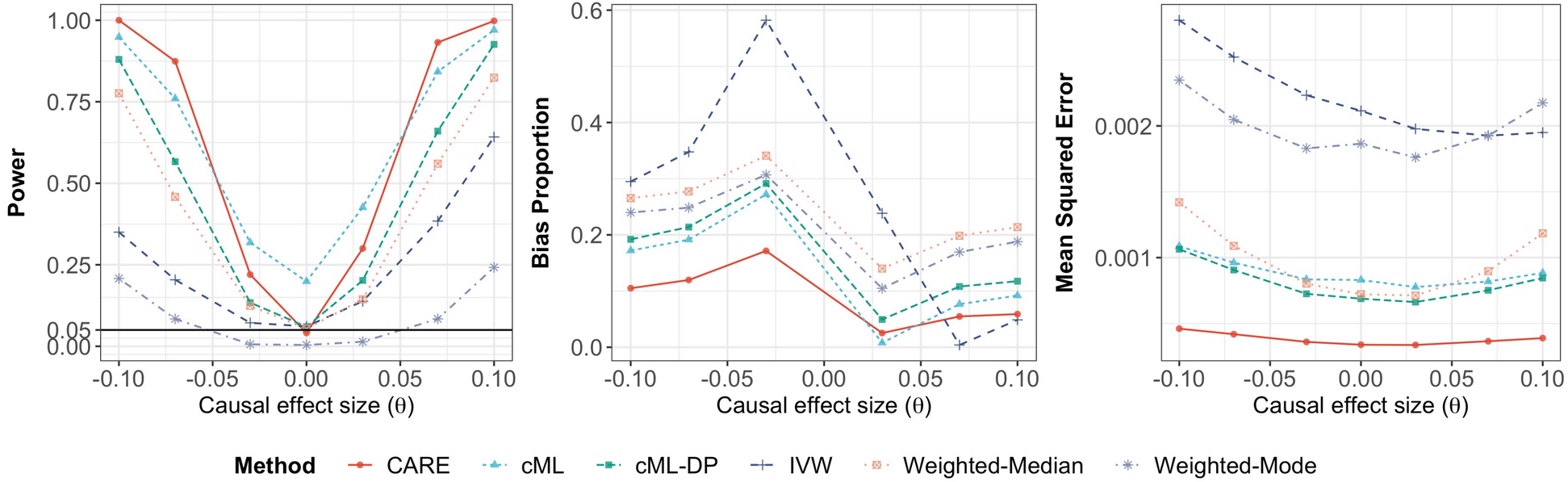
$$\begin{pmatrix}\gamma_j\\ \alpha_j\\ \phi_j\end{pmatrix} \sim \pi_1 \begin{pmatrix}N(0,\sigma_x^2)\\ \delta_0\\ \delta_0\end{pmatrix} + 0.3\pi_2 \begin{pmatrix}N(0,\sigma_x^2)\\ N(0,\sigma_y^2) \notin [-0.01,0]\\ N(0,\sigma_u^2)\end{pmatrix} + 0.7\pi_2 \begin{pmatrix}N(0,\sigma_x^2)\\ N(0,\sigma_y^2)\\ \delta_0\end{pmatrix} + \pi_3 \begin{pmatrix}\delta_0\\ N(0,\sigma_y^2)\\ \delta_0\end{pmatrix} + \pi_4 \begin{pmatrix}\delta_0\\ \delta_0\\ \delta_0\end{pmatrix}$$

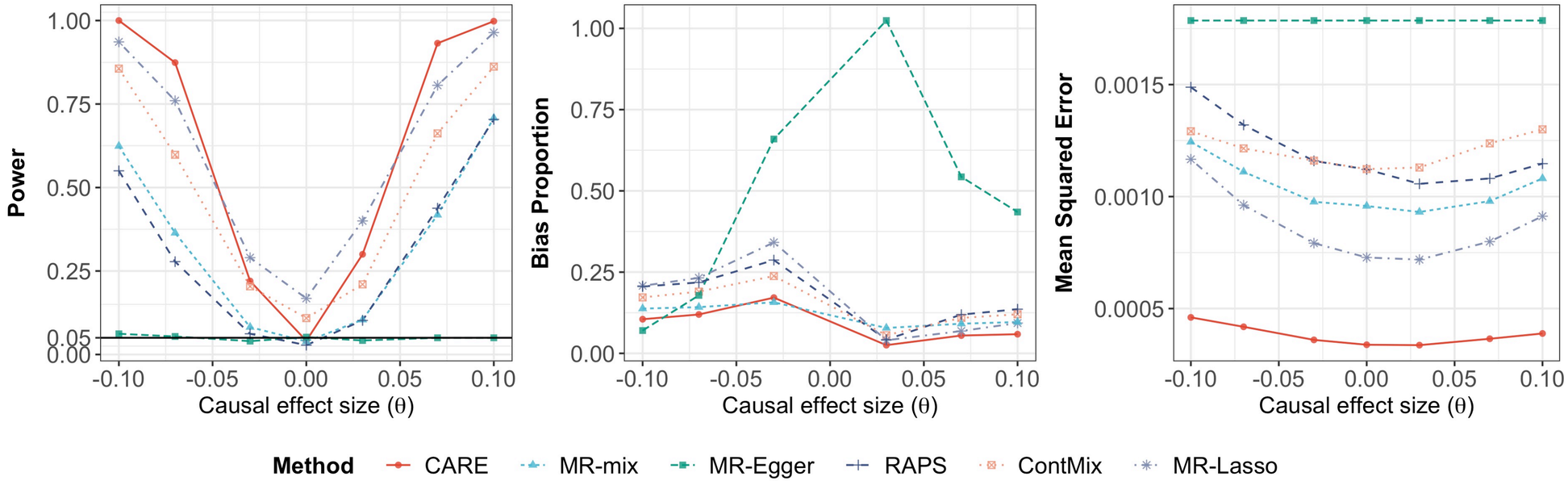
- Measurement errors: $\sigma_{X_j} = 1/\sqrt{n_X}$, $\sigma_{Y_j} = 1/\sqrt{n_Y}$
- Selection cutoff: $\alpha = 5 \times 10^{-8}$ for benchmark methods and $\alpha = 5 \times 10^{-5}$ for the proposed method CARE
- 1,000 replications for type 1 error rates, 500 replications for power

Simulation results: 30% invalid IVs

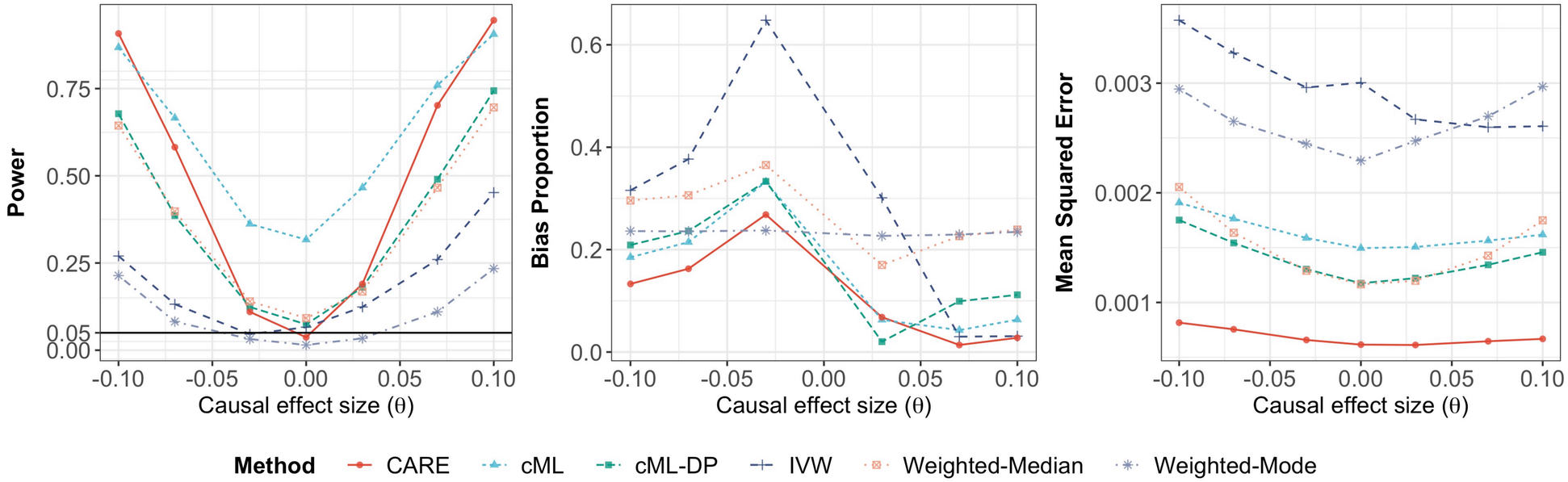


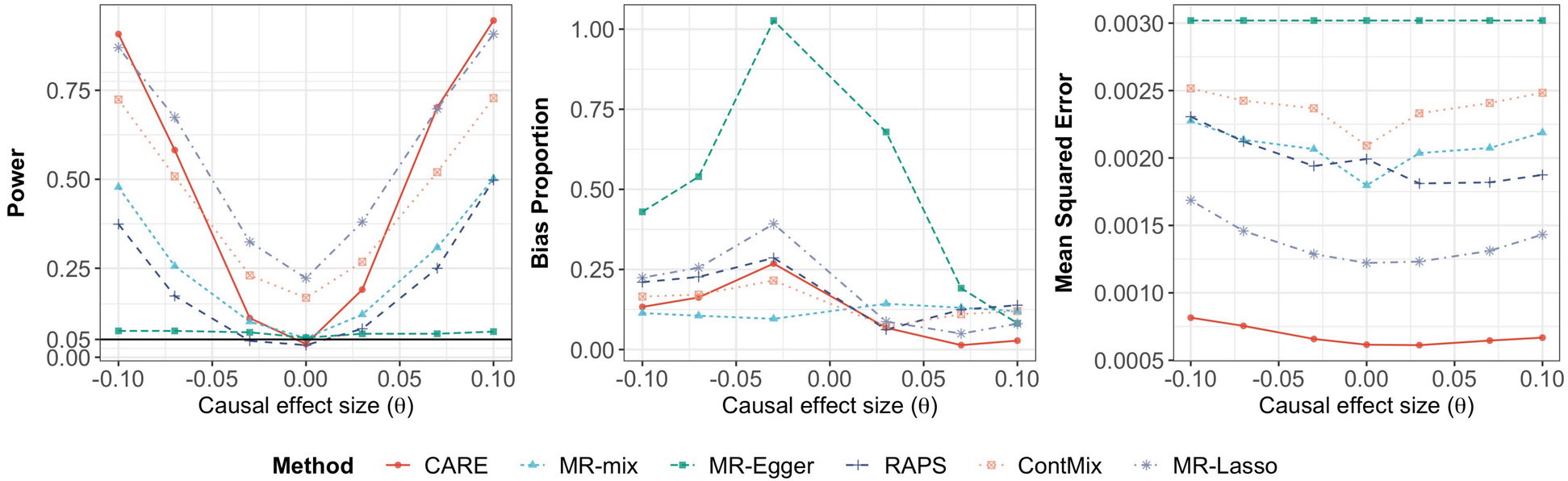
Simulation results: 50% invalid IVs





Simulation results: 70% invalid IVs





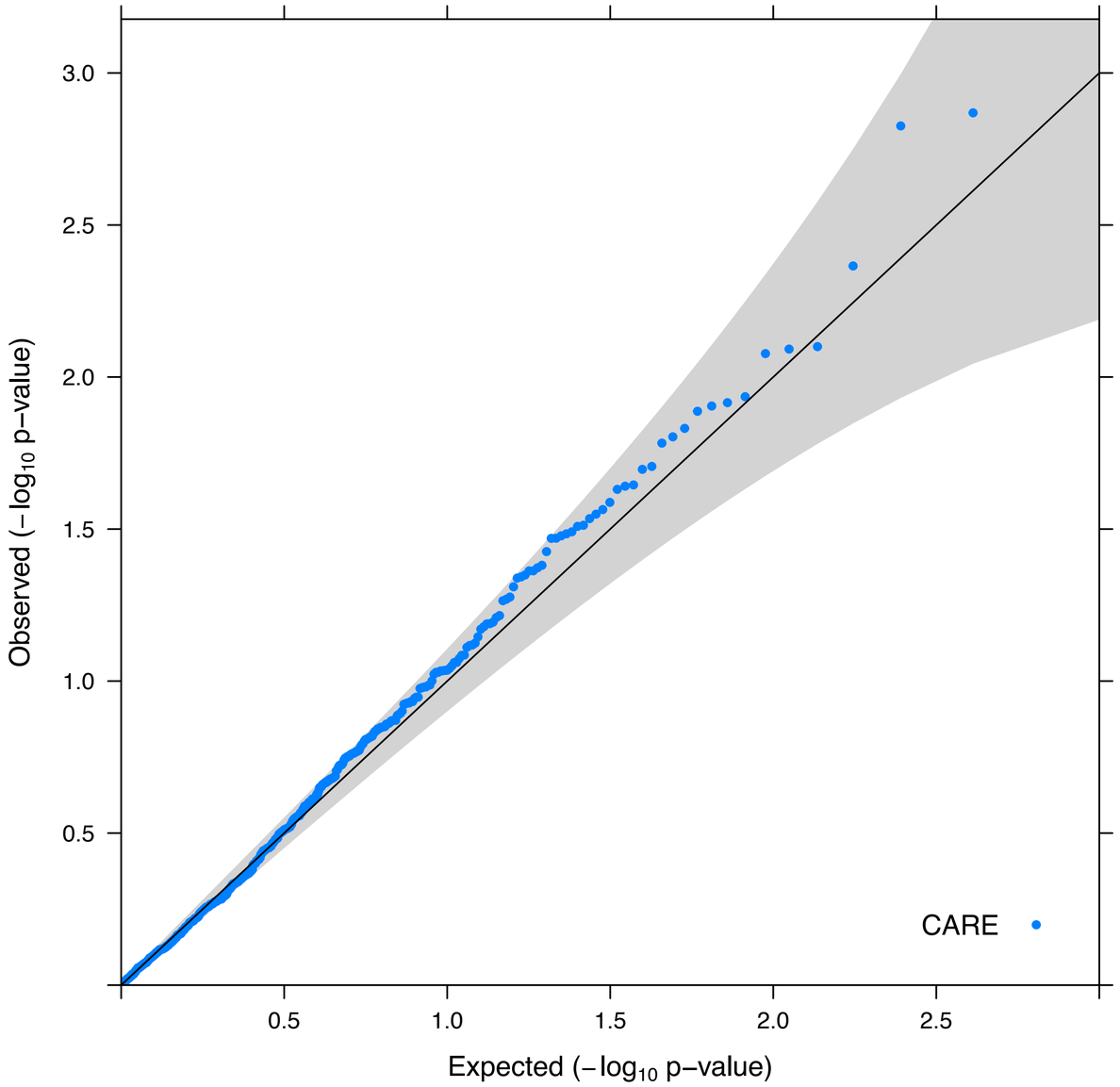
Real data analysis: Negative control outcome analysis

Evaluate Type 1 error rates with real datasets

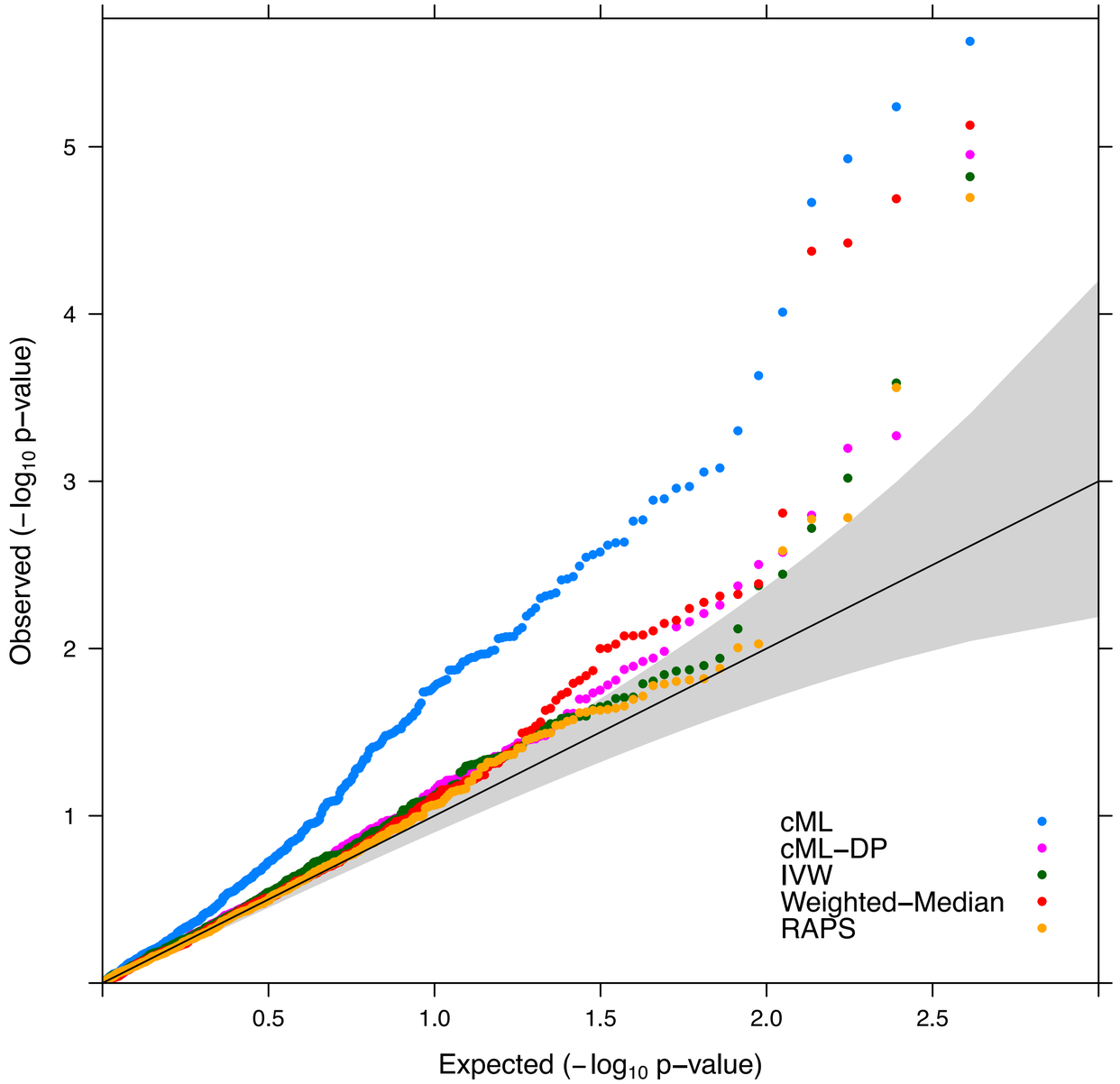
- Exposures: 124 risk factors (BMI, etc.) and diseases (Alzheimer's, etc.)
- **Negative control outcomes**[1]:
  - Natural hair color before greying (black, blonde, dark brown, light brown, and red) is largely **determined at birth** and is not expected to be associated with exposures
  - Based on the UK Biobank study
- We do not expect any causal effect of exposures on negative control outcomes ($\theta = 0$)
- All data were downloaded from the IEU OpenGWAS Project

1 Sanderson, E., Richardson, T. G., Hemani, G., & Smith, G. D. (2021). The use of negative control outcomes in Mendelian Randomisation to detect potential population stratification or selection bias. *International Journal of Epidemiology*, 50(4), 1350–1361

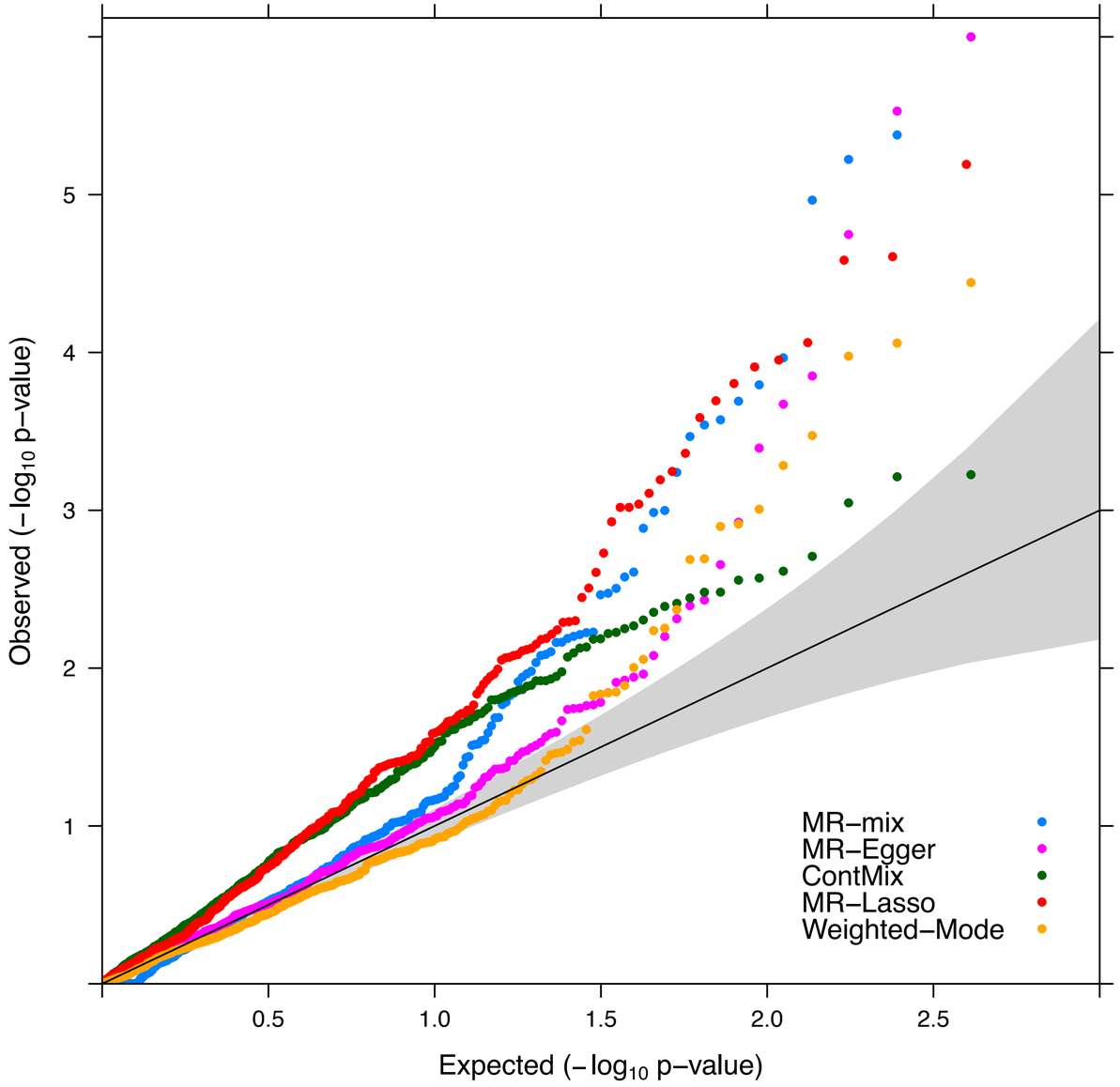
QQ plots for CARE



QQ plots for benchmark methods



QQ plots for benchmark methods



Real data analysis: Risk factors & COVID-19

Identify likely causal risk factors/diseases for COVID-19 severity

- Exposures: 124 risk factors/diseases (BMI, Childhood obesity, birth length, Total cholesterol, waist circumference, overweight, etc.)
- Outcome: COVID-19 severity (B2, Hospitalized COVID-19 vs population)
  - GWAS data from COVID19hg release 6, European ancestry[1]
  - 17,992 cases and 1,810,493 controls
- Use C2 (COVID-19 vs population) for partial validation
  - 87,870 cases and 2,210,804 controls

1 COVID-19 Host Genetics Initiative. (2021). Mapping the human genetic architecture of COVID-19. *Nature*.

Number of significant risk factors/diseases

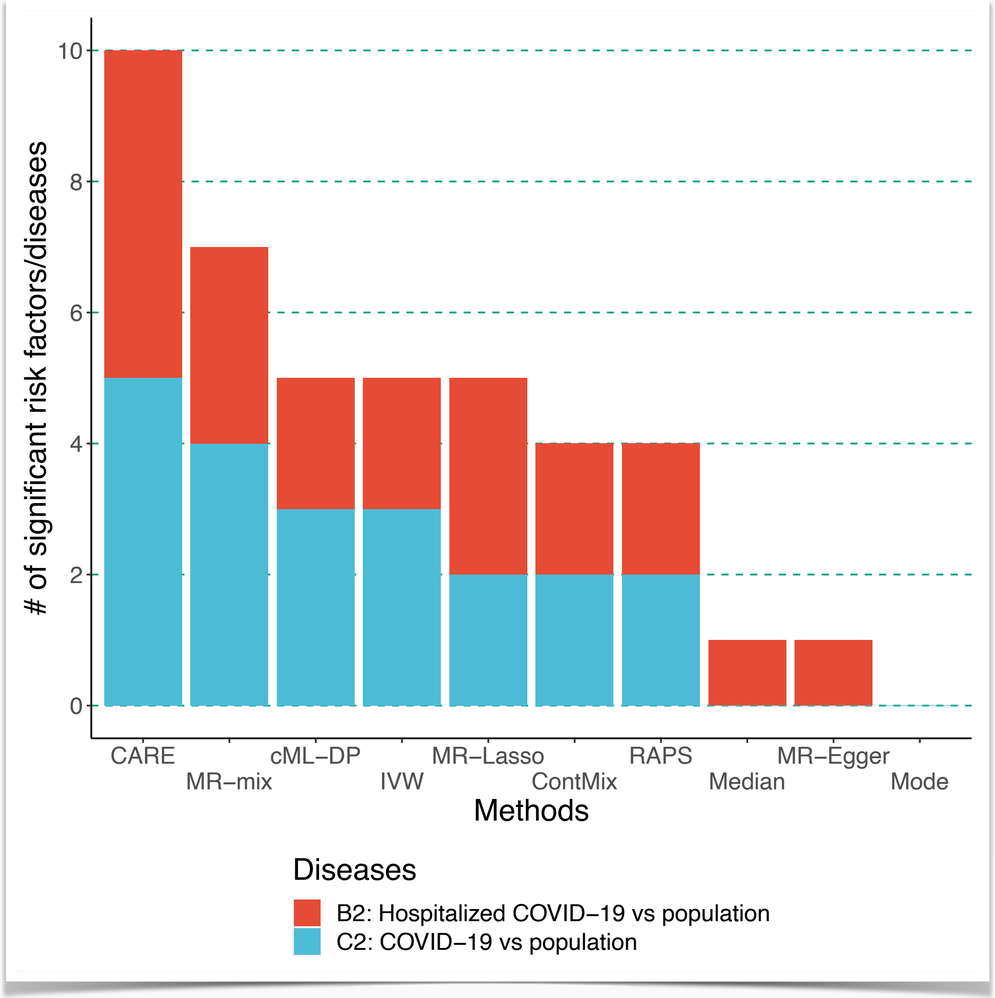


Figure: FDR p $<$ 0.05

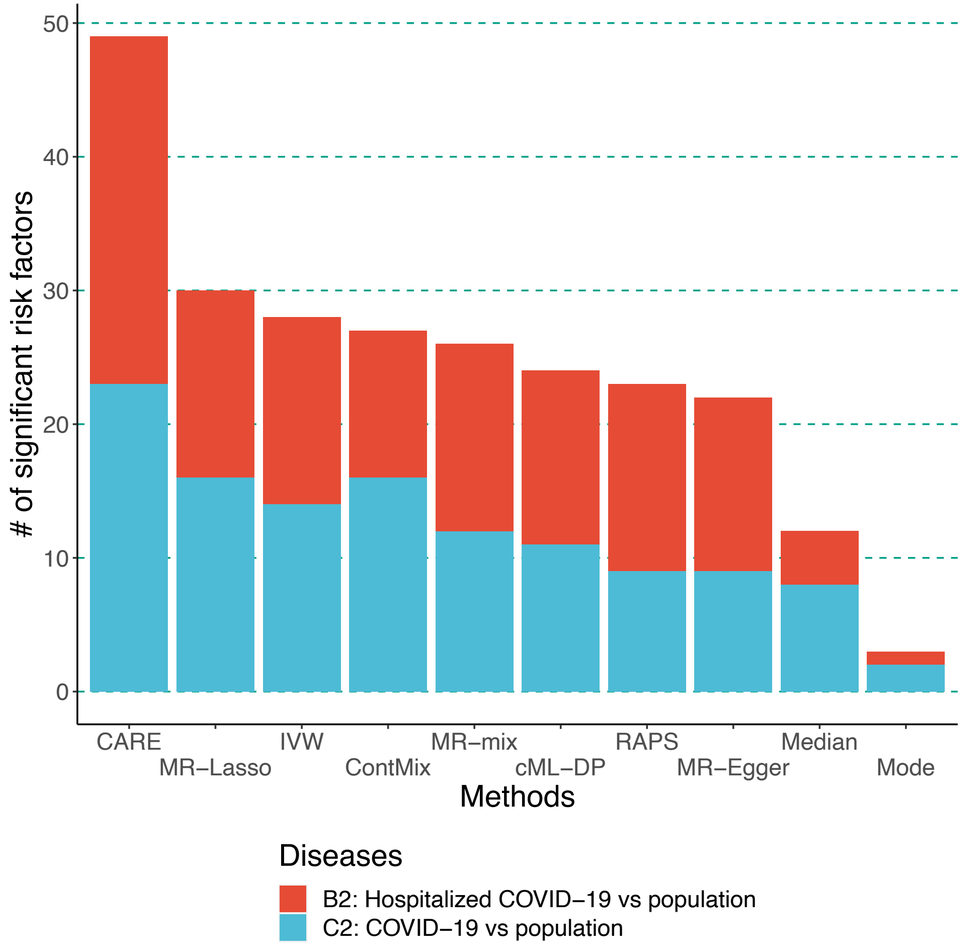
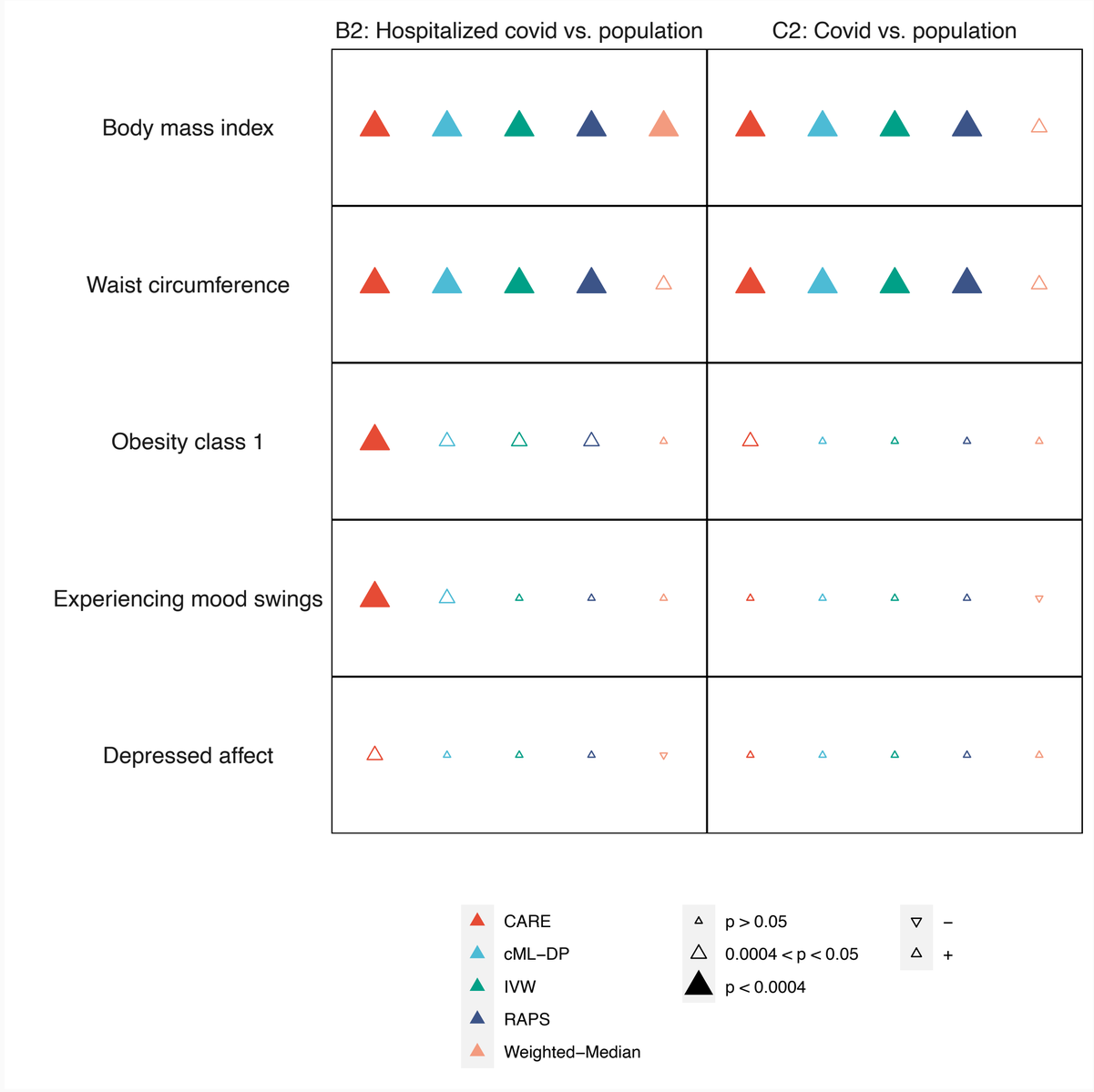


Figure: Suggestive threshold: p $<$ 0.05

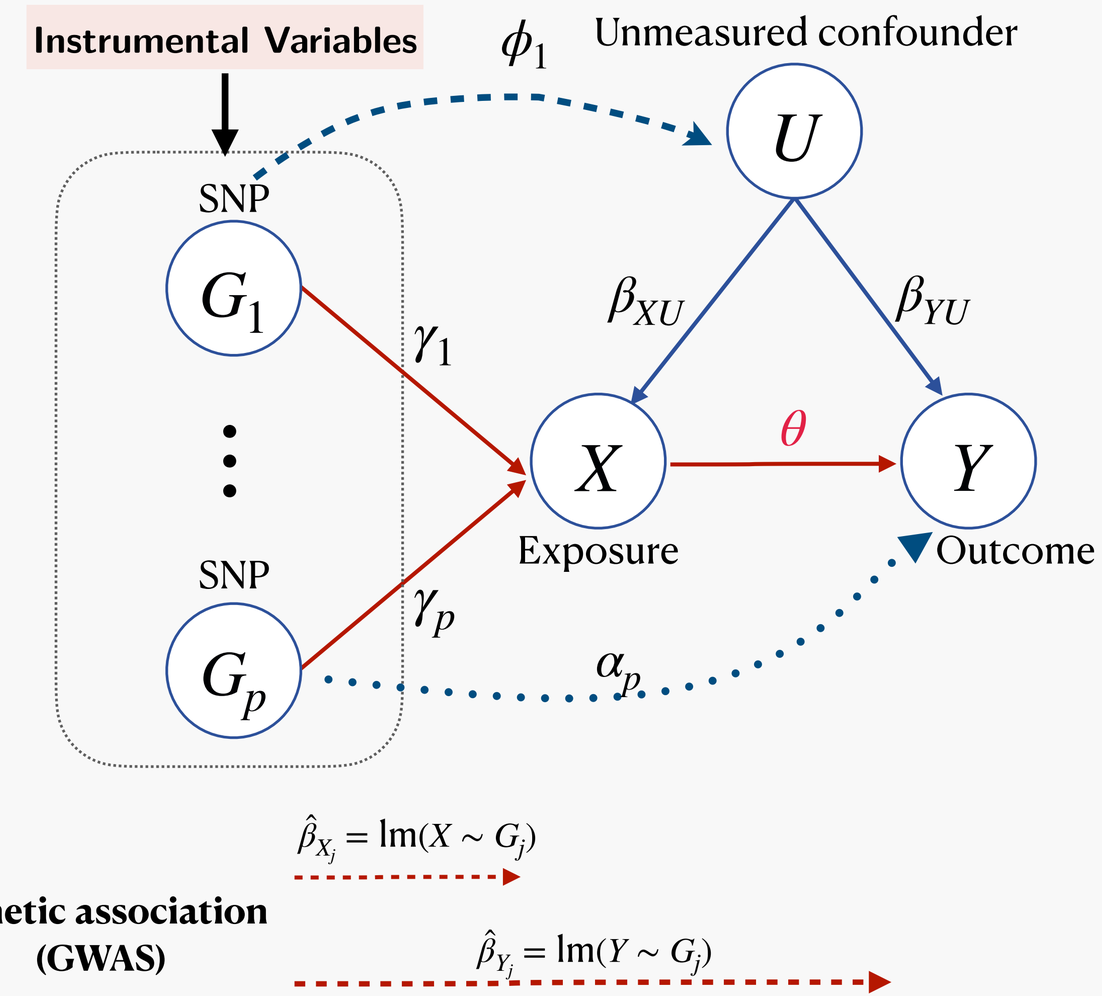
Significant risk factors (FDR $p < 0.05$)



Risk factors for COVID-19 severity (according to CDC):

- Body mass index
- Obesity class 1 (BMI of 30 to 35)
- Depression
- Mental disease

Summary



Assumptions: SNP j is a valid IV if

- Relevance: $\gamma_j \neq 0$
- **Independence:** $\phi_j = 0$
- **Exclusion restriction:** $\alpha_j = 0$

Contribution: A new rerandomization procedure to **break 'winner's curse' bias** in two sample MR

Contribution: A new method (CARE) that removes 'winner's curse' and measurement error bias and is **robust to pleiotropy** and **sample overlap**

Outline

- My research goal and motivation
- Breaking winner's curse in two-sample MR
- Correcting pleiotropy in two-sample MR
- **Other works and future directions**

Identify likely causal biomarkers

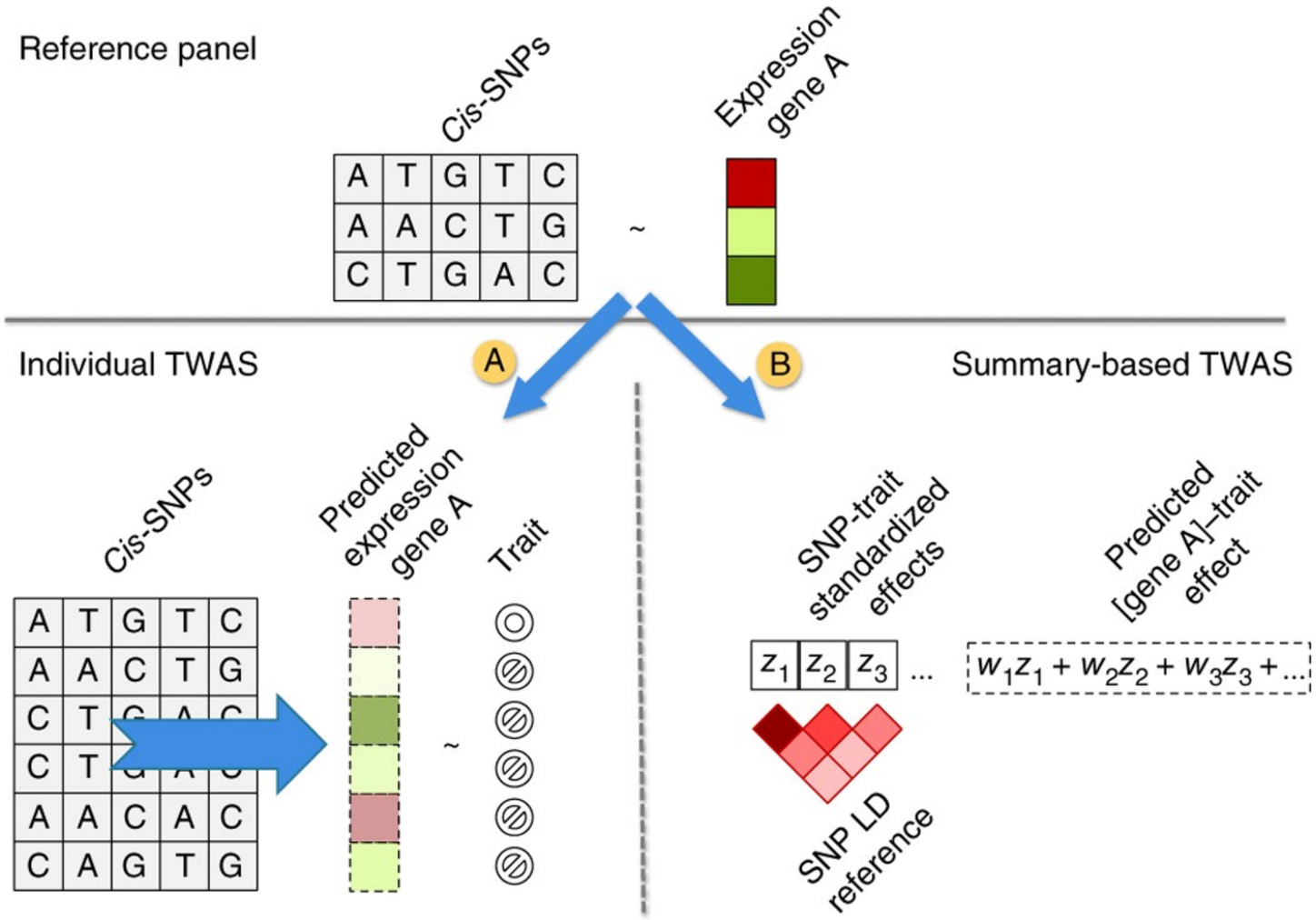


Figure: Workflow of TWAS[1]

1 Gusev, Alexander, et al. "Integrative approaches for large-scale transcriptome-wide association studies." *Nature Genetics* 48.3 (2016): 245-252.

Identify likely causal biomarkers

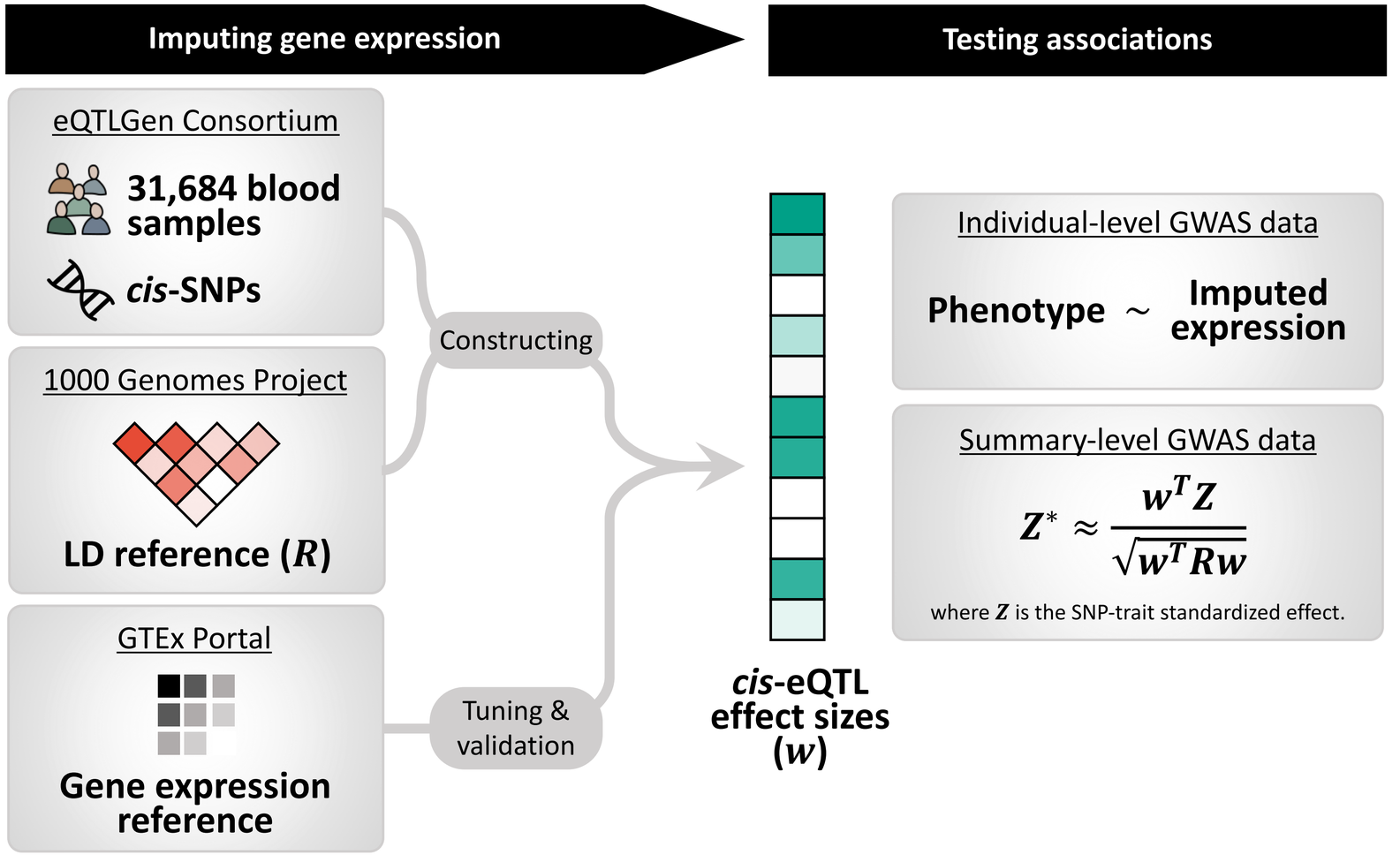
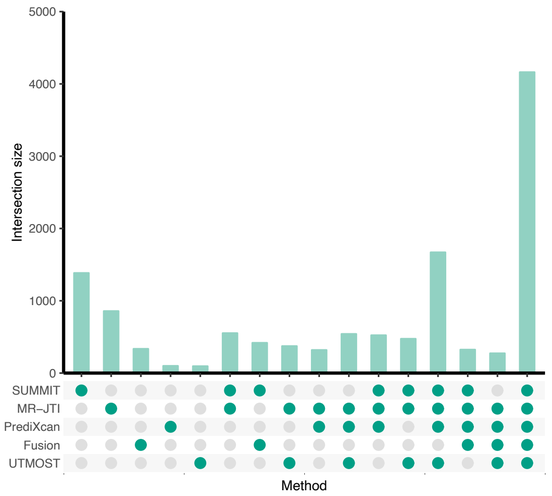


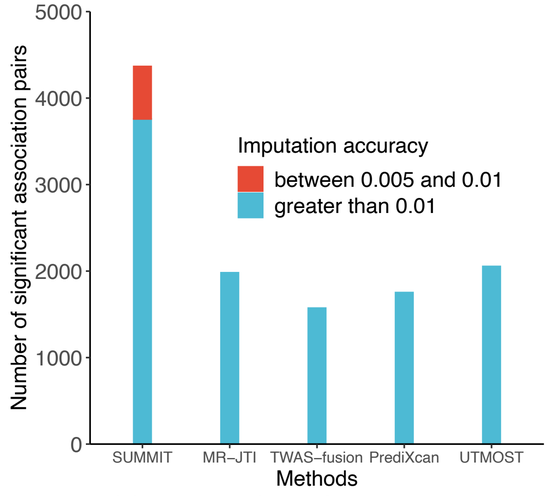
Figure: Workflow of SUMMIT

Zhang, Z.†, Bae, Y.†, Bradley, J., Wu, L, Wu, C.* (2021+). SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification. *Nature Communications*. Under review. Poster talk and Reviewers' Choice at ASHG 2021.

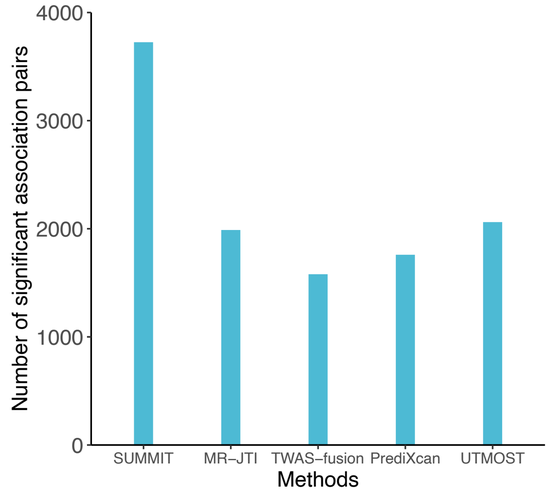
Identify likely causal biomarkers



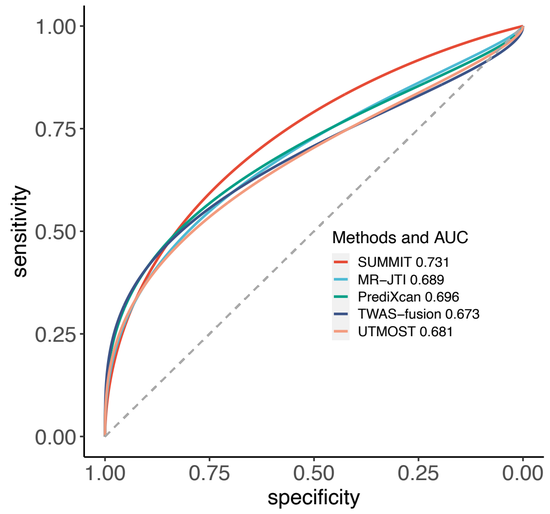
(a) UpSet plot of overlapping imputation models



(b) Number of significant association (evaluated based on all genes)



(c) Number of significant association (evaluated based on common genes)



(d) ROC plot

Zhang, Z.†, Bae, Y.†, Bradley, J., Wu, L, Wu, C.* (2021+). SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification. *Nature Communications*. Under review. Poster talk and Reviewers' Choice at ASHG 2021.

Identify likely causal biomarkers

Ideas:

- TWAS-type methods can be viewed as one type of MR with correlated instrumental variables. Make methods more robust to the violation of IV assumptions
- Consider both cis- and trans-acting elements
- Consider other types of biomarkers (such as proteins) and other ancestries

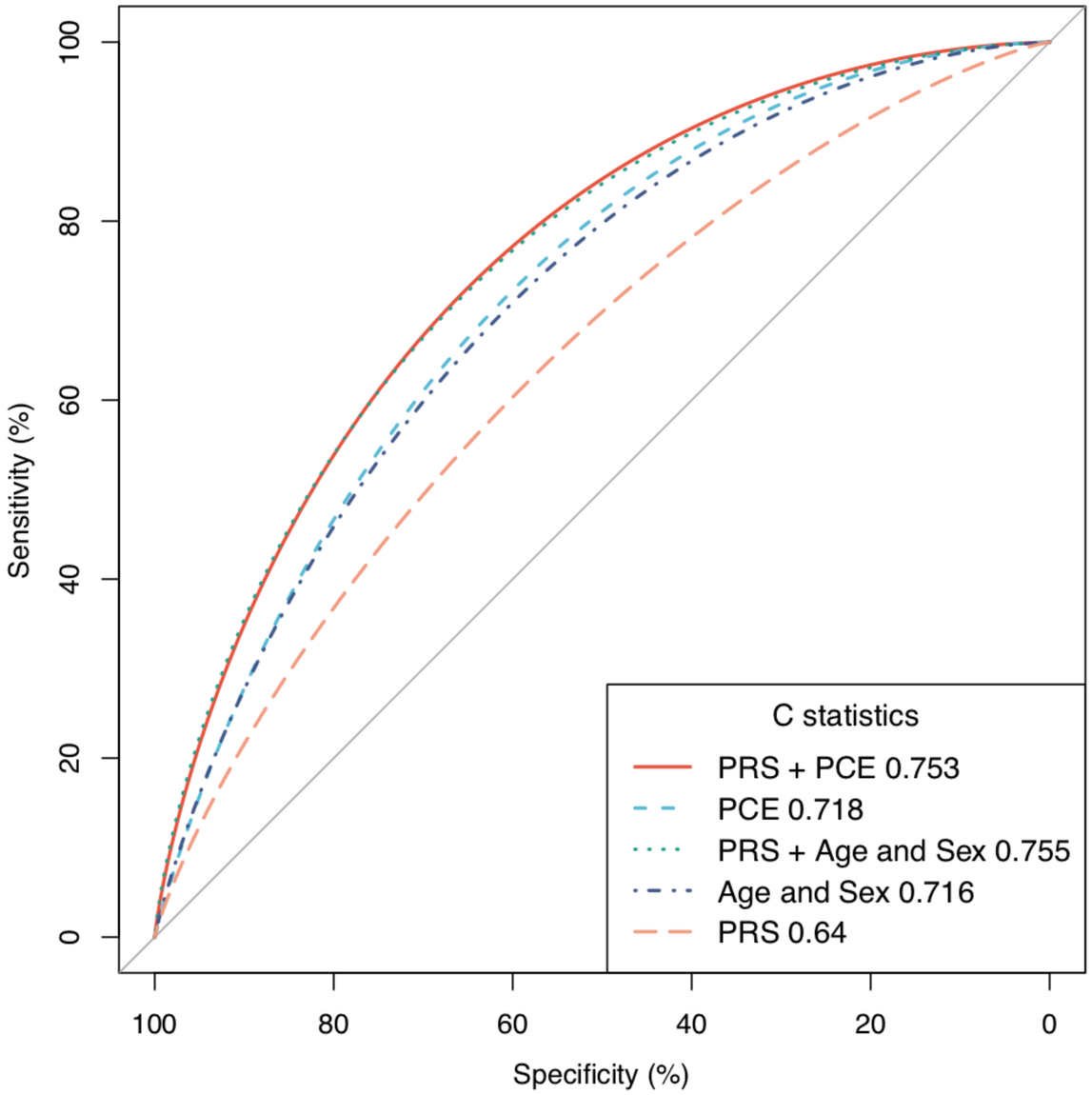
R01: Uncovering causal protein markers to improve prostate cancer etiology understanding and risk prediction in Africans and Europeans (PI: Lang Wu and **Chong Wu**); Put forward for exception funding by PO

Enhance risk prediction

- **Polygenic risk score (PRS):** a risk prediction method by integrating genetic information from all genetic variants
- There are many debates regarding whether PRS is really useful in a clinical setting
- We evaluate this for coronary artery disease[1]
  - Ensemble PRS: combine multiple GWAS datasets and several PRS methods
  - Evaluate if adding ensemble PRS to PCE can improve the risk prediction using the independent White British subjects in UK Biobank
  - Pooled cohort equation (PCE) is a guideline recommended clinical risk score for coronary artery disease

1 King, A.† , Wu, L., Deng, HW., & Wu, C.* (2021+). Polygenic risk score improves the accuracy of a clinical risk score for coronary artery disease. Submitted.

Enhance risk prediction



For incident CAD cases, **14.2%** of individuals correctly reclassified to the higher-risk category and 2.6% incorrectly reclassified to the lower-risk category

Enhance risk prediction

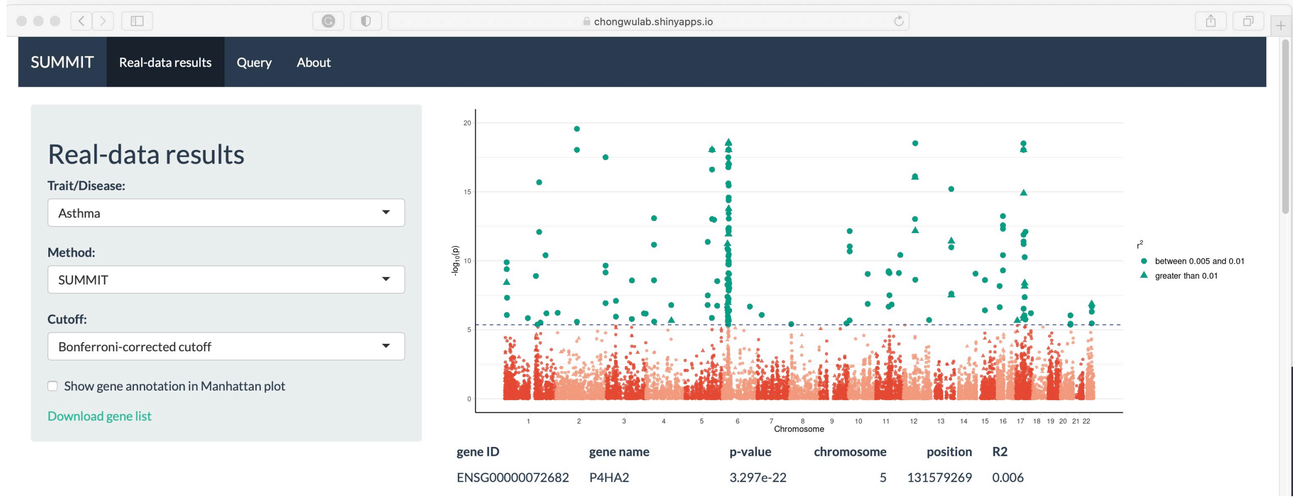
Cross-ancestry PRS:

- Poor performance mainly because minor allele frequency and linkage disequilibrium are different across ancestry
- Causal variants are largely shared across ancestry
- We hypothesize that causal biomarkers are also largely shared across ancestry, and incorporating likely causal biomarkers may be helpful

R01: Uncovering causal protein markers to improve prostate cancer etiology understanding and risk prediction in Africans and Europeans (PI: Lang Wu and **Chong Wu**); Put forward for potential funding by PO

Software and pipeline development

- Online servers (R Shiny) for searching our real data results
- R packages that are available in both GitHub and CRAN
- Tutorials for distributing our software and pipeline



glmtlp: Generalized Linear Models with Truncated Lasso Penalty

Extremely efficient procedures for fitting regularization path with l0, l1, and truncated lasso penalty for linear regression and logistic regression models. This version is a completely new version compared with our previous version, which was mainly based on R. New core algorithms are developed and are now written in C++ and highly optimized.

Version:	2.0.1
Depends:	R (≥ 3.5.0)
Imports:	foreach, doParallel, ggplot2
Suggests:	rmarkdown, knitr, testthat (≥ 3.0.0)
Published:	2021-12-17
Author:	Chunlin Li [aut], Yu Yang [aut, cre], Chong Wu [aut]
Maintainer:	Yu Yang <yang6367 at umn.edu>
License:	GPL-3
URL:	https://yuyangyy.com/glmtlp/
NeedsCompilation:	yes
Materials:	README NEWS
CRAN checks:	glmtlp results

Documentation:

Reference manual:	glmtlp.pdf
Vignettes:	glmtlp

Downloads:

Chong Wu

Home
Research
Grants
Teaching
Students
Software
Presentations

Softwares

I have developed and currently maintain the following software. You can get the latest version from my GitHub repository.

- prclust: R package that provides two algorithms for fitting the penalized regression-based clustering (PRclust). The corresponding paper is Wu, Kwon, Shen and Pan, 2016.
- MiSPU: R package that presents a novel global testing method called aMiSPU, that is highly adaptive and thus high powered across various scenarios, alleviating the issue with the choice of a phylogenetic distance. The corresponding paper is Wu, Chen, Kim and Pan, 2016.
- GLMaSPU: R package that makes it incredibly easy to implement some testing methods under high-dimensional generalized linear models. The corresponding paper is our 2019 Stat Sinica paper.
- glmtlp: R package that makes it easy to implement the truncated lasso penalty under a generalized linear model framework. This package is similar to glmnet but can be applied with a non-convex penalty.

To help researchers from other field use our newly developed method, we have created and maintained the following software and pipeline. We assume no prior knowledge in R and all the following software can be run easily and smoothly once the required packages are installed. Please send me an email （cwu3@fsu.edu) if you have find any bugs when using them.

- IWAS: A software for implementing Imaging-Wide Association Studies (IWAS). The corresponding paper is our 2017 NeuroImage paper.
- TWAS-aSPU: A more powerful gene-based association test to integrate single set or multiple sets of eQTL data with GWAS individual-level data or summary statistics. The corresponding paper is our 2017 Genetics paper.
- aSPUpath2: A new pathway-based method for integrating eQTL data with GWAS summary statistics. This can be viewed as an extension of TWAS to the pathway-based analysis. The corresponding paper is our 2018 Genetic Epidemiology paper.
- egmethyl: A new gene-based test for integrating enhancer-promoter interactions and DNA methylation data with GWAS summary data. The corresponding paper is our 2019 Bioinformatics paper.

We create a GitHub lab page for the software written by the group members

- CMO: Cross Methylome Omnibus (CMO) integrates genetically regulated DNAm in enhancers, promoters, and the gene body to identify additional disease-associated genes.
- FOGS: FOGS is a powerful fine-mapping method that prioritizes putative causal genes by accounting for local LD in TWAS results

Future directions summary

Develop new methods/theory/software to

- Identify likely causal risk factors and biomarkers
- Enhance risk prediction

Extend to other types of big and messy data

- Deep learning

ARTICLE

https://doi.org/10.1038/s41467-021-26643-8 OPEN

Check for updates

Accurate recognition of colorectal cancer with semi-supervised deep learning on pathological images

Gang Yu[1], Kai Sun[1], Chao Xu[2], Xing-Hua Shi[3], Chong Wu[4], Ting Xie[1], Run-Qi Meng[5], Xiang-He Meng[6], Kuan-Song Wang[7✉], Hong-Mei Xiao[6✉] & Hong-Wen Deng[6,8✉]

Acknowledgements

The first part of the talk:

Ma, X., Wang, J.*, & Wu, C.* (2021+). Breaking the winner's curse in Mendelian randomization: Rerandomized inverse variance weighted estimator.

Alphabetical order

The second part of the talk:

Wu, C.* & Wang, J.* (2021+) Accounting for winner's curse and pleiotropy in two-sample Mendelian randomization.

*: corresponding author

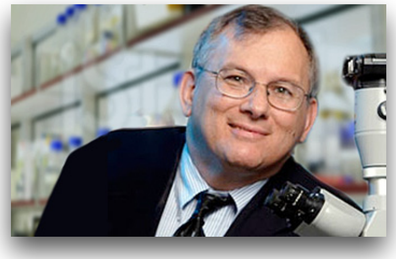
Jingshen Wang
@ UC Berkeley

Xinwei Ma
@ UCSD

Acknowledgements






Collaborators

- Jingshen Wang @ UC Berkeley Biostatistics
- Xinwei Ma @ UCSD Economics
- Lang Wu @ University of Hawaii Cancer Center
- Hong-Wen Deng @ Tulane Biomedical Informatics
- Jon Bradley @ Florida State Statistics
- Richard Nowakowski @ Florida State Biomedical Sciences
- Aliza Wingo @ Emory Psychiatry and Behavioral Sciences
- Thomas Wingo @ Emory Neurology and Human Genetics
- Keenan Walker @ NIA Behavioral Neuroscience

Students

- Zichen Zhang, Austin King, Ye Eun Bae, etc

Thank you!

Chong Wu

Email: cwu3@fsu.edu

Website: https://wuchong.org

Details on $\sigma^2_{X_j,\texttt{RB}}$

$$\sigma^2_{X_j,\texttt{RB}} = \sigma^2_{X_j} - \frac{\sigma^2_{X_j}}{\eta_j^2} E\left[\frac{(\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j})\phi(\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j}) - (-\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j})\phi(-\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j})}{1 - \Phi(\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j}) + \Phi(-\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j})} - \left(\frac{\phi(\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j}) - \phi(-\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j})}{1 - \Phi(\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j}) + \Phi(-\frac{\lambda}{\eta_j} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta_j})} \right)^2 \middle| S_j > 0 \right]$$

$$\hat\sigma^2_{X_j,\texttt{RB}} = \sigma^2_{X_j}\left(1 - \frac{1}{\eta^2} \frac{(\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta})\phi(\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta}) - (-\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta})\phi(-\frac{\lambda}{\eta} - \frac{\hat\gamma_j}{\sigma_{X_j}\eta})}{1 - \Phi(\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta}) + \Phi(-\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta})} + \frac{1}{\eta^2}\left(\frac{\phi(\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta}) - \phi(-\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta})}{1 - \Phi(\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta}) + \Phi(-\frac{\lambda}{\eta} - \frac{\hat\beta_{X_j}}{\sigma_{X_j}\eta})} \right)^2 \right)$$

Back to main slides

Details on assumption

Assumption 1 (Measurement error model)

(i) For any $j \neq j'$, the pairs, $(\hat{\beta}_{Y_j}, \hat{\beta}_{X_j})$ and $(\hat{\beta}_{Y_{j'}}, \hat{\beta}_{X_{j'}})$ are mutually independent

(ii) For each j,

$$\begin{bmatrix} \hat{\beta}_{Y_j} \\ \hat{\beta}_{X_j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \theta\beta_{X_j} \\ \beta_{X_j} \end{bmatrix}, \begin{bmatrix} \sigma_{Y_j}^2 & 0 \\ 0 & \sigma_{X_j}^2 \end{bmatrix} \right)$$

For some $\nu \to 0$, $\{\sigma_{Y_j}/\nu, \sigma_{X_j}/\nu : 1 \leq j \leq p\}$ are uniformly bounded and bounded away from zero

Assumption 2 (Instrument selection): The cutoff value satisfies $\lambda \to \infty$

Assumption 3 (No dominant instrument): The true instrument effect satisfies

$$\max_{j \in S_\lambda} \gamma_j^2 / \left(\sum_{j \in S_\lambda} \gamma_j^2 \right) \xrightarrow{\text{p}} 0$$

When Assumptions 1–3 hold, $p_\lambda = |S_\lambda| \to \infty$ and $\kappa_\lambda/\lambda^2 \to \infty$, where $\kappa_\lambda = \frac{1}{p_\lambda} \sum_{j \in \mathcal{S}_\lambda} (\gamma_j/\sigma_{X_j})^2$, the Theorem holds

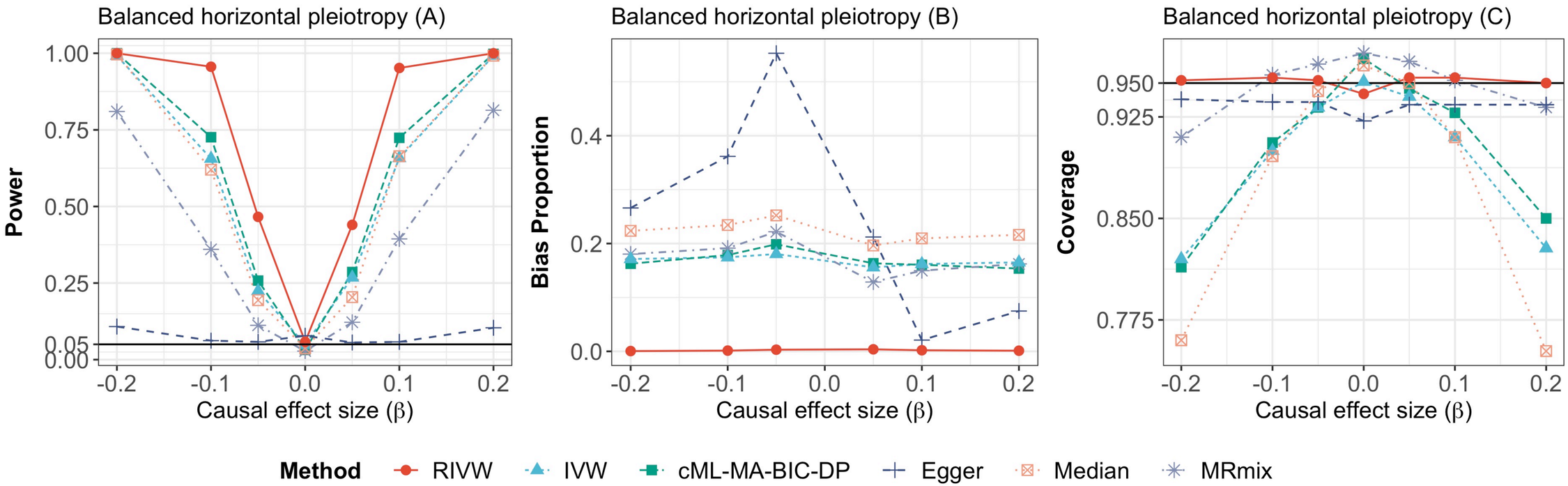
Back to main slides

RIVW Simulation results

Setting (balanced horizontal pleiotropy):

$$\pi_1 = 0.01,\ \varepsilon_x^2 = 5 \times 10^{-5}, \rho = 0.3$$

$$\begin{pmatrix}\gamma_j \\ \alpha_j\end{pmatrix} \sim \pi_1\rho \begin{pmatrix}N(0,\varepsilon_x^2) \\ \delta_0\end{pmatrix} + \pi_1(1-\rho)\begin{pmatrix}N(0,\varepsilon_x^2) \\ N(0,\varepsilon_x^2)\end{pmatrix} + \pi_2 \begin{pmatrix}\delta_0 \\ N(0,\varepsilon_x^2)\end{pmatrix} + \pi_3 \begin{pmatrix}\delta_0 \\ \delta_0\end{pmatrix}$$



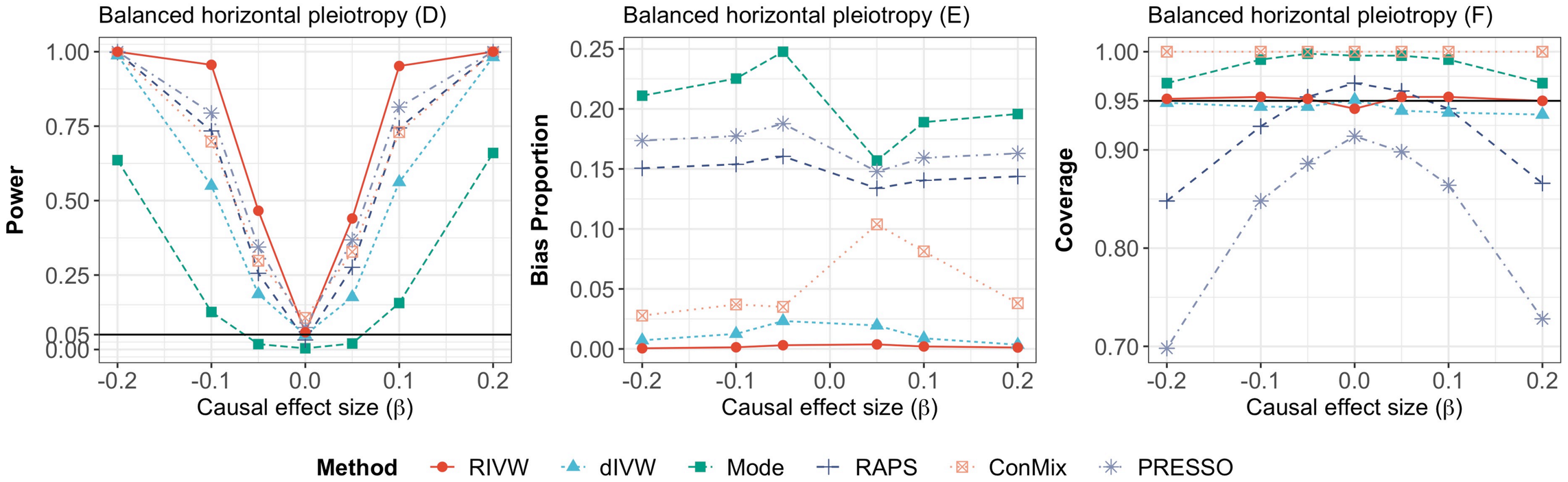
Back to main slides

RIVW Simulation results

Setting (balanced horizontal pleiotropy):

$$\pi_1 = 0.01,\ \varepsilon_x^2 = 5 \times 10^{-5}, \rho = 0.3$$

$$\begin{pmatrix}\gamma_j \\ \alpha_j\end{pmatrix} \sim \pi_1\rho \begin{pmatrix}N(0,\varepsilon_x^2) \\ \delta_0\end{pmatrix} + \pi_1(1-\rho)\begin{pmatrix}N(0,\varepsilon_x^2) \\ N(0,\varepsilon_x^2)\end{pmatrix} + \pi_2 \begin{pmatrix}\delta_0 \\ N(0,\varepsilon_x^2)\end{pmatrix} + \pi_3 \begin{pmatrix}\delta_0 \\ \delta_0\end{pmatrix}$$



Back to main slides

Details on independent IV selections

- In MR studies, we often require independent IVs
- To achieve this, one often applies the clumping (select ones with the smallest p values in a region) to select independent IVs: really hard to deal with the $\max_{j \in \mathcal{S}} \hat{\beta}_{X_j}$
- In our RIVW, we propose a modified clumping (select ones with the smallest estimated variance in a region) to select independent IVs
  - Benefits: our method and theory can go through with this procedure
  - Disadvantages: lose power compared to the original clumping procedure

Back to main slides

The benefits of rerandomization

Back to main slides

Conditional on the selection, $\hat{\beta}_{X_j}$ follows a truncated normal distribution and we can reduce bias by[1]

$$\hat{\beta}_{X_j} = E\left(\hat{\beta}_{X_j} \middle| |\hat{\beta}_{X_j}/\sigma_{X_j}| > \lambda\right)$$

Simulations:

- Dimension, sample size: $p = 200{,}000$, $n_X = 100{,}000$
- Threshold: $\lambda = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = 4.06$, $\alpha = 5 \times 10^{-5}$
- $\beta_{X_j} \sim \pi \cdot N\left(0.01, \varepsilon_x^2\right) + (1-\pi) \cdot \delta_0, \qquad \varepsilon_x^2 = 1 \times 10^{-7}, \pi = 0.02$
- Measurement errors: $\sigma_{X_j} = 1/\sqrt{n_X}$

Simulation results: Standardized Bias: $(\tilde{\beta}_{X_j} - \beta_{X_j})/\sigma_{X_j}$

Two-sample: 1.26; Three-sample: -0.020; Bias-reduced: -0.658; **Rerandomization: -0.027**

1 Zhong, H., & Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, *9*(4), 621-634.

Details on bias-corrected least squares function

When we have true β_{X_j}, the least squares function is: $\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\beta_{X_j}-r_j\right)^2}{\sigma^2_{\hat{Y}_j}}$

$$E\left[\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\hat{\beta}_{X_j,\text{RB}}-r_j\right)^2}{\sigma^2_{\hat{Y}_j}}\right]=E\left[\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j}-\theta\cdot(\beta_{X_j}-u_j)-r_j\right)^2}{\sigma^2_{\hat{Y}_j}}\right]$$

$$=E\left[\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\beta_{X_j}-r_j\right)^2}{\sigma^2_{\hat{Y}_j}}\right]+E\left[\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\theta\cdot u_j\right)^2}{\sigma^2_{\hat{Y}_j}}\right]$$

$$=E\left[\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\beta_{X_j}-r_j\right)^2}{\sigma^2_{\hat{Y}_j}}\right]+\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\theta^2\sigma^2_{X_j,RB}}{\sigma^2_{\hat{Y}_j}}$$

So $\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\beta_{X_j}-r_j\right)^2}{\sigma^2_{\hat{Y}_j}}\sim\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\left(\hat{\beta}_{Y_j}-\theta\cdot\hat{\beta}_{X_j,\text{RB}}-r_j\right)^2}{\sigma^2_{\hat{Y}_j}}-\frac{1}{2}\sum_{j\in\mathcal{S}_\lambda}\frac{\theta^2\sigma^2_{X_j,RB}}{\sigma^2_{\hat{Y}_j}}$

Back to main slides

Details on coordinate descent algorithm

- We start with an initial guess of θ, denoted as $\theta^{(0)}$, which can be either 0 or generated from a distribution: $\theta^{(0)} \sim \text{Uniform}\left(\min_{1\le j\le M} \hat{\beta}_{Yj}/\hat{\beta}_{Xj}, \max_{1\le j\le M} \hat{\beta}_{Yj}/\hat{\beta}_{Xj}\right)$

- At iteration $k+1$, we update r_j as follows. We order decreasingly

$$\frac{\left(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j,\text{RB}} - r_j\right)^2 - \theta^2 \cdot \sigma^2_{X_j,\text{RB}}}{\sigma^2_{\hat{Y}_j}}, \qquad j = 1,2,\ldots,M\,.$$

Then we set $r_j^{(k+1)} = \hat{\beta}_{Yj} - \theta^{(k)} \cdot \hat{\beta}^{(k)}_{X_j,\text{RB}}$ for the largest K component $j = 1,\ldots,K$ and $r_j^{(k+1)} = 0$ for $j = K+1,\ldots,M$

- We next update θ by RIVW formula

- We iterate the above two steps to update r_j and θ coordinately until the difference between $\theta^{(k+1)}$ and $\theta^{(k)}$ is small

Back to main slides

Discussion

- Applied MR studies require domain expertise; valid IV selections for MR analyses involve many steps (e.g., removing IVs with potential pleiotropic effect, etc.);
- Researchers report IVW estimators as their main results and use sensitive analyses and robust MR methods to confirm their findings
- Our new method can serve as one type of robust MR method, which considers winner's curse bias, measurement errors in IVs, and relax valid IV assumptions.

Back to main slides