Leverage Mendelian Randomization to Learn Meaningful Representations (LMR$\times$2)

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

December 14, 2021

Outline

What is MR?

- Wikipedia definition:

  In epidemiology, Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in observational studies.

What is MR?

- Wikipedia definition:

  In epidemiology, Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in observational studies.

- Folk definition:

$$\text{MR} = \text{Use genetic variation as instrumental variables.}$$

What is MR?

- Wikipedia definition:

  In epidemiology, Mendelian randomization is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease in observational studies.

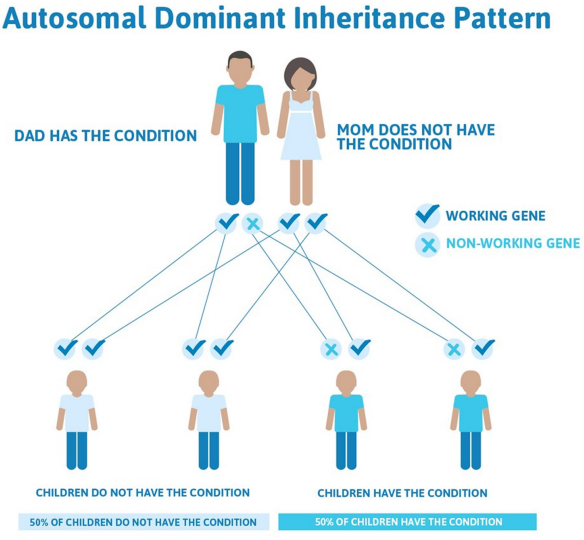
- Folk definition:

  MR = Use genetic variation as instrumental variables.

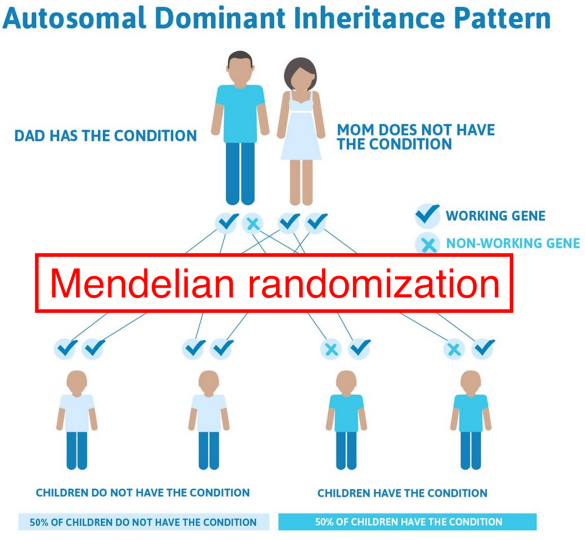
- A more informative definition:

  MR = Base causal inference on randomness in Mendelian inheritance.

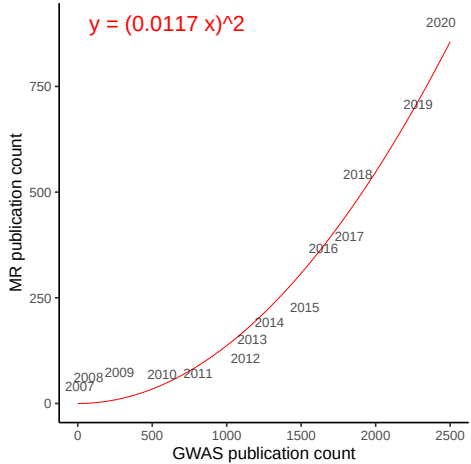
Heredity as a natural experiment



Heredity as a natural experiment



Surging popularity of MR

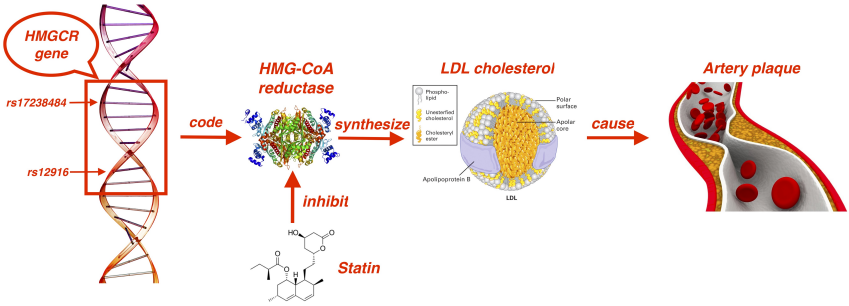


- Applications of MR are fueled by the increasing availability of GWAS datasets.[1]

[1] Data are obtained from Web of Science (`https://www.webofknowledge.com/`).

Example: Causal effect of the "bad" cholesterol

A well understood pathway of heart disease

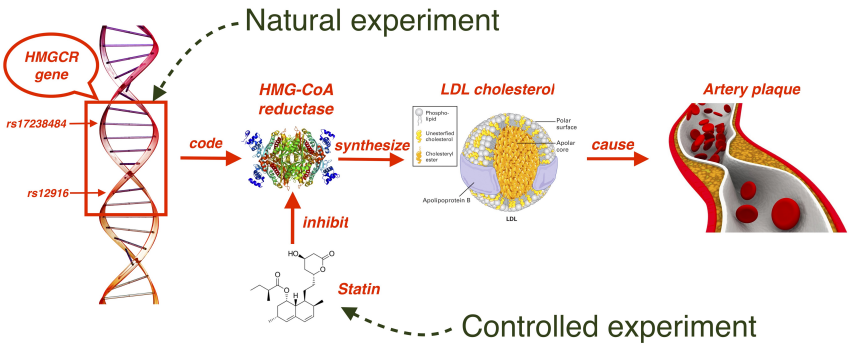


Basic idea

People who inherited certain alleles of *rs17238484* and *rs12916* have **naturally** higher concentration of LDL cholesterol.

Example: Causal effect of the "bad" cholesterol

A well understood pathway of heart disease

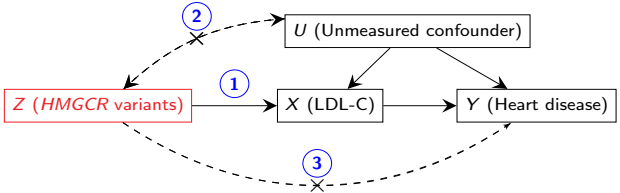


Basic idea

People who inherited certain alleles of *rs17238484* and *rs12916* have **naturally** higher concentration of LDL cholesterol.

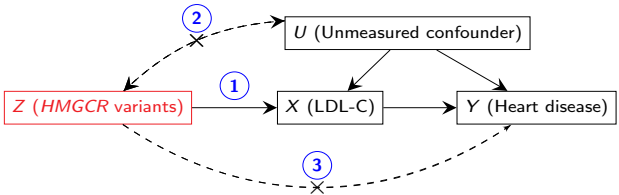
When do genetic instruments give correct answers?

The IV diagram



When do genetic instruments give correct answers?

The IV diagram

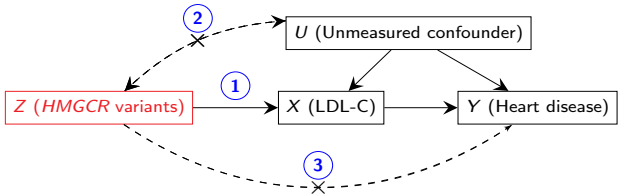


Must assume 3 core IV assumptions $\Longrightarrow$ Partial identification

1. **Relevance:** $Z \not\perp\!\!\!\perp X$.
2. **Exogeneity (natural experiment):** $Z \perp\!\!\!\perp U$.
3. **Exclusion restriction:** Z has no direct effect on Y.

When do genetic instruments give correct answers?

The IV diagram



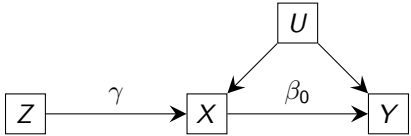
Must assume 3 core IV assumptions $\Longrightarrow$ Partial identification

1. **Relevance:** $Z \not\perp\!\!\!\perp X$.
2. **Exogeneity (natural experiment):** $Z \perp\!\!\!\perp U$.
3. **Exclusion restriction:** Z has no direct effect on Y.

Plus 1 extra assumption $\Longrightarrow$ Point identification

Could be linearity, monotonicity (Angrist, Imbens & Rubin, 1996), or homogeneity (Hernán & Robins, 2006; Wang & Tchetgen Tchetgen, 2018).

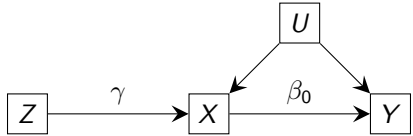
Basic idea: division



The Wald estimator

$$\text{Causal effect of } X \text{ on } Y\ (\beta_0) = \frac{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{Y}\ (\Gamma = \gamma \cdot \beta_0)}{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{X}\ (\gamma)}.$$

Basic idea: division



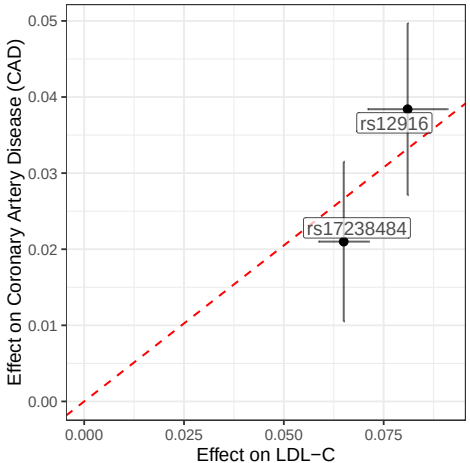
The Wald estimator

$$\text{Causal effect of } X \text{ on } Y\ (\beta_0) = \frac{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{Y}\ (\Gamma = \gamma \cdot \beta_0)}{\textbf{Causal effect of } \boldsymbol{Z} \textbf{ on } \boldsymbol{X}\ (\gamma)}.$$

Heuristic: Linear structural equation model

$$\begin{aligned} X &= \gamma Z + \eta_X U + E_X, \\ Y &= \beta_0 X + \eta_Y U + E_Y \\ &= (\beta_0 \gamma) Z + \underbrace{f(U, E_X, E_Y)}_{\text{independent of } Z} \end{aligned}$$

Example: Causal effect of LDL-cholesterol



Division $\overset{\text{in statistics}}{=\!=\!=}$ Regression with no intercept.

A main challenge to MR

Violation of exclusion restriction due to pleiotropy (multiple functions of genes)

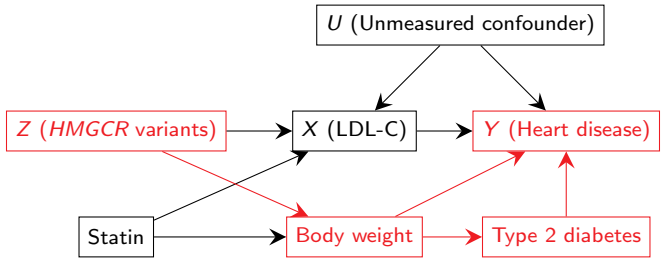
[2] Swerdlow, D. I., et al. "HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials." *Lancet* (2015).

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

A main challenge to MR

Violation of exclusion restriction due to pleiotropy (multiple functions of genes)

Example: *HMGCR* is associated with body weight[2]



- Recent genetic studies show that pleiotropy is indeed wide-spread.[3]

[2] Swerdlow, D. I., et al. "HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials." *Lancet* (2015).

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

Two ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Two ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-but-not-broken calipers!!

Two ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-but-not-broken calipers!!

Question 2: When is that enough?

Two ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-but-not-broken calipers!!

Question 2: When is that enough?

1. $< 50\%$ of the calipers are broken (Kang et al., 2016); or
2. Rusty readings are balanced around the truth (Bowden et al., 2015).

Two ideas to deal with pleiotropy

Useful metaphor: genetic instruments are rusty.

Question 1: What would you do if you have a rusty caliper?

Today's Answer: Find many rusty-but-not-broken calipers!!

Question 2: When is that enough?

1. $< 50\%$ of the calipers are broken (Kang et al., 2016); or
2. Rusty readings are balanced around the truth (Bowden et al., 2015).

Remaining issues

1. Both situations are common in MR.
2. Need to deal with many weak instruments.

Three-sample summary-data MR

- Sample 1: Select genetic variants associated with the hypothesized cause (LDL-C in the previous example; epidemiologists call this **exposure**).
- Sample 2: Obtain the GWAS summary data $(\hat{\gamma}_j, \sigma_{Xj}), j = 1, \ldots, p$ for the gene-exposure associations.
- Sample 3: Obtain the GWAS summary data $(\hat{\Gamma}_j, \sigma_{Yj}), j = 1, \ldots, p$ for the gene-outcome associations.

This is crucial for eliminating selection bias and the dependence between $\hat{\gamma}_j$ and $\hat{\Gamma}_j$.

Assumptions

Assumption 1: Measurement error model

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

Assumption 2: Random rusty calipers

The causal effect β satisfies $\boldsymbol{\Gamma} \approx \boldsymbol{\beta_0}\boldsymbol{\gamma}$. Specifically, let $\alpha_j = \Gamma_j - \beta\gamma_j$. Then we assume

- InSIDE (Instrument Strength Independent of Direct Effect): α_j is independent of γ_j;
- Most $\alpha_j \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \tau^2)$, but a few $|\alpha_j|$ might be very large.

These assumptions are based on extensive exploratory data analyses.

Robust adjusted profile scores (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1 + \beta^2\sigma_{Xj}^2 + \tau^2\sigma_{Yj}^2}}$.

[4] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

Robust adjusted profile scores (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1 + \beta^2\sigma_{Xj}^2 + \tau^2\sigma_{Yj}^2}}$.
- For some **robust loss ρ (let $\psi = \rho'$)**, the RAPS equations are

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

$$\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} t_j \cdot \psi(t_j) - \mathbb{E}[T\psi(T)], \text{ for } T \sim \mathrm{N}(0, 1).$$

[4] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

Robust adjusted profile scores (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1 + \beta^2\sigma_{Xj}^2 + \tau^2\sigma_{Yj}^2}}$.
- For some **robust loss ρ (let $\psi = \rho'$)**, the RAPS equations are

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

$$\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \boldsymbol{t_j \cdot \psi(t_j)} - \mathbb{E}[T\psi(T)], \text{ for } T \sim \mathrm{N}(0, 1).$$

- Roughly speaking, the first equation means that

$$\sum_{j=1}^{p} \begin{pmatrix}\textbf{Estimated quality}\\ \textbf{of instrument } \boldsymbol{j}\end{pmatrix} \cdot \begin{pmatrix}\textbf{Estimated error}\\ \textbf{of instrument } \boldsymbol{j}\end{pmatrix} = 0.$$

[4] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

Robust adjusted profile scores (RAPS)

- Define standardized residual: $t_j(\beta, \tau^2) = \dfrac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{1 + \beta^2\sigma_{Xj}^2 + \tau^2\sigma_{Yj}^2}}$.
- For some **robust loss ρ (let $\psi = \rho'$)**, the RAPS equations are

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

$$\psi_2^{(\rho)}(\beta, \tau^2) = \sum_{j=1}^{p} \boldsymbol{t_j \cdot \psi(t_j)} - \mathbb{E}[T\psi(T)], \text{ for } T \sim \mathrm{N}(0, 1).$$

- Roughly speaking, the first equation means that

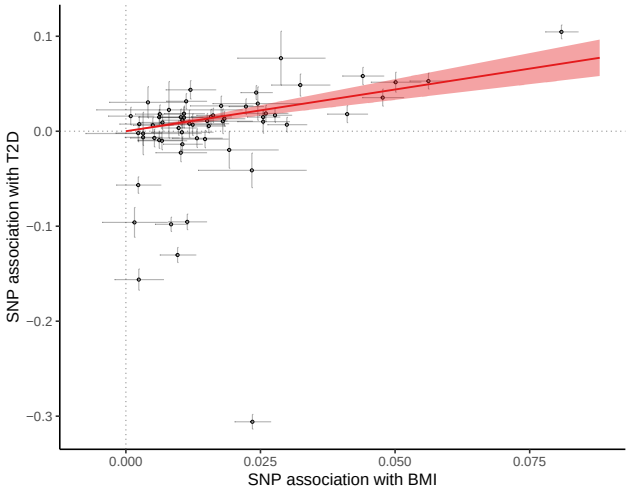
$$\sum_{j=1}^{p} \begin{pmatrix}\textbf{Estimated quality}\\ \textbf{of instrument } \boldsymbol{j}\end{pmatrix} \cdot \begin{pmatrix}\textbf{Estimated error}\\ \textbf{of instrument } \boldsymbol{j}\end{pmatrix} = 0.$$

- Estimated quality of the instruments can be improved by empirical Bayes, which works really well with many weak instruments.[4]

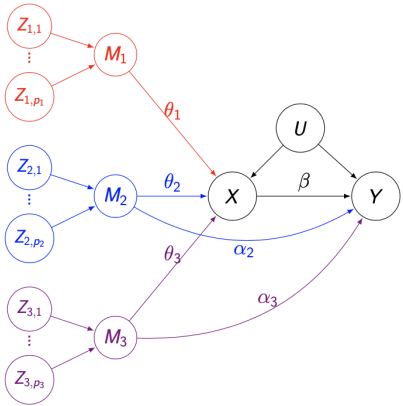
[4] Zhao, Q. et al. (2019). "Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization". *International Journal of Epidemiology*, 48(5):1478-1492.

Outline

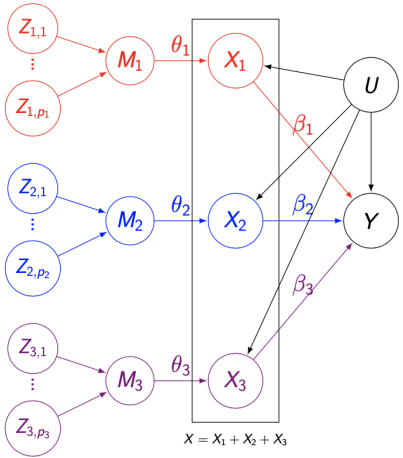
Motivating example: BMI and type 2 diabetes (T2D)



Two scenarios of mechanistic heterogeneity



(a) Scenario 1: Multiple pathways of horizontal pleiotropy.



(b) Scenario 2: Multiple mechanisms for the exposure X.

What would happen in each case?

If each diagram is interpretated as a linear structural equations model, we can derive the Wald ratio for each pathway.

Instruments Z	**Pathway** M	**Effect of** M **on** X	**Effect of** M **on** Y	**Wald estimand**
		Scenario 1		
$Z_{1,1}, \ldots, Z_{1,p_1}$	M_1	θ_1	$\theta_1\beta$	β
$Z_{2,1}, \ldots, Z_{2,p_2}$	M_2	θ_2	$\theta_2\beta + \alpha_2$	$\beta + \alpha_2/\theta_2$
$Z_{3,1}, \ldots, Z_{3,p_3}$	M_3	θ_3	$\theta_3\beta + \alpha_3$	$\beta + \alpha_3/\theta_3$
		Scenario 2		
$Z_{1,1}, \ldots, Z_{1,p_1}$	M_1	θ_1	$\theta_1\beta_1$	β_1
$Z_{2,1}, \ldots, Z_{2,p_2}$	M_2	θ_2	$\theta_2\beta_2$	β_2
$Z_{3,1}, \ldots, Z_{3,p_3}$	M_3	θ_3	$\theta_3\beta_3$	β_3

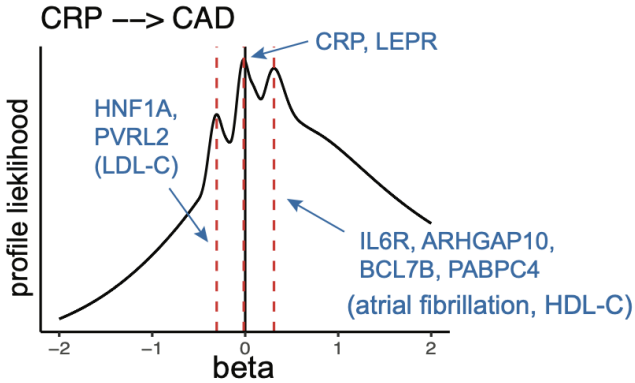
What would happen in each case?

If each diagram is interpretated as a linear structural equations model, we can derive the Wald ratio for each pathway.

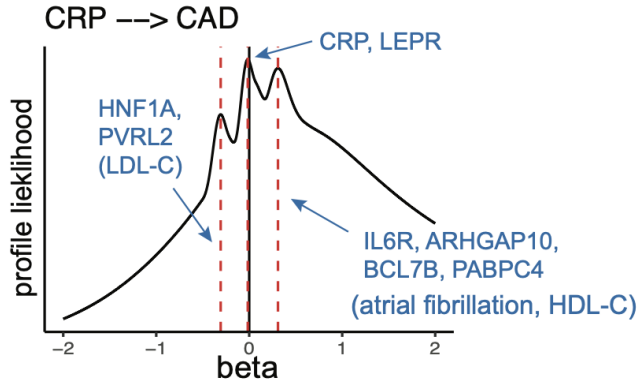
Instruments Z	**Pathway** M	**Effect of** M **on** X	**Effect of** M **on** Y	**Wald estimand**
		Scenario 1		
$Z_{1,1}, \ldots, Z_{1,p_1}$	M_1	θ_1	$\theta_1\beta$	β
$Z_{2,1}, \ldots, Z_{2,p_2}$	M_2	θ_2	$\theta_2\beta + \alpha_2$	$\beta + \alpha_2/\theta_2$
$Z_{3,1}, \ldots, Z_{3,p_3}$	M_3	θ_3	$\theta_3\beta + \alpha_3$	$\beta + \alpha_3/\theta_3$
		Scenario 2		
$Z_{1,1}, \ldots, Z_{1,p_1}$	M_1	θ_1	$\theta_1\beta_1$	β_1
$Z_{2,1}, \ldots, Z_{2,p_2}$	M_2	θ_2	$\theta_2\beta_2$	β_2
$Z_{3,1}, \ldots, Z_{3,p_3}$	M_3	θ_3	$\theta_3\beta_3$	β_3

- SNPs on the same pathway have the same Wald estimand, while SNPs across different pathways generally have different estimands.
- Mechanistic heterogeneity can arise even when all SNPs are valid instruments (Scenario 2).

Solution 1: Robust likelihood plot



Solution 1: Robust likelihood plot



- More detail: Wang, J., Zhao, Q., Bowden, J., Hemani, G., Smith, G. D., Small, D. S., & Zhang, N. R. (2021). Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLOS Genetics*. DOI:10.1371/journal.pgen.1009575.
- Also contains methods for multiple exposures and overlapping samples.

Solution 2: Modelling the effect along each path

Modified model

- GWAS summary data:

$$\begin{pmatrix} \hat{\gamma}_j \\ \hat{\Gamma}_j \end{pmatrix} \overset{\text{indep.}}{\sim} N\left(\begin{pmatrix} \gamma_j \\ \beta_j \gamma_j \end{pmatrix}, \begin{pmatrix} \sigma^2_{X_j} & 0 \\ 0 & \sigma^2_{Y_j} \end{pmatrix} \right), \quad j = 1, \ldots, p,$$

- **Mixture model** for path-specific effects:

$$\begin{aligned} Z_j &\sim \text{Categorical } (\pi_1, \ldots, \pi_K), \\ \beta_j | Z_j = k &\sim N(\mu_k, \sigma_k^2), \quad k = 1, \ldots, K. \end{aligned}$$

Solution 2: Modelling the effect along each path

Modified model

- GWAS summary data:

$$\begin{pmatrix} \hat{\gamma}_j \\ \hat{\Gamma}_j \end{pmatrix} \overset{\text{indep.}}{\sim} N\left(\begin{pmatrix} \gamma_j \\ \beta_j \gamma_j \end{pmatrix}, \begin{pmatrix} \sigma_{X_j}^2 & 0 \\ 0 & \sigma_{Y_j}^2 \end{pmatrix} \right), \quad j = 1, \ldots, p,$$

- **Mixture model** for path-specific effects:

$$\begin{aligned} Z_j &\sim \text{Categorical } (\pi_1, \ldots, \pi_K), \\ \beta_j | Z_j = k &\sim N(\mu_k, \sigma_k^2), \quad k = 1, \ldots, K. \end{aligned}$$

- More detail: Iong, D., Zhao, Q., & Chen, Y. (2020). A latent mixture model for heterogeneous causal mechanisms in Mendelian randomization. arXiv:2007.06476.

Solution 2: Modelling the effect along each path

Modified model

- GWAS summary data:

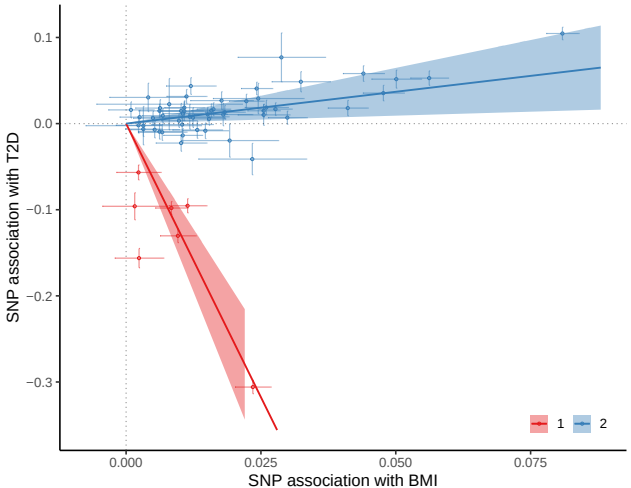
$$\begin{pmatrix} \hat{\gamma}_j \\ \hat{\Gamma}_j \end{pmatrix} \overset{\text{indep.}}{\sim} N\left(\begin{pmatrix} \gamma_j \\ \beta_j \gamma_j \end{pmatrix}, \begin{pmatrix} \sigma^2_{X_j} & 0 \\ 0 & \sigma^2_{Y_j} \end{pmatrix} \right), \quad j = 1, \ldots, p,$$

- **Mixture model** for path-specific effects:

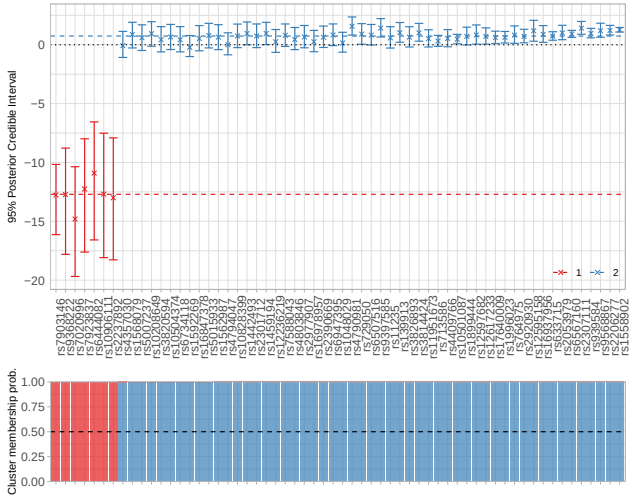
$$\begin{aligned} Z_j &\sim \text{Categorical}\ (\pi_1, \ldots, \pi_K), \\ \beta_j | Z_j = k &\sim N(\mu_k, \sigma^2_k), \quad k = 1, \ldots, K. \end{aligned}$$

- More detail: Iong, D., Zhao, Q., & Chen, Y. (2020). A latent mixture model for heterogeneous causal mechanisms in Mendelian randomization. arXiv:2007.06476.
- Alternative solution: **Bayesian model averaging**. See Shapland, C. Y., Zhao, Q., & Bowden, J. (2020). Profile-likelihood Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of horizontal pleiotropy. BioRxiv:2020.02.11.943712.

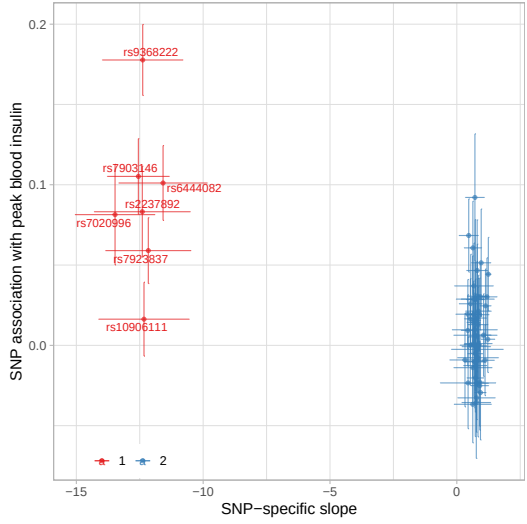
BMI-T2D example: Two-cluster fit



BMI-T2D example: Posterior intervals

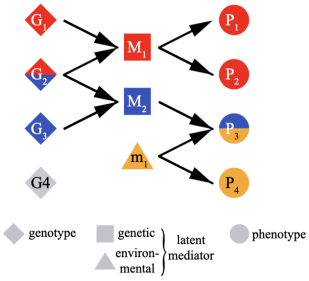


BMI-T2D example: A possible explanation



How can we discover the latent pathways? Examine the phenonome!

a1 Graphical representation

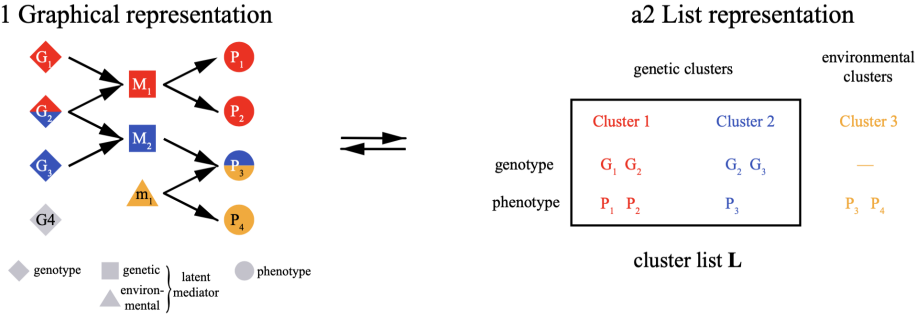


a2 List representation

	genetic clusters		environmental clusters
	Cluster 1	Cluster 2	Cluster 3
genotype	G_1 G_2	G_2 G_3	—
phenotype	P_1 P_2	P_3	P_3 P_4

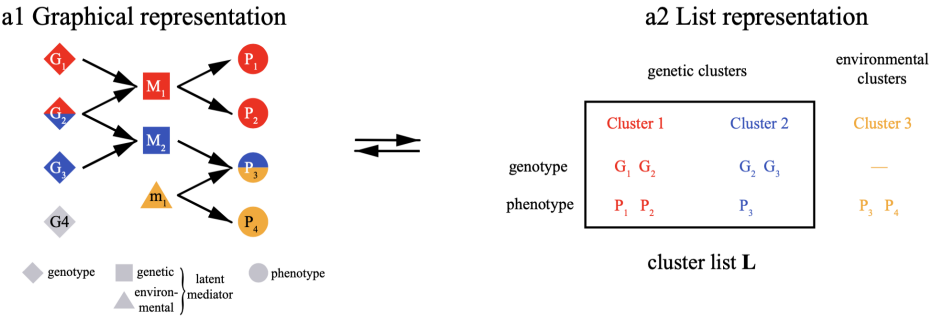
cluster list **L**

How can we discover the latent pathways? Examine the phenonome!



- If we let $\hat{\beta}$ denote the genome-phenome matrix of GWAS coefficients, then $\hat{\beta}\hat{\beta}^T$ should exhibit a **low-rank structure**.

How can we discover the latent pathways? Examine the phenonome!

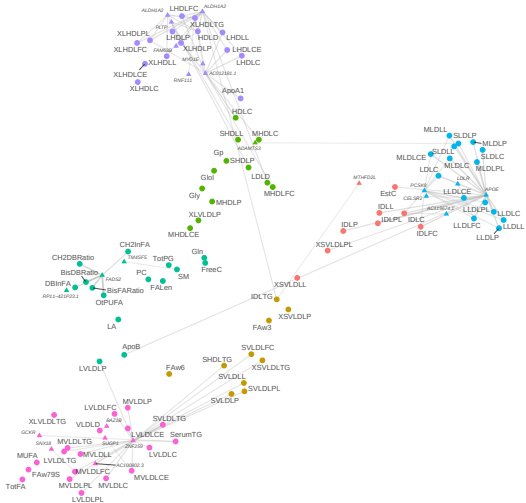


- If we let $\hat{\boldsymbol{\beta}}$ denote the genome-phenome matrix of GWAS coefficients, then $\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T$ should exhibit a **low-rank structure**.

Additional ideas (Ongoing work)

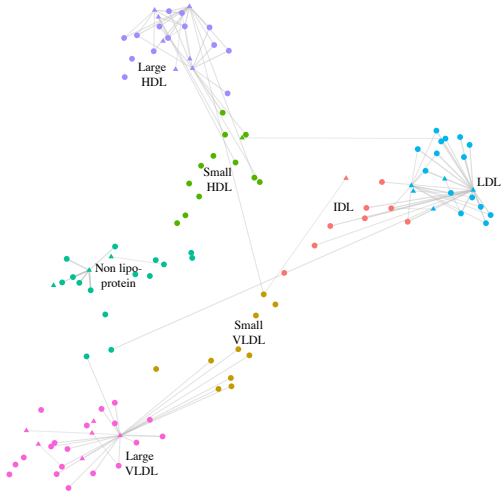
- To remove the environmental factors, contrast "signal" loci with "noise" loci.
- To stablize the results, use "bagging" (bootstrap aggregating).
- To visualise the results, use lower-dimensional embedding such as the UMAP.

Preliminary results: Metabolome GWAS



GWAS data: Kettunen et al. (2016) DOI:10.1038/ncomms11122.

Preliminary results: Metabolome GWAS



GWAS data: Kettunen et al. (2016) DOI:10.1038/ncomms11122.

Preliminary results: UK BioBank

b3 Cluster list

Cluster	Phenotype	Genotype	Cluster	Phenotype	Genotype
Fat-free	Body fat-free mass Standing height Basal metabolic rate	*GDF5* *ZBTB38* *ID4*	Fat	BMI Leg/arm/trunk fat mass	*FTO* *ADH1B* *AC09082*
Skin/hair color	Hair color Skin color Ease of skin tanning	*HERC2* *DEF8* *TPCN2*	Cardio-vascular	CHD MI High cholesterol Simvastatin	*PXDN* *SPOCK3*
Platelet	PLT count Circulatory diseases Heart rate	*JMJD1C* *CTC-454M9.1*	Diabetes	Diabetes Metformin Alcohol addiction	*TPTE2* *RP1-116*
Red cell	RBC MCH Genetic haemato-logical disorder	*HBS1L* *ODF3B* *SLC17A3*	Lymp-hocyte	Mono count Eos count ANC	*ITGA4* *CYP8B1* *GFI1*
Venous thrombo-embolism	PE DVT	*ABO* *SDK1*	Smoking	Ever smoked Never smoked	—
Bone mineral density	BMD T-score BMD QUI	*SACS* *FMN2*	Skin neoplasm	Malignant skin neoplasm	*PAX5* *FOXP1*

LMR×2

Summary

- Mendelian randomization provides genetic anchors to learn meaningful (and likely causal) representations of life.
- Many challenges remain:
  1. Pleiotropy;
  2. Non-linear structures and interactions;
  3. High-dimensionality;
  4. Low signal-to-noise ratio.

LMR×2

Summary

- Mendelian randomization provides genetic anchors to learn meaningful (and likely causal) representations of life.
- Many challenges remain:
  1. Pleiotropy;
  2. Non-linear structures and interactions;
  3. High-dimensionality;
  4. Low signal-to-noise ratio.

Collaborators in the works presented here

Jingshu Wang (Chicago); Dylan Small, Nancy Zhang (UPenn); Gibran Hemani, George Davey Smith (Bristol); Jack Bowden (Exeter); Daniel Iong, Yang Chen (Michigan); Zijun Gao, Trevor Hastie (Stanford).