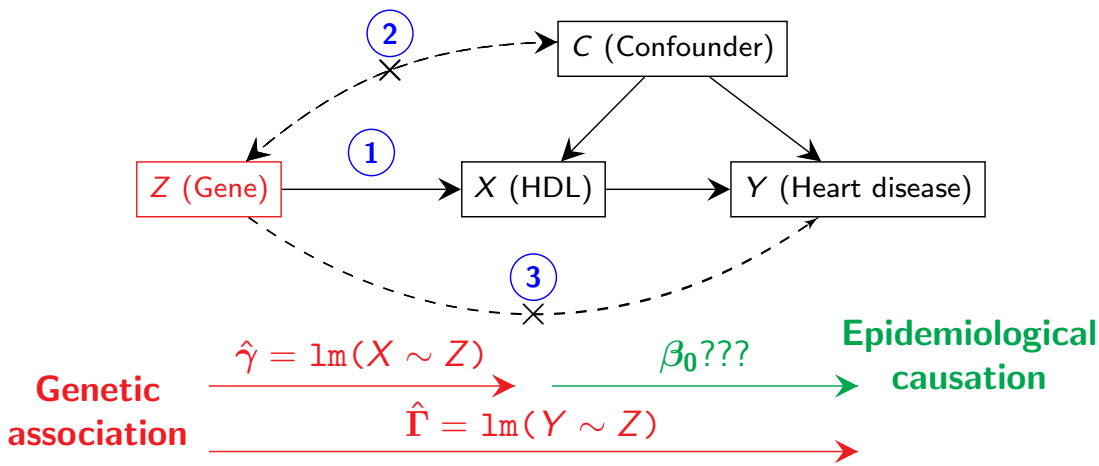
Mendelian randomization: From genetic association to epidemiological causation

Qingyuan Zhao

Department of Statistics, The Wharton School, University of Pennsylvania

April 24, 2018



Motivation: Epidemiology of cardiovascular diseases

- Cardiovascular diseases take the lives of 17.7 million people every year, 31% of all global deaths.[1]
- Risk factors: hypertension, high cholesterol, smoking, ...
- Ascertainment of a risk factor requires a large body of studies.

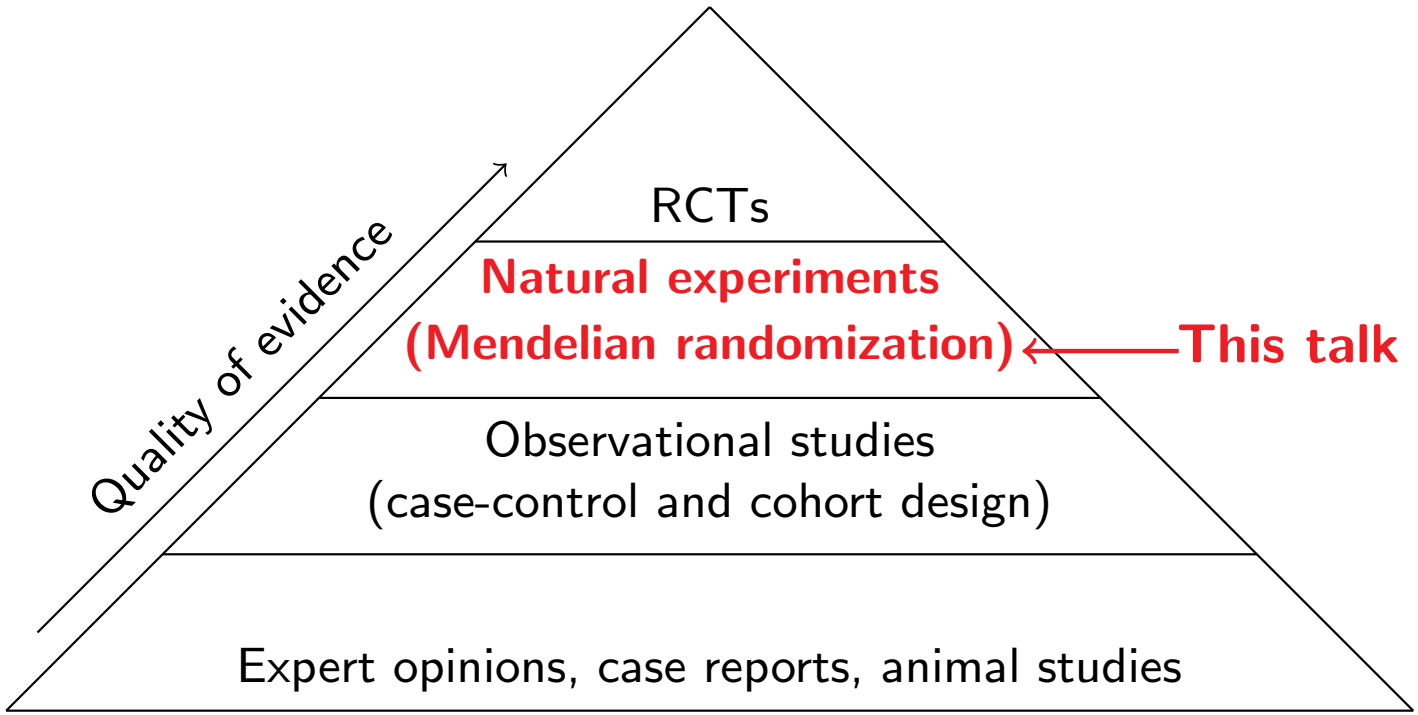


Figure: (A rough) Hierarchy of evidence.[2]

[1] Source: World Health Organization `www.who.int/cardiovascular_diseases/en/`

[2] Based on: American Academy of Pediatrics clinical guidelines. Gidding, et al. (2012). "Developing the 2011 Integrated Pediatric Guidelines for Cardiovascular Risk Reduction." *Pediatrics* 129(5).

The Lipid Hypothesis

"Decreasing blood cholesterol significantly reduces the risk of cardiovascular diseases."[3]

1913	First evidence from a **rabbit study**.
1950s – 1980s	Accumulation of evidence from **observational studies**. Transformation to **the LDL hypothesis**.
1970s	Discoveries of the regulation of LDL cholesterol → Brown and Goldstein winning the Nobel prize in 1985.
1980s	More evidence from **US Coronary Primary Prevention Trial**.
1990s	Skepticism continue until landmark **statin trials**.
2010s	Reaffirmation from **Mendelian randomization**.

However, the role of **HDL cholesterol** remains quite controversial.

[3] History based on: Academy of Medical Sciences Working Group (2007). "Identifying the environmental causes of disease: how should we decide what to believe and when to take action?" Academy of Medical Sciences.

The HDL Hypothesis

"HDL is protective against heart diseases."[4]

1960s Formulation of the hypothesis from **observational studies**. The inverse association has been firmly established over the years.

1980s Supporting evidence from **animal studies**.

But... 2000s Null findings from **studies of Mendelian disorders**.

2010s Failed **RCTs**, though each has its own caveats.

2010s Null findings from **Mendelian randomization**.

> *"I'd say* **the HDL hypothesis is on the ropes right now**,*" said Dr. James A. de Lemos . . . Dr. Kathiresan said. "I tell them, ' It means you are at increased risk, but* **I don't know if raising it will affect your risk.**"' — *New York Times, May 16, 2012.*

- Reasons of null findings: flawed design, lack of power, HDL function hypothesis . . .
- **We will reassess the evidence for HDL using a new design and new statistical methods of Mendelian randomization.**

[4] History based on: Rader and Hovingh (2014). "HDL and cardiovascular disease" *Lancet* 384.

Fundamental challenge of observational studies

"Correlation is not causation".

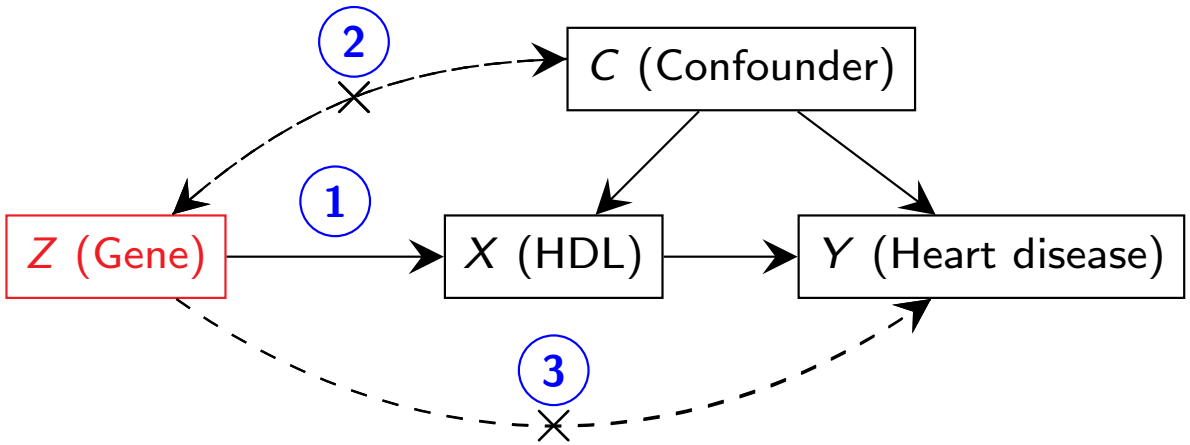
Observational studies = Enumerating confounders

- Idea: Conditioning on possible sources of spurious correlation.
- For HDL and heart disease, confounders include:
  - Age.
  - Sex.
  - Smoking status.
  - Diabetes.
  - Blood pressure.
  - ...
- **Fundamental challenge: We can never be sure this list is complete.**
- The promise of Mendelian randomization: unbiased estimation of causal effect without enumerating confounders.

What is Mendelian randomization (MR)?

"Using genetic variants as instrumental variables."

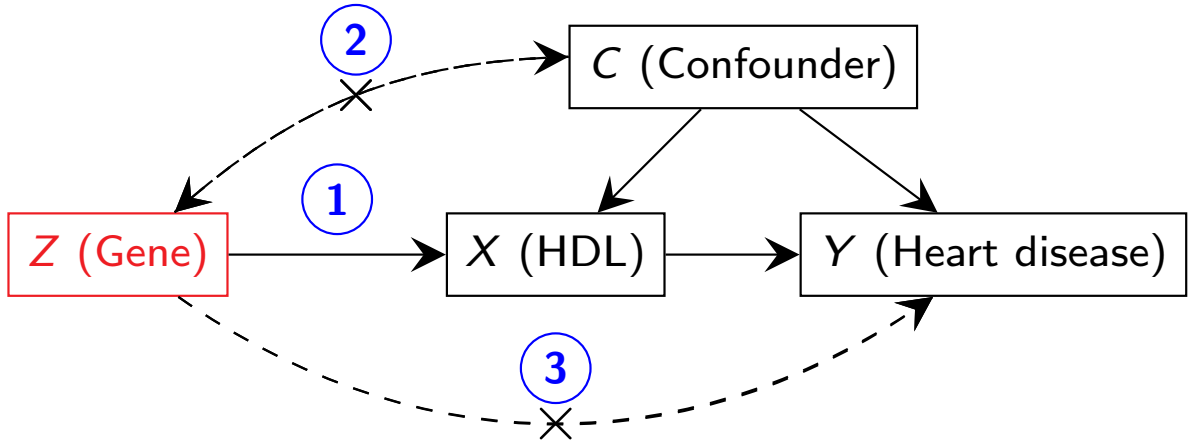
Causal diagram for instrumental variables (IV)



Core IV assumptions

1. **Relevance**: Z is associated with the exposure (X).
2. **Effective random assignment**: Z is independent of the unmeasured confounder (C).
3. **Exclusion restriction**: Z cannot have any direct effect on the outcome (Y).

Examine the core IV assumptions for MR



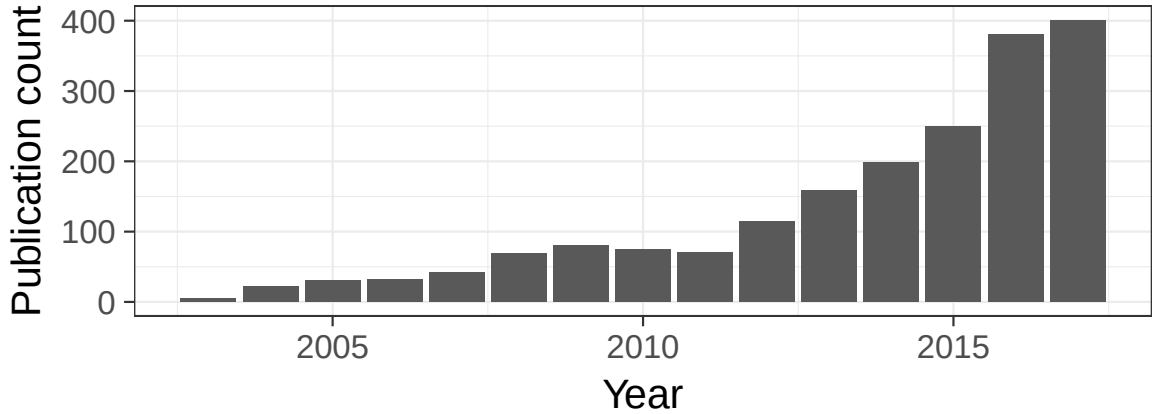
Criterion (1)	✓	Massive pool of potential IVs, Large-scale GWAS identifies many causal variants
Criterion (2)	✓	Due to Mendel's Second Law
Criterion (3)	?	Problematic because of wide-spread **pleiotropy** (multiple functions of genes).

Additional challenges

- Many genetic variants are only weakly associated with X.
- Most GWAS data come in summary-statistics format due to privacy.

MR studies in epidemiology

Surging interest in MR[5]



- MR methods are also increasingly used in human genetics.[6]

Conventional design: a 2012 MR study of HDL in *Lancet*[7]

> *Methods . . . First, we used as an instrument* **a single nucleotide polymorphism (SNP)** *in the endothelial lipase gene (LIPG Asn396Ser) . . . Second, we used as an instrument* **a genetic score consisting of 14 common SNPs** *that exclusively associate with HDL cholesterol . . .*

[5] Thomson Reuters *Web of Science*, topic "Mendelian randomization", `www.webofknowledge.com`.

[6] Gamazon, E. et al. (2015). "A gene-based association method for mapping traits using reference transcriptome data." *Nature Genetics* 47.

[7] Example from: Voight et al. (2012). "Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study." *Lancet* 380: 572–580.

New methods for MR

Part 1: Increased robustness to pleiotropy

We will derive an estimator that is robust to **both**

1. **Sparse** pleiotropy/invalid IV.
   - Works of Hyunseung Kang and coauthors.[8]
2. **Dense but balanced** pleiotropy.
   - Works of Jack Bowden, Stephen Burgess and coauthors (e.g. MR-Egger).[9]

Part 2: Increased efficiency in genome-wide MR

- Due to "missing heritability", we would like to use as many SNPs as possible to gain statistical power.
- Example: for height, there are extremely large number of causal variants **tiny effect sizes**, **spreading widely** across the genome.[10]
- Statistical insights are needed to guarantee increased efficiency.

[8] Kang, H. et al. (2016). "Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization." *Journal of American Statistical Association*, 111.

[9] Bowden, J. et al. (2015). "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression." *International Journal of Epidemiology*, 44.

[10] Shi, H. et al. (2016). "Contrasting the genetic architecture of 30 complex traits from summary association data." *American Journal of Human Genetics*, 99. See also a 2017 *Cell* paper by Boyle et al.

Rest of the talk

Part 0: Data Structure & Modeling Assumptions

Part 1: Increased robustness to pleiotropy

Part 2: Increased efficiency in genome-wide MR

Outline

Part 0: Data Structure & Modeling Assumptions

Part 1: Increased robustness to pleiotropy

- Evolution of pleiotropy models: Assumption 2.1 $\rightarrow$ 2.2 $\rightarrow$ 2.3
- Evolution of statistical methods: PS $\rightarrow$ APS $\rightarrow$ RAPS
- Example: BMI and blood pressure

Part 2: Increased efficiency in genome-wide MR

- RAPS with Empirical Partially Bayes
- Example: HDL and Coronary Heart Disease

Working example

Instrumental variables $\boldsymbol{Z}_{1:p}$: Single nucleotide polymorphisms (SNPs).

Exposure variable X: Body mass index (BMI).

Outcome variable Y: Systolic blood pressure (SBP).

Data preprocessing for two-sample summary-data MR

Dataset	`BMI-FEM`	`BMI-MAL`	`SBP-UKBB`
Source	GIANT (female)	GIANT (male)	UK BioBank
Sample size	171977	152893	317754
GWAS	$\text{lm}(X \sim Z_j)$	$\text{lm}(X \sim Z_j)$	$\text{lm}(Y \sim Z_j)$
Coefficient	Used for selection	$\hat{\gamma}_j$	$\hat{\Gamma}_j$
Std. Err.		σ_{Xj}	σ_{Yj}

Step 1 Use `BMI-FEM` to select significant (p-value $\leq 5 \times 10^{-8}$) and independent SNPs ($p = 25$).

Step 2 Use `BMI-MAL` to obtain $(\hat{\gamma}_j, \sigma_{Xj})$, $j = 1 : p$.

Step 3 Use `SBP-UKBB` to obtain $(\hat{\Gamma}_j, \sigma_{Yj})$, $j = 1 : p$.

Assumption 1

Measurement error model

$$\begin{pmatrix}\hat{\boldsymbol{\gamma}}\\ \hat{\boldsymbol{\Gamma}}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\gamma}\\ \boldsymbol{\Gamma}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_X & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_Y\end{pmatrix}\right), \quad \begin{aligned}\boldsymbol{\Sigma}_X &= \mathrm{diag}(\sigma^2_{X1}, \ldots, \sigma^2_{Xp}),\\ \boldsymbol{\Sigma}_Y &= \mathrm{diag}(\sigma^2_{Y1}, \ldots, \sigma^2_{Yp}).\end{aligned}$$

Pre-processing warrants Assumption 1

Dataset	`BMI-FEM`	`BMI-MAL`	`SBP-UKBB`
GWAS	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(X \sim Z_j)$	$\mathrm{lm}(Y \sim Z_j)$
Coefficient	Used for selection	$\hat{\gamma}_j$	$\hat{\Gamma}_j$
Std. Err.		σ_{Xj}	σ_{Yj}

- Large sample size $\Rightarrow$ CLT.
- Independence due to
  1. Non-overlapping samples (in all three datasets).
  2. Independent SNPs.

Assumption 2

Linking the genetic associations

The causal effect β_0 satisfies $\boldsymbol{\Gamma} \approx \beta_0 \boldsymbol{\gamma}$. This contains two claims:

1. The relationship is **approximately linear**.
2. The slope β_0 has a **causal interpretation**.

Heuristic: Linear structural equation model

Assume all the IVs are valid.

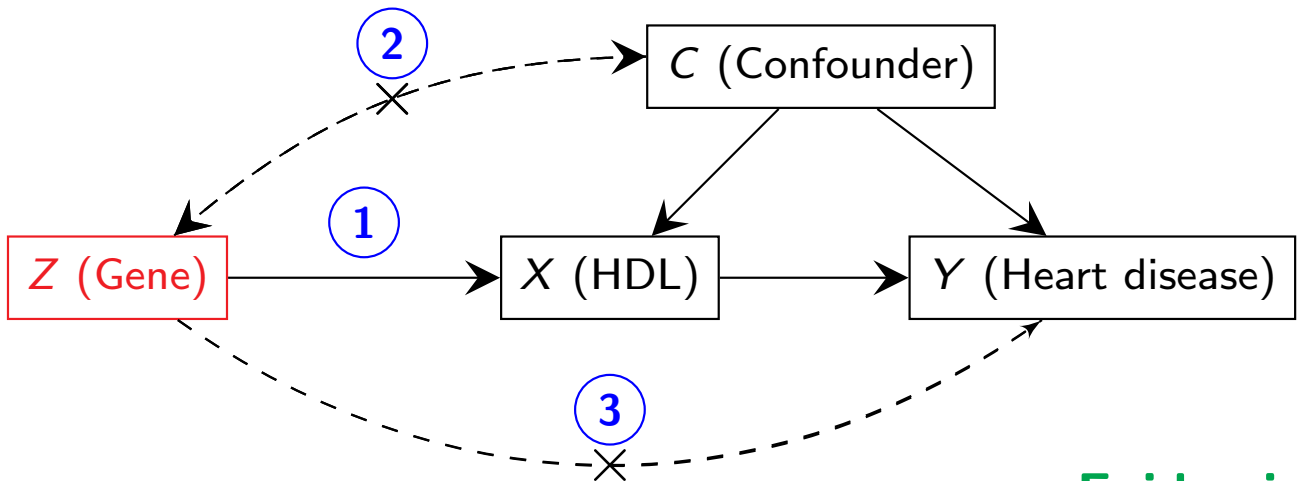
$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X C + E_X,$$

$$Y = \beta_0 X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y C + E_Y$$

$$= \sum_{j=1}^{p} \underbrace{(\beta_0 \gamma_j)}_{\boldsymbol{\Gamma}_j} Z_j + \underbrace{\sum_{j=1}^{p} \alpha_j Z_j}_{\textbf{0 by exclusion restriction}} + \underbrace{f(C, E_X, E_Y)}_{\textbf{independent of Z}}$$

Statistical problem

Genetic association $(\hat{\gamma}_j, \hat{\Gamma}_j, \sigma_{Xj}, \sigma_{Yj})_{j=1:p}$ $\overset{\text{inference}}{\Longrightarrow}$ Epidemiological causation β_0



Genetic association $\hat{\gamma} = \texttt{lm}(X \sim Z)$ β_0??? **Epidemiological causation**

$\hat{\Gamma} = \texttt{lm}(Y \sim Z)$

Outline

Part 0: Data Structure & Modeling Assumptions

Part 1: Increased robustness to pleiotropy

- Evolution of pleiotropy models: Assumption 2.1 $\rightarrow$ 2.2 $\rightarrow$ 2.3
- Evolution of statistical methods: PS $\rightarrow$ APS $\rightarrow$ RAPS
- Example: BMI and blood pressure

Part 2: Increased efficiency in genome-wide MR

- RAPS with Empirical Partially Bayes
- Example: HDL and Coronary Heart Disease

Assumptions 1 & 2.1 $\Longrightarrow$ Profile score (PS)

Asumption 2.1: No pleiotropy

The linear relation $\Gamma_j = \beta_0 \gamma_j$ is true for every $j = 1, \ldots, p$.

- Log-likelihood of the data (up to additive constant):

$$l(\beta, \gamma_1, \ldots, \gamma_p) = -\frac{1}{2}\left[\sum_{j=1}^{p} \frac{(\hat{\gamma}_j - \gamma_j)^2}{\sigma^2_{Xj}} + \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \gamma_j \beta)^2}{\sigma^2_{Yj}}\right].$$

- Profile likelihood: $l(\beta) = \max_{\boldsymbol{\gamma}} l(\beta, \boldsymbol{\gamma}) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{\sigma^2_{Yj} + \beta^2 \sigma^2_{Xj}}$.
- The MLE solves the profile score (PS) equation $l'(\hat{\beta}_{\mathrm{PS}}) = 0$.
- This estimator is an extension of the **limited information maximum likelihood (LIML)** of Anderson and Rubin (1949)[11] to the two-sample summary-data setting.
- Consistency and asymptotic normality can be proven.

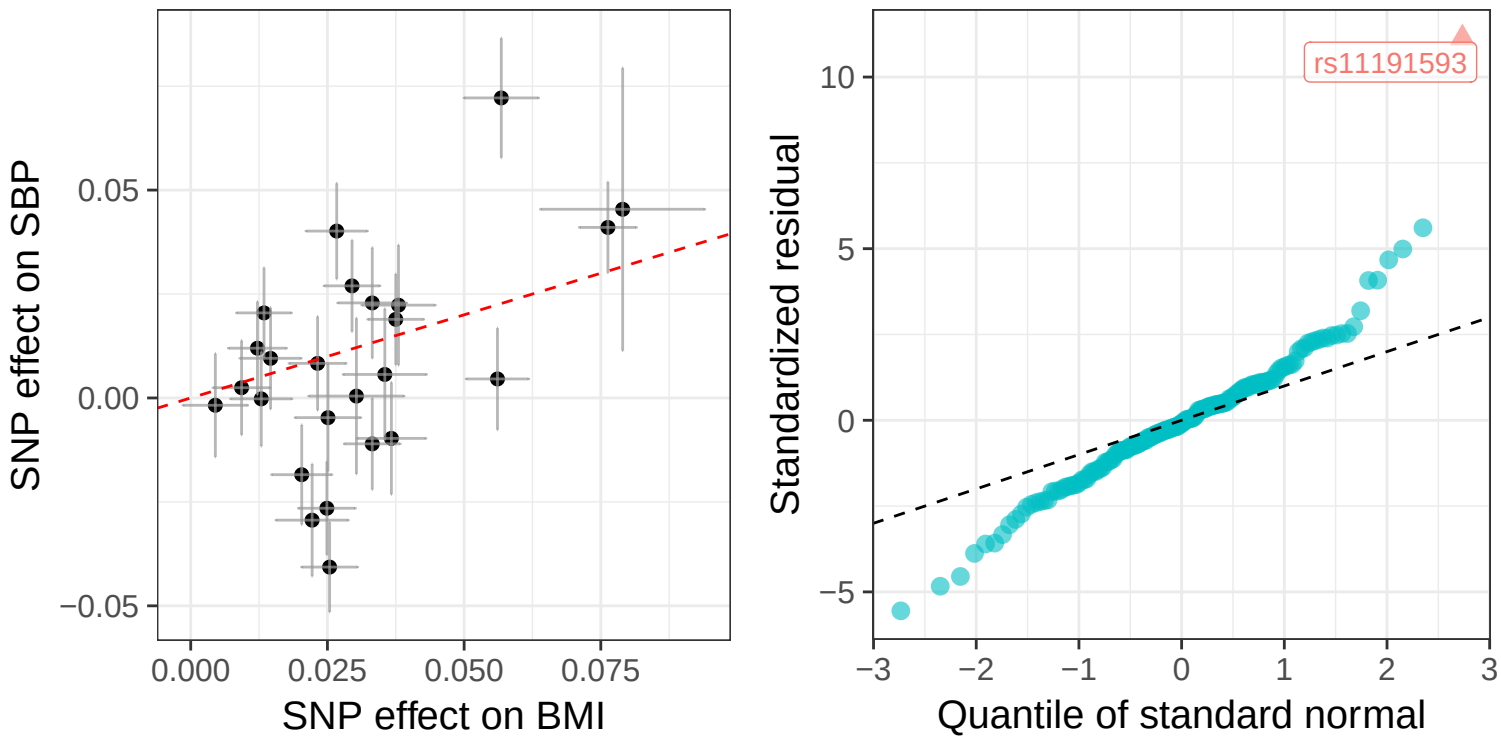
[11] Anderson, T. , & Rubin, H. (1949). "Estimation of the parameters of a single equation in a complete system of stochastic equations." *Annals of Mathematical Statistics*, 20.

Diagnostic plots show clear overdispersion

BMI-SBP Example (continued)

- Left ($p = 25$, $p_{\text{sel}} < 5 \cdot 10^{-8}$): Scatter-plot of GWAS summary data.
- Right ($p = 160$, $p_{\text{sel}} < 10^{-4}$): Q-Q plot of standardized residual

$$\hat{t}_j = \frac{\hat{\Gamma}_j - \hat{\beta}\hat{\gamma}_j}{\sqrt{\hat{\beta}^2\sigma_{Xj}^2 + \sigma_{Yj}^2}}.$$



Why Assumption 2.1 failed?

Answer: pleiotropy (direct effect on the outcome).

Heuristic: Linear structural equation model (with invalid IVs)

$$X = \sum_{j=1}^{p} \gamma_j Z_j + \eta_X C + E_X,$$

$$Y = \beta_0 X + \sum_{j=1}^{p} \alpha_j Z_j + \eta_Y C + E_Y$$

$$= \sum_{j=1}^{p} \underbrace{(\beta_0 \gamma_j + \alpha_j)}_{\Gamma_j} Z_j \quad + \quad \underbrace{f(C, E_X, E_Y)}_{\text{independent of } \boldsymbol{Z}}$$

Assumption 2.2: Random independent pleiotropy

Assume $\alpha_j = \Gamma_j - \beta_0 \gamma_j$ is independent of γ_j and $\alpha_j \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau_0^2)$.

Assumption 2.2 is consistent with genetic theory

This ubiquitous pleiotropy model is consistent (or not inconsistent) with the current understanding of genetic effects:

- Fisher's infinitesimal model (1918).
- Leading edge perspective on pleiotropy[12]

> *"In summary, the omnigenic model of complex disease proposes that essentially* **any gene with regulatory variants in at least one tissue that contributes to disease pathogenesis is likely to have nontrivial effects** *on risk for that disease. Furthermore, the relative effect sizes are such that, since core genes are hugely outnumbered by peripheral genes,* **a large fraction of the total genetic contribution to disease comes from peripheral genes that do not play direct roles in disease**."

[12] Boyle, E. et al. (2017). "An expanded view of complex traits: from polygenic to omnigenic". *Cell* 169, p1177–1186

Back to statistics: Failure of the profile likelihood

- The profile likelihood under Assumption 2.2 is given by

$$l(\beta, \tau^2) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{(\sigma_{Yj}^2 + \tau^2) + \beta^2\sigma_{Xj}^2} + \log(\sigma_{Yj}^2 + \tau^2),$$

- Easy to verify

$$\mathbb{E}\Big[\frac{\partial}{\partial\beta} l(\beta_0, \tau_0^2)\Big] = 0.$$

- But **the other score function is biased**:

$$\frac{\partial}{\partial\tau^2} l(\beta, \tau^2) = \frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\big[(\sigma_{Yj}^2 + \tau^2) + \beta^2\sigma_{Xj}^2\big]^2} - \frac{1}{\sigma_{Yj}^2 + \tau^2}.$$

- This is not too surprising as we are profiling out p nuisance parameters $\gamma_1, \cdots, \gamma_p$ (**the Neyman-Scott problem**).

Assumptions 1 & 2.2 $\Longrightarrow$ Adjusted profile score (APS)

- We take the approach of McCullagh & Tibshirani (1990)[13] to modify the profile score

$$\psi_1(\beta, \tau^2) = -\frac{\partial}{\partial \beta} l(\beta, \tau^2),$$

$$\psi_2(\beta, \tau^2) = \sum_{j=1}^{p} \sigma_{Xj}^2 \left\{ \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\left[(\sigma_{Yj}^2 + \tau^2) + \beta^2\sigma_{Xj}^2\right]^2} - \frac{1}{(\sigma_{Yj}^2 + \tau^2) + \beta^2\sigma_{Xj}^2} \right\}.$$

- Trivial roots: $\beta \to \pm\infty$ or $\tau^2 \to \infty$.
- Let $\hat{\beta}_{\text{APS}}$ be the non-trivial finite solution.

Theorem

Let Assumptions 1 & 2.2 be given and assume $\sigma^2 = O(1/n)$ *and* $(\beta_0, p\tau_0^2)$ *is in a bounded set* $\mathcal{B}$*. If* $p \to \infty$ *and* $p/n^2 \to 0$*, then*

1. *With probability going to* 1 *there exists a solution in* $\mathcal{B}$*.*
2. *All solutions in* $\mathcal{B}$ *are consistent:* $\hat{\beta}_{\text{APS}} \xrightarrow{p} \beta_0$ *and* $p\hat{\tau}_{\text{APS}}^2 - p\tau_0^2 \xrightarrow{p} 0$*.*

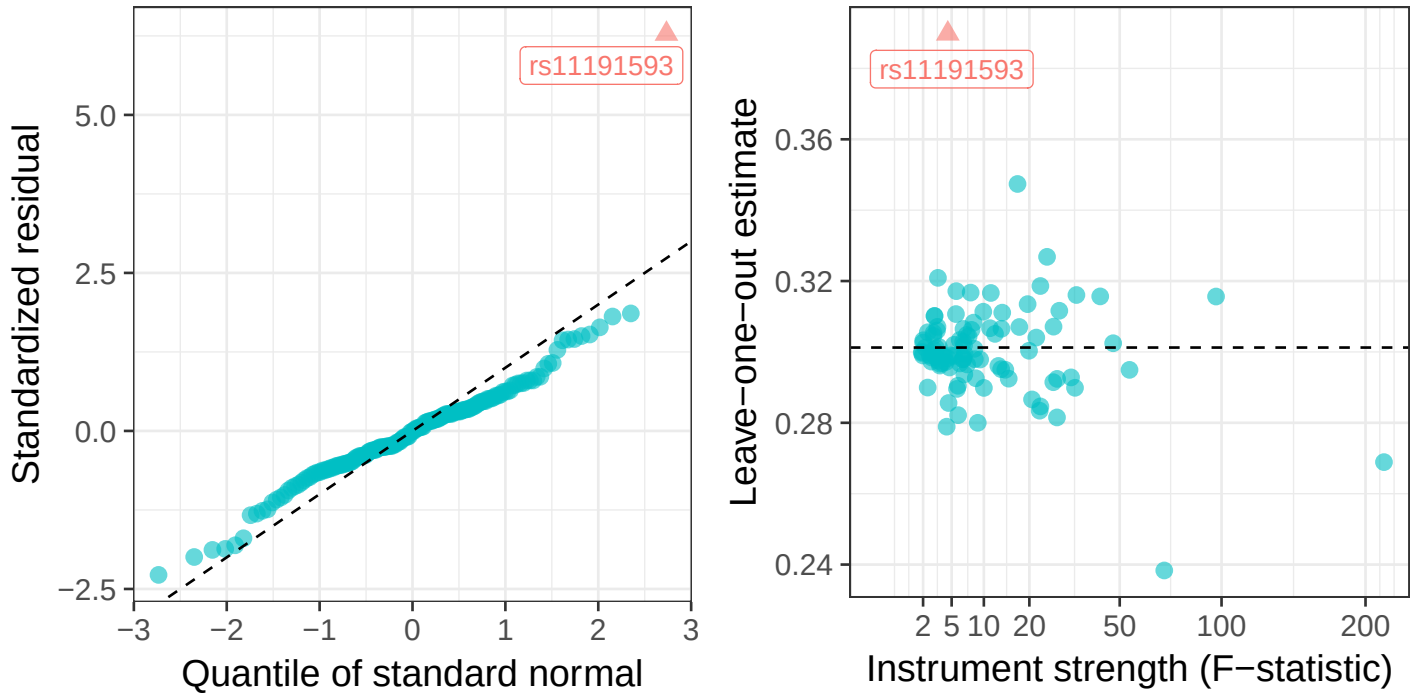
- Can obtain asymptotic normality assuming $p/n \to \lambda \in (0, \infty)$.

[13] McCullagh, P. & Tibshirani, R. (1990). "A simple method for the adjustment of profile likelihoods". *Journal of the Royal Statistical Society. Series B (Methodological)*, 52.

Diagnostic plots show influential outlier

- Same 160 SNPs ($p_{\text{sel}} < 10^{-4}$).

Left: Q-Q plot of std. residuals; Right: Influence of a single SNP.



- A clear outlier: rs11191593, with high influence (right plot).
- A GWAS catalog search reveals that this SNP is strongly associated with reticulocyte (immature red blood cell) count.[14]
- Slightly underdispersed (probably because β is underestimated).

[14] Astle, W. et al. (2016). "The allelic landscape of human blood cell trait variation and links to common complex disease." Cell 167: 1415-1429.

Assumptions 1 & 2.3 $\Longrightarrow$ RAPS

Assumption 2.3: Random pleiotropy with outliers

Most $\alpha_j \sim \mathrm{N}(0, \tau_0^2)$, but some $|\alpha_j|$ might be very large.

Robust adjusted profile score (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{(\sigma_{Yj}^2 + \tau^2) + \beta^2\sigma_{Xj}^2}}$.
- For some robust loss function ρ, the RAPS are

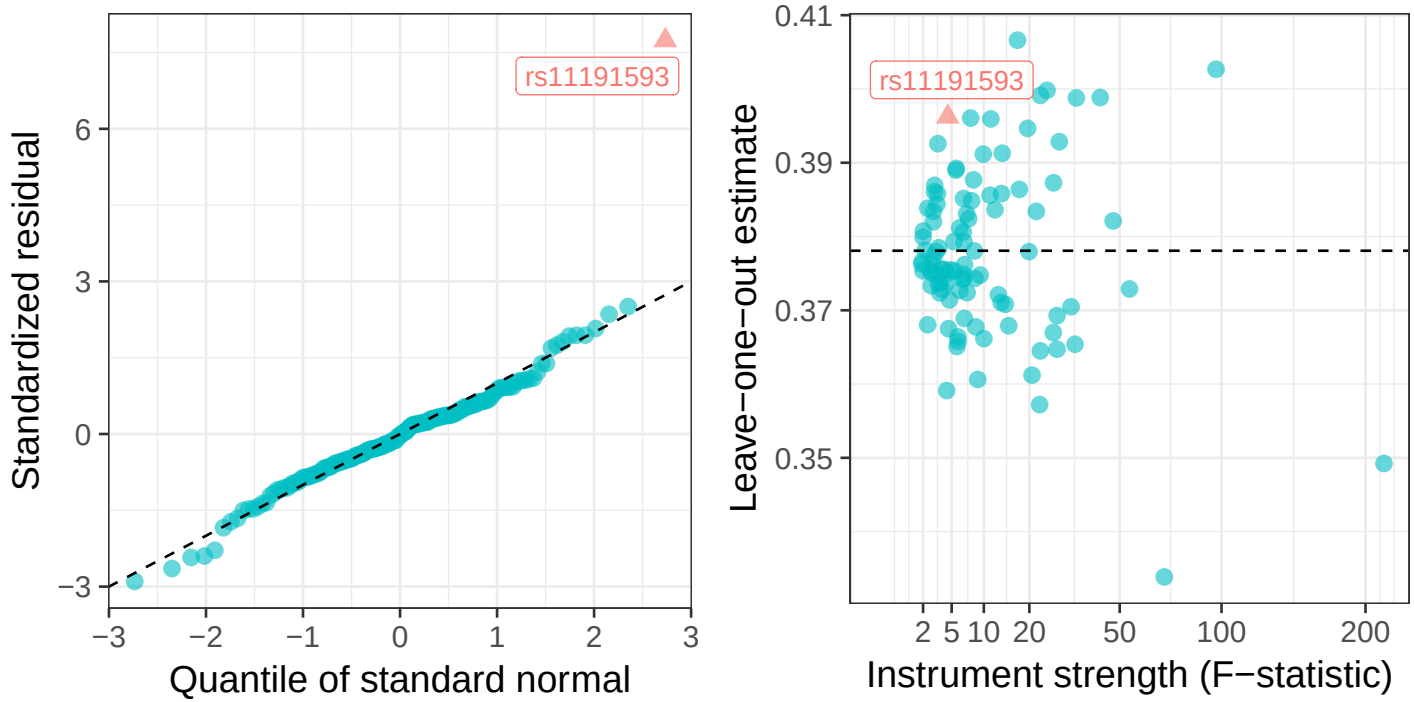
$$\psi_1^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \rho'(t_j) \cdot \frac{\partial}{\partial\beta} t_j,$$

$$\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \sigma_{Xj}^2 \frac{t_j \cdot \rho'(t_j) - \mathbb{E}[T\rho'(T)]}{(\sigma_{Yj}^2 + \tau^2) + \beta^2\sigma_{Xj}^2}, \text{ for } T \sim \mathrm{N}(0, 1).$$

- Reduces to APS when $\rho(t) = t^2/2$.
- General theory is quite difficult, but local identifiability can be prove.
- Asymptotic normality can be established assuming consistency and additional technical conditions.

Diagnostic plots show satisfactory fit

- Same 160 SNPs, now using RAPS with Huber's loss function.



- Influence of the outlier `rs11191593` is limited.
- Can further reduce its influence using redescending score (e.g. Tukey's biweight).

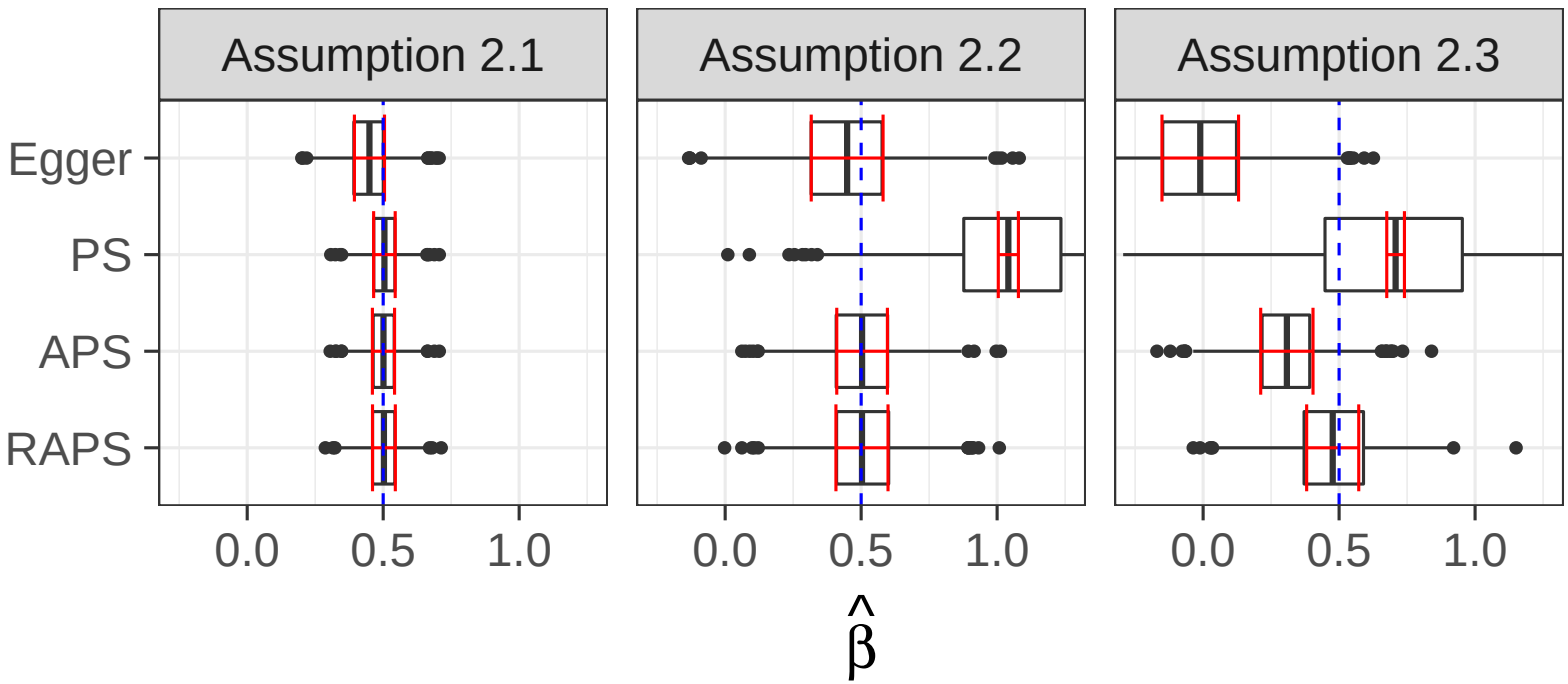
Comparison of the methods

In the BMI-SBP example

- **MR-Egger**: Weighted least squares of $\hat{\Gamma}_j$ against $\hat{\gamma}_j$ (ignoring measurement error in $\hat{\gamma}_j$ and weak IV bias).

MR-Egger	$\hat{\beta} = 0.51$ (SE 0.10)
Profile score (PS)	$\hat{\beta} = 0.61$ (SE 0.05)
Adjusted PS (APS)	$\hat{\beta} = 0.30$ (SE 0.16)
Robust APS (RAPS) w. Huber	$\hat{\beta} = \mathbf{0.38}$ **(SE 0.12)**

In a simulation study



Outline

Part 0: Data Structure & Modeling Assumptions

Part 1: Increased robustness to pleiotropy

- Evolution of pleiotropy models: Assumption 2.1 $\rightarrow$ 2.2 $\rightarrow$ 2.3
- Evolution of statistical methods: PS $\rightarrow$ APS $\rightarrow$ RAPS
- Example: BMI and blood pressure

Part 2: Increased efficiency in genome-wide MR

- RAPS with Empirical Partially Bayes
- Example: HDL and Coronary Heart Disease

Towards genome-wide MR

Unsatisfactory property of profile score (PS)

- Using Taylor's expansion we can show $\mathrm{Var}(\hat{\beta}_{\mathrm{PS}}) \approx V_1/V_2^2$, where

$$V_1 = \sum_{j=1}^{p} \frac{\gamma_j^2\sigma_{Yj}^2 + \Gamma_j^2\sigma_{Xj}^2 + \sigma_{Xj}^2\sigma_{Yj}^2}{(\sigma_{Yj}^2 + \beta_0^2\sigma_{Xj}^2)^2}, \quad V_2 = \sum_{j=1}^{p} \frac{\gamma_j^2\sigma_{Yj}^2 + \Gamma_j^2\sigma_{Xj}^2}{(\sigma_{Yj}^2 + \beta_0^2\sigma_{Xj}^2)^2}.$$

- Paradoxical observation: adding a new SNP Z_{p+1} with $\gamma_{p+1} \approx 0$ increases the variance.
- This prohibits a truly "genome-wide" design of MR.

Semiparametric mixture model

- Why? Maximum likelihood is **not efficient** when $p \to \infty$!
- Key idea: do not maximize the likelihood over $\boldsymbol{\gamma}$, but over the (empirical) distribution of $\boldsymbol{\gamma}$.[15]
- However, the computation is intractable.

[15] Kiefer, J., & Wolfowitz, J. (1956). "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". *Annals of Mathematical Statistics*, 27, 887–906.

An alternative approach: Conditional score[16]

- Recall the log-likelihood of SNP j under **Assumption 2.1** is

$$l_j(\beta, \boldsymbol{\gamma}) = -\frac{(\hat{\gamma}_j - \gamma_j)^2}{2\sigma_{Xj}^2} - \frac{(\hat{\Gamma}_j - \gamma_j\beta)^2}{2\sigma_{Yj}^2}.$$

- Sufficient statistic for γ_j: $\hat{\gamma}_{j,\mathrm{MLE}}(\beta) = \dfrac{\hat{\gamma}_j/\sigma_{Xj}^2 + \beta\hat{\Gamma}_j/\sigma_{Yj}^2}{1/\sigma_{Xj}^2 + \beta^2/\sigma_{Yj}^2}$.
- Conditional score is defined as

$$C_j(\beta) = \frac{\partial}{\partial\beta} l_j(\beta, \boldsymbol{\gamma}) - \mathbb{E}\left[\frac{\partial}{\partial\beta} l_j(\beta, \boldsymbol{\gamma}) \,\middle|\, \hat{\gamma}_{j,\mathrm{MLE}}(\beta)\right] = \frac{\boldsymbol{\gamma_j}(\hat{\Gamma}_j - \beta\hat{\gamma}_j)}{\sigma_{Yj}^2 + \beta^2\sigma_{Xj}^2}.$$

- Observation 1: $\boldsymbol{\gamma_j}$ **only appears as weight** to "residual" $\hat{\boldsymbol{\Gamma}}_\mathbf{j} - \beta\hat{\boldsymbol{\gamma}}_\mathbf{j}$.
- Observation 2: $\hat{\boldsymbol{\gamma}}_{\mathbf{j},\mathbf{MLE}}(\boldsymbol{\beta})$ **is independent of** $\hat{\boldsymbol{\Gamma}}_\mathbf{j} - \boldsymbol{\beta}\hat{\boldsymbol{\gamma}}_\mathbf{j}$.
- A general class of unbiased estimating equations:

$$\sum_{j=1}^{p} \frac{\boldsymbol{f}(\hat{\gamma}_{j,\mathrm{MLE}}(\beta)) \cdot (\hat{\Gamma}_j - \beta\hat{\gamma}_j)}{\sigma_{Yj}^2 + \beta^2\sigma_{Xj}^2} = 0$$

- Reduces to MLE/profile score if f is identity.

[16] This method is based on Lindsay, B. (1985). "Using empirical partially Bayes inference for increased efficiency". *Annals of Statistics*, 13, 914–931.

Empirical partially Bayes

- Lindsay showed that the optimal choice is the Bayes estimate of γ_j

$$f(\hat{\gamma}_{j,\mathrm{MLE}}) = \mathbb{E}[\gamma_j | \hat{\gamma}_{j,\mathrm{MLE}}(\beta)].$$

- Since the distribution of $\boldsymbol{\gamma}$ is unknown, he suggested to use empirical Bayes.
- The entire approach is partially Bayes because only the nuisance parameters $\boldsymbol{\gamma}$ are modeled in a Bayesian way.

Implementation to genome-wide MR

- It is more convenient to model the effect sizes γ_j/σ_{Xj}.
- We find that a good prior is the **spike-and-slab Gaussian mixture**.
- **Selective shrinkage** is important to increase efficiency.
- The whole approach can be extended to Assumptions 2.2 & 2.3 to account for pleiotropy.

Application to HDL and coronary heart disease

Dataset

- Used a 2010 GWAS of blood lipids to select 1122 independent SNPs not associated with LDL or triglycerides (p-value ≥ 0.01).
- 23 SNPs were genome-wide significant for HDL.
- HDL dataset: an non-overlapping 2013 GWAS of blood lipids.
- Coronary artery disease dataset: CARDIoGRAMplusC4D consortium.

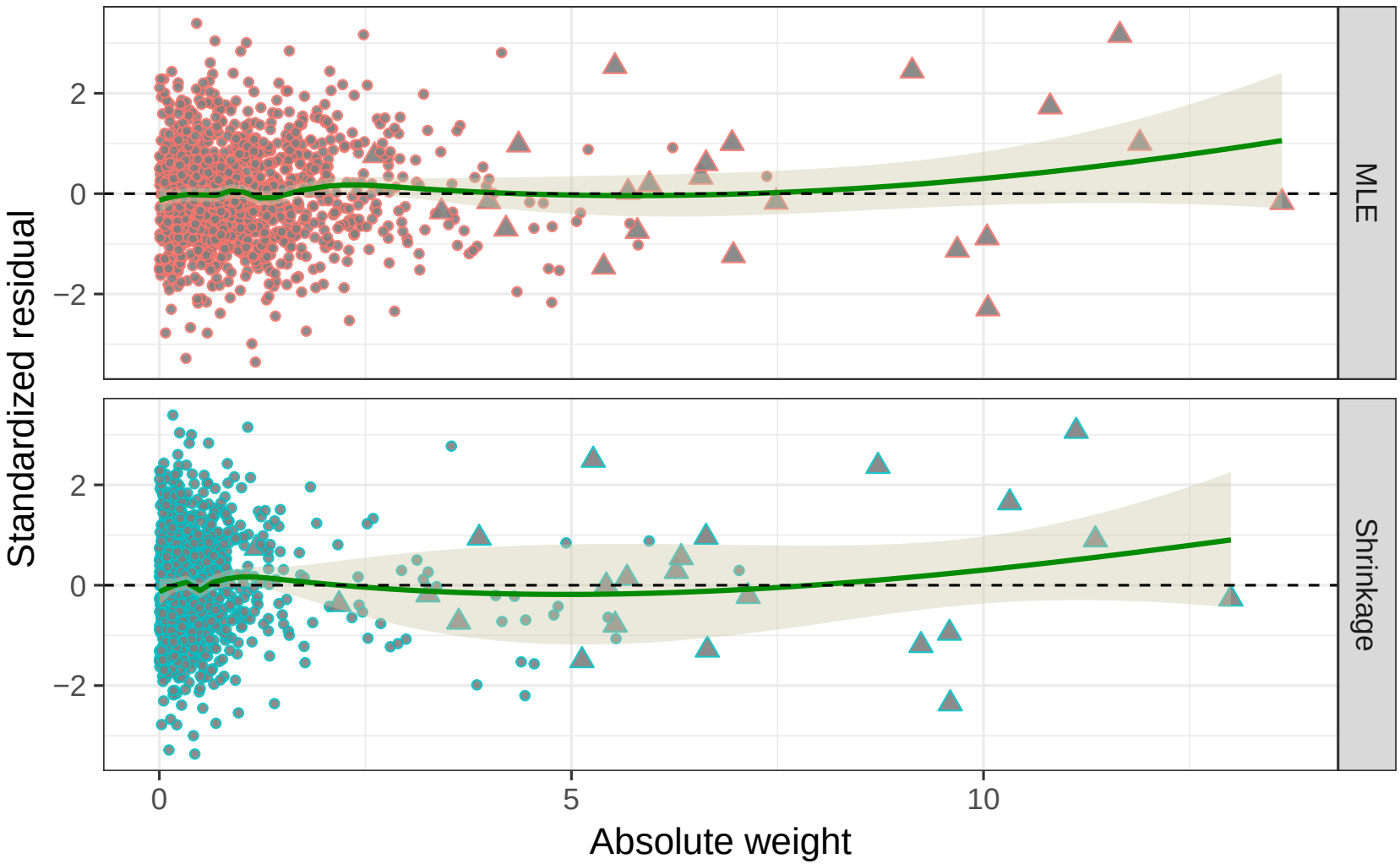
Fitted prior for γ_j/σ_{Xj}

- Spike: $p_1 = 0.91$, $\sigma_1 = 0.73$;
- Slab: $p_2 = 0.09$, $\sigma_2 = 4.57$.

Increase of efficiency (rough estimates from simulation)

Conventional MR $\overset{\uparrow 100\%}{\Longrightarrow}$ Genome-wide MR $\overset{\uparrow 20\%}{\Longrightarrow}$ Empirical partially Bayes.

Visualization of empirical partially Bayes



▲ 23 genome-wide significant SNPs in the selection GWAS.

• Rest 1099 SNPs.

Results

- Method: RAPS with Huber's loss + empirical partially Bayes.
- Scale: Odds ratio (95% CI) per 1 SD increase of LDL/HDL.

	LDL	HDL
Observational study		
2009 *JAMA*	**1.50 (1.39–1.61)**	**0.78 (0.74–0.82)**
Previous MR		
2012 *Lancet*	2.13 (1.69–2.69)	0.93 (0.68–1.26)
2016 *JAMA Cardiology*	1.68 (1.51–1.87)	0.95 (0.85–1.06)
New MR		
Significant SNPs	1.76 (1.53–2.03)	0.88 (0.74–1.04)
All SNPs	**1.61 (1.45–1.80)**	**0.82 (0.73–0.91)**

Caveats and future work

- It is unclear if the selected SNPs are truly unrelated to LDL or triglycerides $\Longrightarrow$ Our pleiotropy model might be insufficient.
- Estimates using strong instruments and weak instruments are not identical $\Longrightarrow$ The causal mechanism might be heterogeneous.
- The CARDIoGRAMplusC4D seems to have a small fraction of overlapping samples $\Longrightarrow$ We are working on a correction.

Conclusion

It is perhaps too soon to give up hope on the HDL hypothesis.

Acknowledgment

Collaborators

Jingshu Wang (Penn), Dylan Small (Penn), Jack Bowden (Bristol), Gibran Hemani (Bristol), Yang Chen (Michigan).

References

- Statistical inference in two-sample Mendelian randomization using robust adjusted profile score. arXiv: 1801.09652.
- A genome-wide design and an empirical partially Bayes approach to increase the power of Mendelian randomization, with application to the effect of blood lipids on cardiovascular disease. arXiv: 1804.07371.

Software

- R package `mr.raps` is available on CRAN.
- Can be directly called from the `TwoSampleMR` platform (`https://github.com/MRCIEU/TwoSampleMR`).