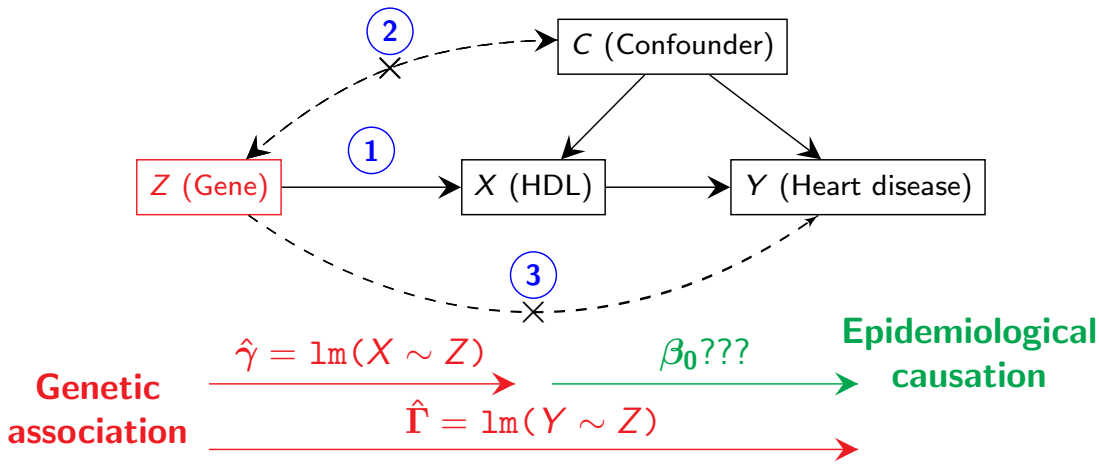
Mendelian randomization

Qingyuan Zhao

Department of Statistics, The Wharton School, University of Pennsylvania

July 12, 2019 @ Causal Inference Summer Institute



Outline

Outline

Hierarchy of evidence

When the goal is to infer causation...

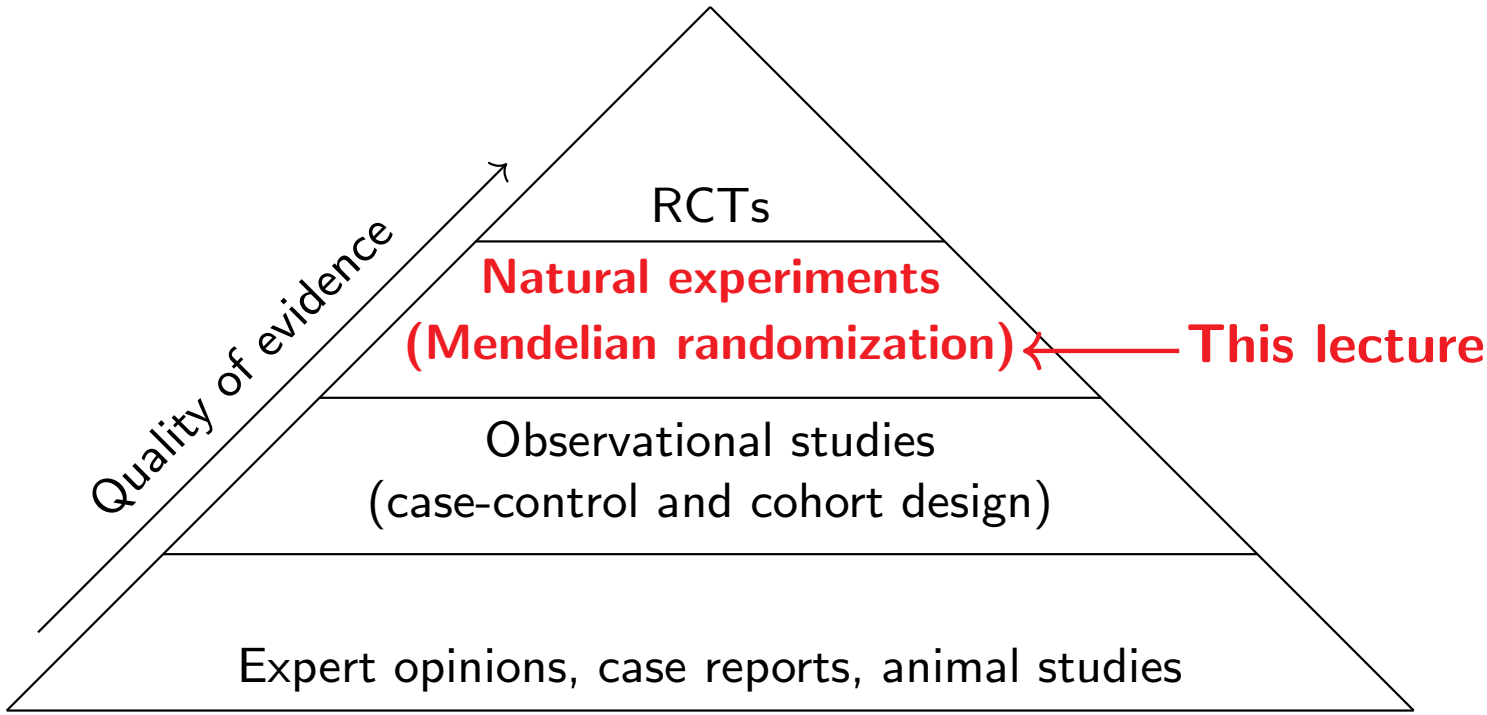


Figure: (A rough) Hierarchy of evidence in medical studies.[1]

[1] 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).

Fundamental challenge of observational studies

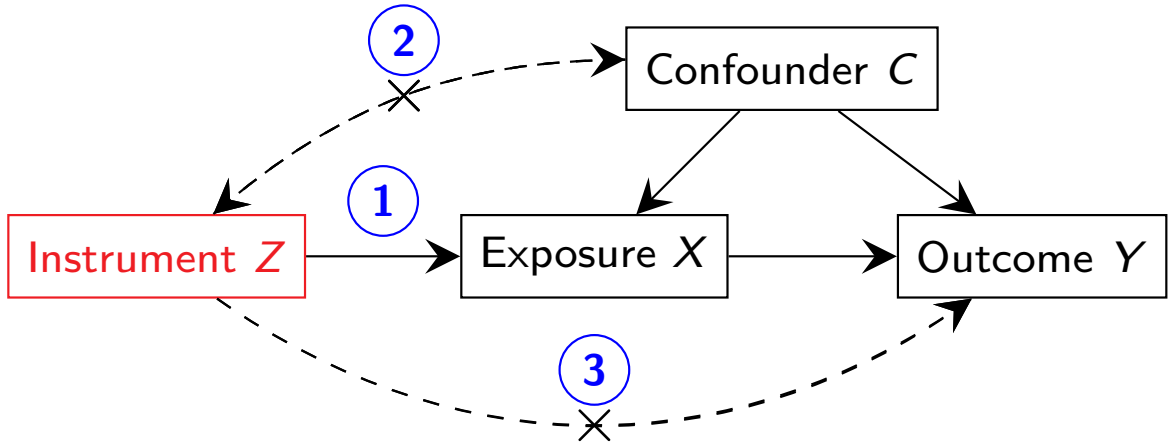
"Correlation does not imply causation".

Observational studies = Enumerating confounders

- Idea: Conditioning on possible sources of spurious correlation.
- Example: Possible confounders between smoking and lung cancer:
  - Age.
  - Sex.
  - Urban/Rural.
  - Working environment.
  - Socioeconomic class.
  - ...
- **Fundamental challenge: We can never be sure this list is complete.**
- The promise of instrumental variables: unbiased estimation of causal effect without enumerating confounders.

What is an instrument variable (IV)?

Causal diagram for IV



Core IV assumptions

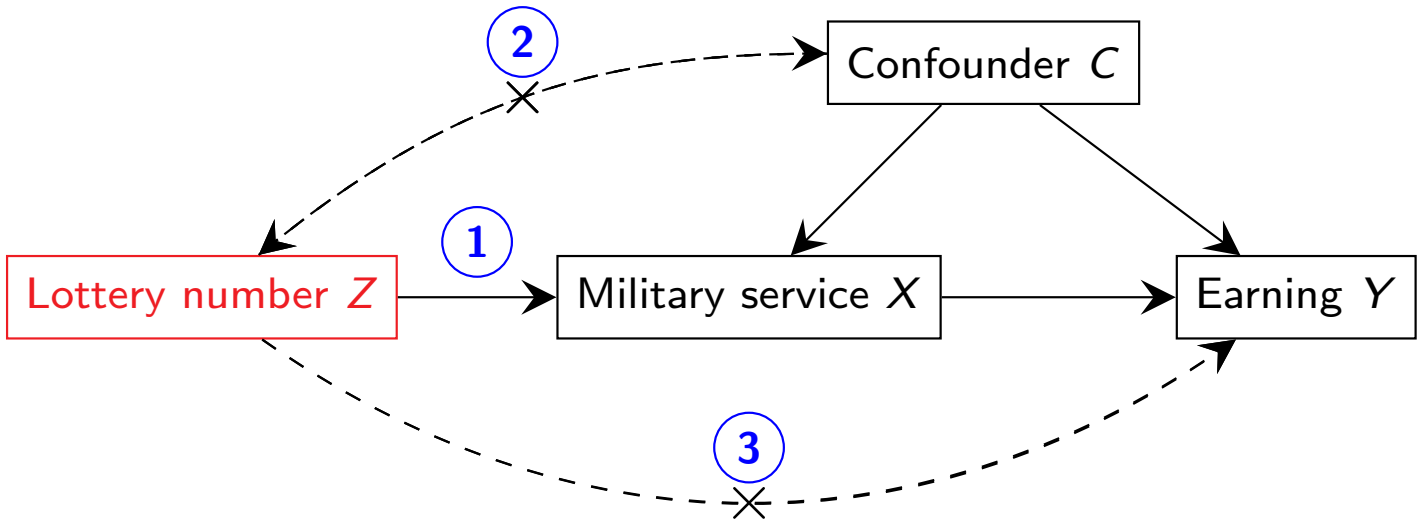
1. **Relevance**: Z is associated with the exposure (X).
2. **Effective random assignment**: Z is independent of the unmeasured confounder (C).
3. **Exclusion restriction**: Z cannot have any direct effect on the outcome (Y).

Wald's estimator based on Intention-to-treat (ITT) analysis

$$\text{Causal effect of } X \text{ on } Y \approx \frac{\text{ITT Effect of } Z \text{ on } Y}{\text{ITT Effect of } Z \text{ on } X}$$

Outline

IV in Economics: Effect of military service on earnings[2]



- In 1970, the U.S. government conducted draft lottery to determine priority of conscription for the Vietnam war.
- Exercise: Justify the core IV assumptions.
- The draft lottery can be regarded as a "natural experiment" of military service.

[2] Angrist, J. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *American Economic Review*, 80(3), 313–336.

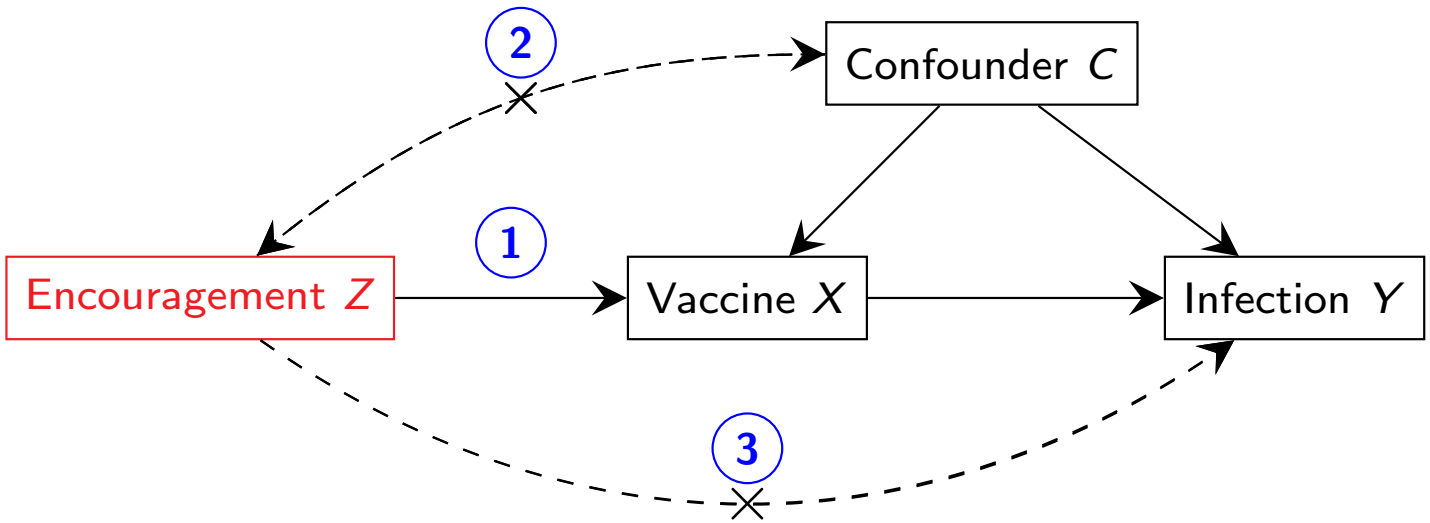
Results of the Vietnam-war lottery study

Table 4.1.3: Wald estimates of the effects of military service on the earnings of white men born in 1950

	Earnings		Veteran Status		Wald Estimate of Veteran Effect
Earnings year	Mean	Eligibility Effect	Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8	0.267	0.159	-2,741
		(210.5)		(0.040)	(1,324)
1971	3,338	-325.9			-2050
		(46.6)			(293)
1969	2,299	-2.0			
		(34.5)			

Notes: Adapted from Angrist (1990), Tables 2 and 3. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

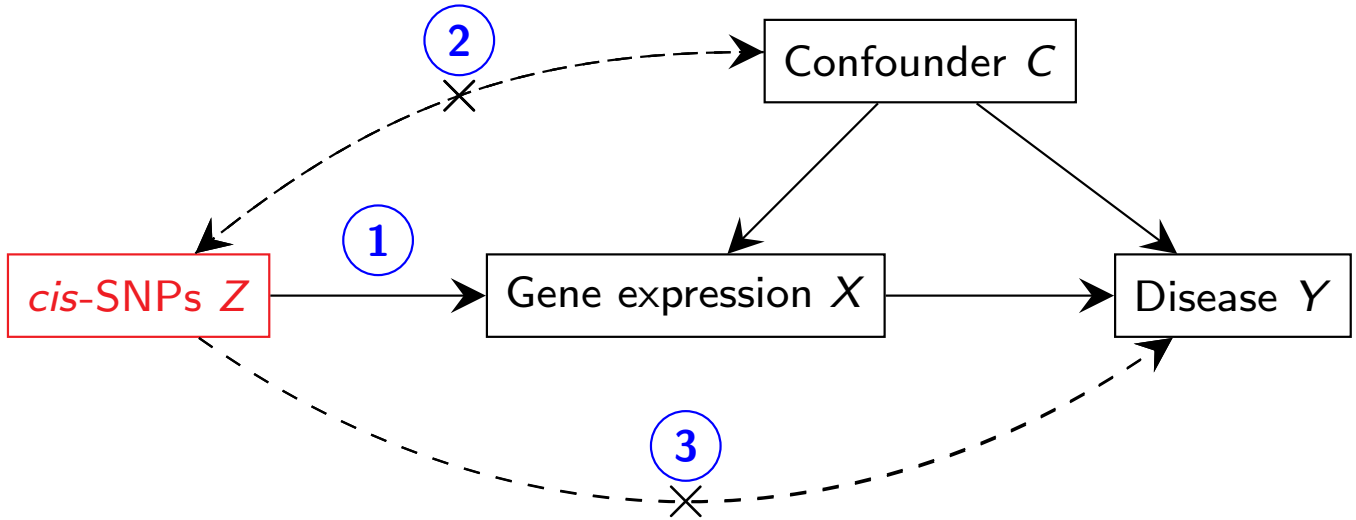
IV in Public Health: Effectiveness of vaccine[3]



- This is also called randomized encouragement design.
- The same idea can be applied to RCTs with non-compliance.

[3] Hirano, K. et al. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1), 69–88.

IV in Human Genetics: Gene testing[4]



- Compared to *trans*-SNPs, *cis*-SNPs are more likely to satisfy exclusion restriction (criterion 3).
- This is a special case of "Mendelian randomization" where genetic variation is used as IV and typically X is an epidemiological risk factor (more downstream).

[4] Gamazon, E. et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9).

Outline

Linear IV model

- The Wald ratio estimator becomes inadequate when $\boldsymbol{Z}$ and $\boldsymbol{X}$ are multivariate.
- The most commonly used IV estimators are based on the following linear model:

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\alpha} + U_i,$$
$$\boldsymbol{X}_i = \boldsymbol{Z}_i^T \boldsymbol{\gamma} + \boldsymbol{V}_i.$$

IV assumptions in the linear model

1. Relevance: $\boldsymbol{\gamma} \neq 0$;
2. Exogeneity: $\boldsymbol{Z}_i \perp\!\!\!\perp (U_i, \boldsymbol{V}_i)$;
3. Exclusion restriction: $\boldsymbol{\alpha} = \mathbf{0}$.

- The exposure variable $\boldsymbol{X}_i$ is called *confounded* or *endogenous* if it is correlated with U_i (or equivalently, if $\boldsymbol{V}_i$ is correlated with U_i).

Identification of causal effect

Under the linear IV model, the causal effect $\boldsymbol{\beta}$ satisfies

$$\mathbb{E}[\boldsymbol{Z}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})] = \mathbf{0}.$$

- Notice how this is different from the usual normal equation

$$\mathbb{E}[\boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})] = \mathbf{0}.$$

- To identify $\boldsymbol{\beta}$, we need $\dim(\boldsymbol{Z}_i) \geq \dim(\boldsymbol{X}_i)$.
- Just-identified case: When $\dim(\boldsymbol{Z}_i) = \dim(\boldsymbol{X}_i)$, we can estimate $\boldsymbol{\beta}$ by solving

$$\sum_{i=1}^{n} \boldsymbol{Z}_i(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}) = \mathbf{0}.$$

  The solution in matrix-form is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\boldsymbol{Y}.$$

- Over-identified case: When $\dim(\boldsymbol{Z}_i) > \dim(\boldsymbol{X}_i)$, we have some freedom to choose which (linear combinations of) equations to solve.

Two-stage least squares (TSLS)

- In the over-identified case, here is a general class of IV estimator: Let $\boldsymbol{f} : \mathbb{R}^{\dim(\boldsymbol{Z}_i)} \mapsto \mathbb{R}^{\dim(\boldsymbol{X}_i)}$ be any function that maps from the space of $\boldsymbol{Z}$ to $\boldsymbol{X}$. Then $\boldsymbol{\beta}$ satisfies

$$\mathbb{E}[\boldsymbol{f}(\boldsymbol{Z}_i) \cdot (Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})] = \mathbf{0}.$$

- The most efficient choice of $\boldsymbol{f}$ is $\boldsymbol{f}(\boldsymbol{Z}_i) = \mathbb{E}[\boldsymbol{X}_i | \boldsymbol{Z}_i] = \boldsymbol{Z}_i^T \boldsymbol{\gamma}$.
- The nuisance parameter $\boldsymbol{\gamma}$ is not known but can be estimated from the data. The most common estimator is least squares:

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{X}.$$

- Thus the IV estimator of $\boldsymbol{\beta}$ is given by (let $\boldsymbol{P_Z} = \boldsymbol{Z}(\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T$)

$$\hat{\boldsymbol{\beta}} = \left[(\boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T \boldsymbol{X}\right]^{-1} (\boldsymbol{Z}\hat{\boldsymbol{\gamma}})^T \boldsymbol{Y} = (\boldsymbol{X}^T \boldsymbol{P_Z} \boldsymbol{X})^{-1} (\boldsymbol{X}^T \boldsymbol{P_Z} \boldsymbol{Y}).$$

- This is called two-stage least squares, because (let $\hat{\boldsymbol{X}} = \boldsymbol{P_Z} \boldsymbol{X}$)

$$\hat{\boldsymbol{\beta}} = \texttt{lm}(\boldsymbol{Y} \sim \hat{\boldsymbol{X}}) = \texttt{lm}(\boldsymbol{Y} \sim \texttt{predict}(\texttt{lm}(\boldsymbol{X} \sim \boldsymbol{Z})))$$

- However, standard error of $\hat{\boldsymbol{\beta}}$ cannot be obtained directly from `lm` because $\hat{\boldsymbol{\gamma}}$ is estimated from the data.

Limited information maximum likelihood (LIML)

- Recall the linear IV model:

$$Y_i = \boldsymbol{X}_i^T \beta + U_i,$$
$$\boldsymbol{X}_i = \boldsymbol{Z}_i^T \gamma + \boldsymbol{V}_i.$$

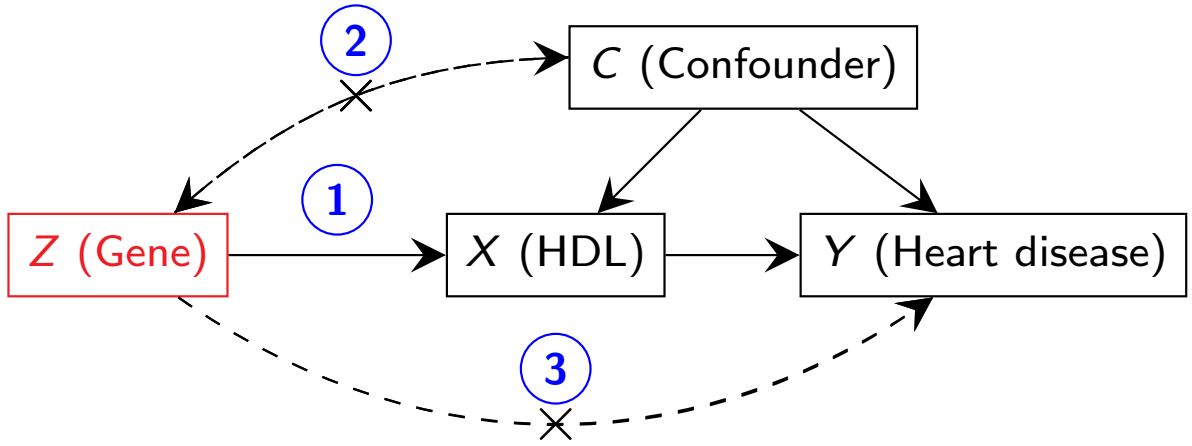
- The LIML estimator assumes the noise variables $(U_i, \boldsymbol{V}_i)$ are jointly normal with mean $\mathbf{0}$ and covariance Σ.
- LIML maximizes the log-likelihood of this problem:

$$l(\beta, \gamma, \Sigma) = -\frac{1}{2} \sum_{i=1}^{n} \log |\Sigma^{-1}| + \begin{pmatrix} Y_i - \boldsymbol{X}_i^T \beta \\ \boldsymbol{X}_i - \boldsymbol{Z}_i^T \gamma \end{pmatrix}^T \Sigma^{-1} \begin{pmatrix} Y_i - \boldsymbol{X}_i^T \beta \\ \boldsymbol{X}_i - \boldsymbol{Z}_i^T \gamma \end{pmatrix}.$$

- TSLS and LIML are asymptotically equivalent (when $n \to \infty$ and $\dim(\boldsymbol{X}_i)$ and $\dim(\boldsymbol{Z}_i)$ are fixed).
- LIML is more robust to weak instruments (small γ).

Outline

MR = Using genetic variation as IV



Examine the core IV assumptions

Criterion (1)	✓	Modern GWAS have identified many causal variants
Criterion (2)	✓	Almost Comes for free due to Mendel's Second Law Minor concern: population stratification
Criterion (3)	?	Problematic because of wide-spread **pleiotropy** (multiple functions of genes).

- Exercise: what if there are two SNPs in LD?

Non-conventional challenges in MR

Weak instruments Many genetic variants are only weakly associated with the exposure.

- Solution: Use LIML-type estimator instead of TSLS.

Two-sample IV/MR Association data for X and Y often come from different population.

- Need to justify the causal structure is invariant.[5]

Summary-data MR Most GWAS data come in summary-statistics format due to privacy.

- Solution: Develop statistical methods that can be applied to summary statistics.

Pleiotropy Exclusion restriction is likely violated for many genetic IVs.

- Solution: Use more robust methods that account for pleiotropy.

[5] Zhao, Q. et al. (2017). Two-sample instrumental variable analyses using heterogeneous samples. arXiv:1709.00081.

A general workflow of two-sample summary-data MR

1. Select independent IVs for the exposure (using **GWAS-E1**).
2. Extract GWAS summary statistics of the selected IVs for the **exposure** (using **GWAS-E2**).
3. Extract GWAS summary statistics of the selected IVs for the **outcome** (using **GWAS-O**).
4. **Harmonize** data in steps 2 and 3 so the reference allele is the same.
5. Perform statistical analysis.

Open-source software

A one-stop solution is being developed in the `R` package `TwoSampleMR`:[6]

- A large number of public GWAS summary datasets being collected.
- Convenient wrapper of LD clumping and data harmonization.
- Functions for statistical analysis.

Caveat

`TwoSampleMR` does not differentiate between **GWAS-E1** and **GWAS-E2**, which may introduce selection bias (also called winner's curse).

[6] `https://github.com/MRCIEU/TwoSampleMR`

Modeling assumptions for GWAS summary data

Dataset: estimated effects $(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\Gamma}})$ and standard errors $(\boldsymbol{\sigma}_X, \boldsymbol{\sigma}_Y)$.

Assumption 1: Measurement error model

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

Pre-processing warrants Assumption 1

- Large sample size $\Rightarrow$ CLT.
- (Approximate) independence due to
  1. Non-overlapping samples (in GWAS-E1, GWAS-E2, GWAS-O).
  2. Independent SNPs.

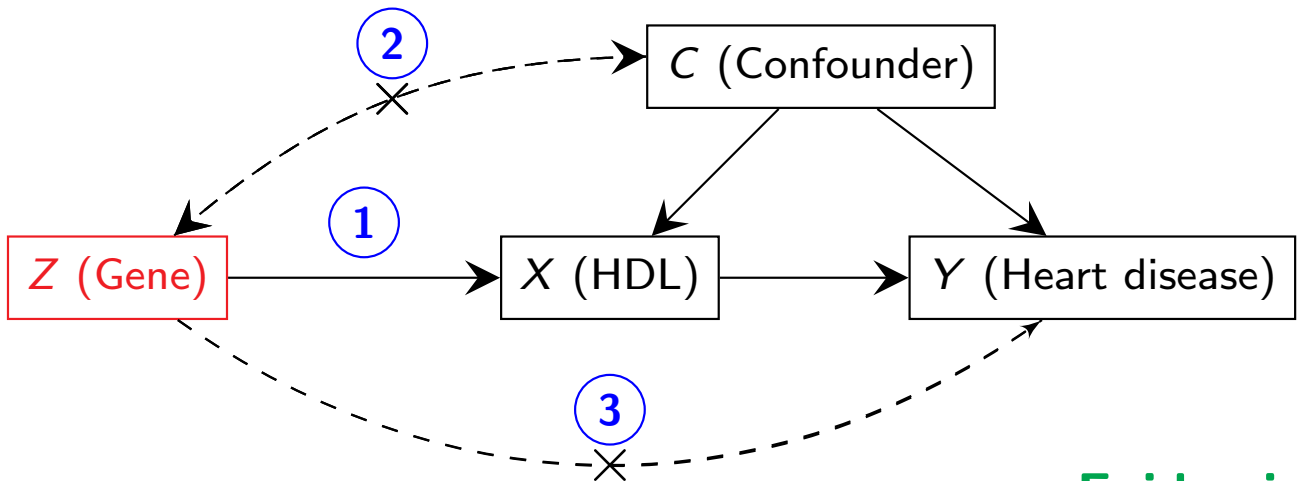
Assumption 2: Linking the genetic associations (ITT effects)

The causal effect β satisfies $\boldsymbol{\Gamma} \approx \beta\boldsymbol{\gamma}$. In particular, we have found a reasonable model for $\boldsymbol{\alpha} = \boldsymbol{\Gamma} - \beta\boldsymbol{\gamma}$ is universal pleiotropy with outliers

1. Most $\alpha_j \perp\!\!\!\perp \gamma_j$ and $\alpha_j \overset{i.i.d.}{\sim} N(0, \tau^2)$ for some small τ^2.
2. A few $|\alpha_j|$ might be very large.

Statistical problem

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

Meta-analysis methods

- Each SNP produces an independent Wald estimator: $\hat{\beta}_j = \hat{\Gamma}_j / \hat{\gamma}_j$.
- Using Delta method and assuming $\Gamma_j \equiv \beta_0 \gamma_j$, we can obtain

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} = \frac{\Gamma_j + \epsilon_{Yj}}{\gamma_j + \epsilon_{Xj}} \approx \mathrm{N}\left(\frac{\Gamma_j}{\gamma_j}, \frac{\sigma^2_{Xj} + \beta^2 \sigma^2_{Yj}}{\gamma_j^2} := \sigma_j^2 \right).$$

- Next combine the individual estimates using meta-analysis:
  1. Inverse-variance weighting:

$$\hat{\beta}_{\mathrm{IVW}} = \texttt{Mean}(\hat{\beta}_j, \texttt{weights} = 1/\sigma_j^2).$$

  2. Weighted median:

$$\hat{\beta}_{\mathrm{WMED}} = \texttt{Median}(\hat{\beta}_j, \texttt{weights} = 1/\sigma_j^2).$$

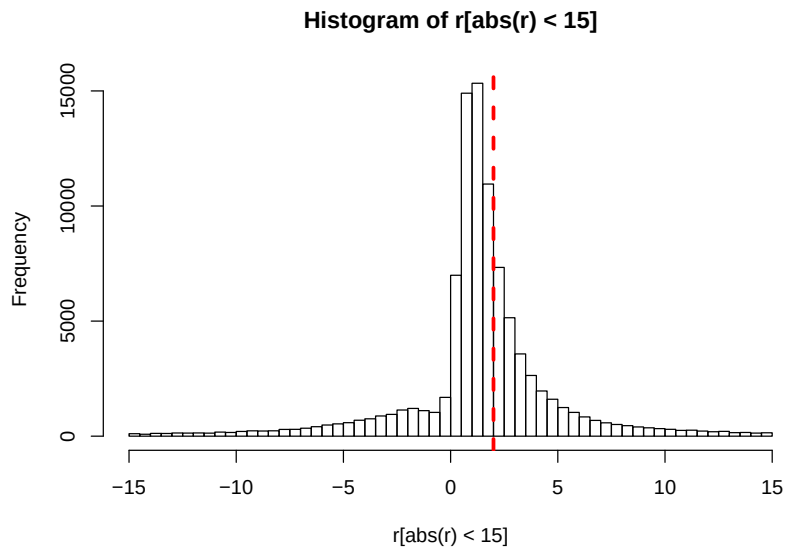
  3. MR-Egger regression:

$$\hat{\beta}_{\mathrm{Egger}} = \texttt{lm}(\hat{\Gamma}_j \sim \hat{\gamma}_j, \texttt{weights} = 1/\sigma_j^2).$$

Weak instrument bias

- The main issue of meta-analysis methods is that the Delta method approximation is not accurate if $|\gamma_j|/\sigma_{Xj}$ is small.
- In general, the distribution of $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j$ is a mixture of Cauchy distribution and a bi-modal distribution.[7]

```
> p <- 100000; r <- (2 + rnorm(p))/(1 + rnorm(p));
> hist(r[abs(r) < 15], 100); abline(v = 2, lty = "dashed", col = "red", lwd = 3);
> median(r)
[1] 1.327066
```



- This problem is known as weak instrument bias in the IV literature.

[7] Marsaglia, G. (2006). Ratios of normal variables. *Journal of Statistical Software*, 16(4), 1–10.

Likelihood methods[8]

- Using the ratio $\hat{\Gamma}_j/\hat{\gamma}_j$ is a little silly if $\hat{\gamma}_j$ is small.
- We can pool information from multiple weak IVs using the likelihood.
- Assuming $\Gamma_j \equiv \beta\gamma_j$, the log-likelihood of SNP j is

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

- 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

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

- Observation 1: γ_j **only appears as weight** to "residual" $\hat{\Gamma}_j - \beta\hat{\gamma}_j$.
- Observation 2: $\hat{\gamma}_{j,\mathrm{MLE}}(\beta)$ **is independent of** $\hat{\Gamma}_j - \beta\hat{\gamma}_j$.

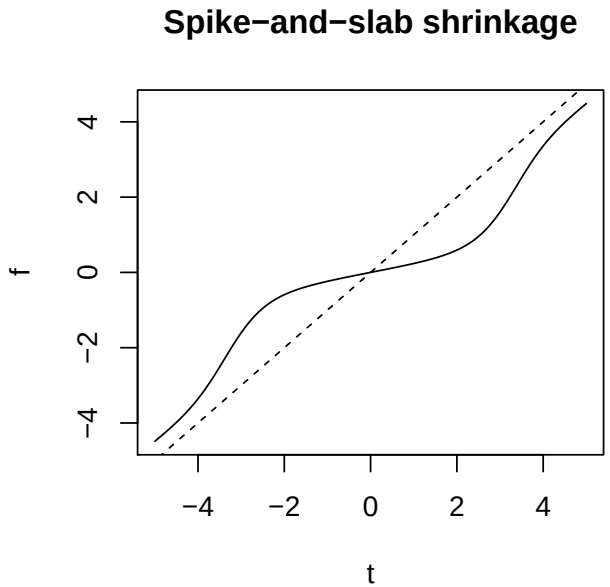
[8] Based on Lindsay, B. (1985). "Using empirical partially Bayes inference for increased efficiency". *Annals of Statistics*, 13, and Zhao, Q. et al. (2018). arXiv:1804.07371 (to appear in *IJE*).

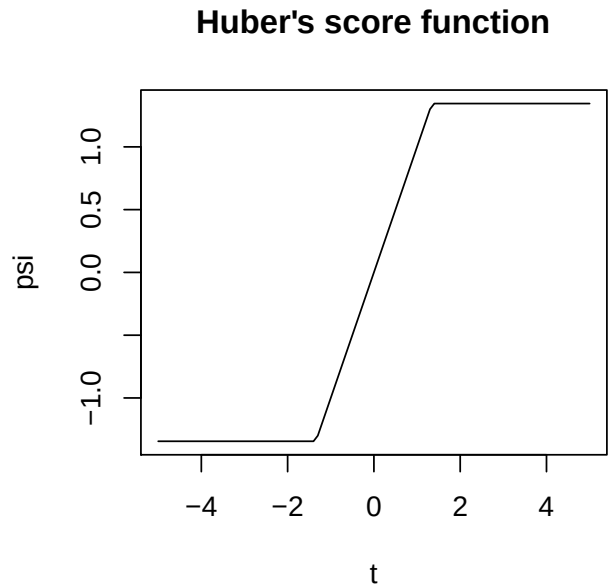
Increased efficiency and robustness

- The observations above motivate a general class of unbiased estimating equations:

$$\sum_{j=1}^{p} f_j(\hat{\gamma}_{j,\mathrm{MLE}}(\beta)) \cdot \psi\left(\frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{\sigma^2_{Yj} + \beta^2\sigma^2_{Xj}}}\right) = 0$$

- Heuristic: choose f_j to increase efficiency; choose bounded ψ to be robust against outliers.
- Example: $f_j(\hat{\gamma}_{j,\mathrm{MLE}})$ is the spike-and-slab shrinkage estimate of γ_j.
- Example: ψ is Huber's score function.



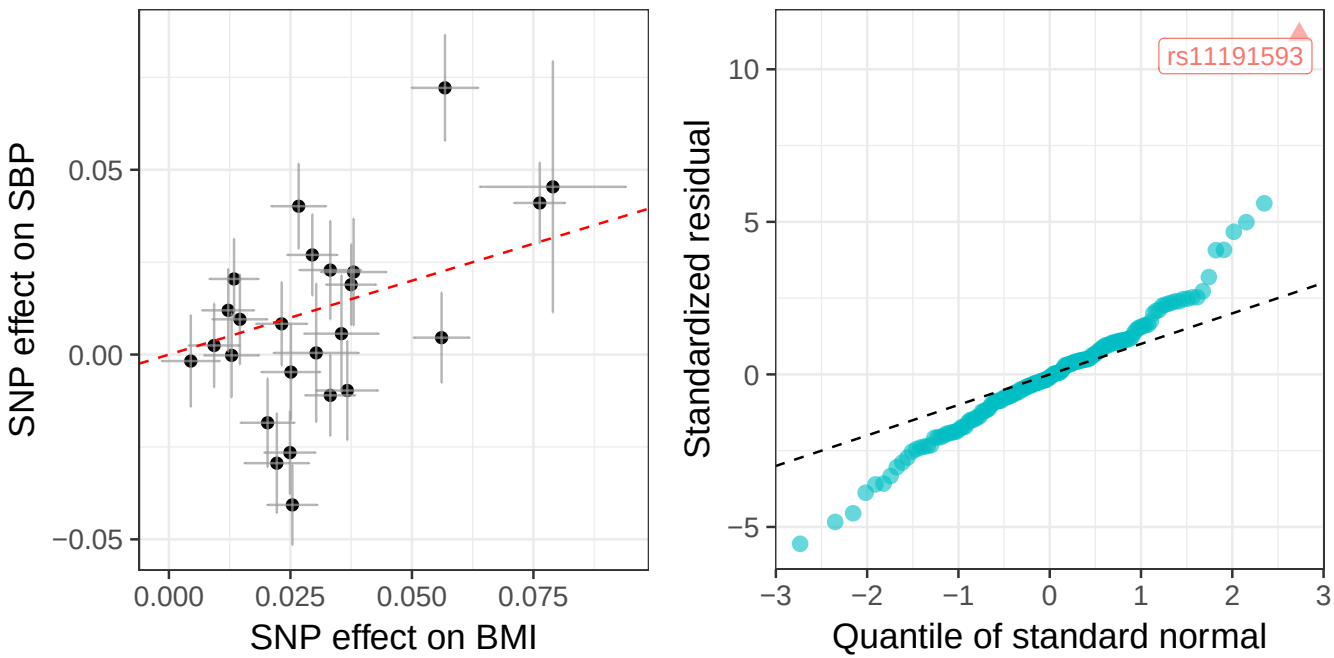


An overdispersion phenomenon

- So far we have assumed $\Gamma_j = \beta\gamma_j$ (at least for most j).
- However we have found that $\alpha_j = \Gamma_j - \hat{\beta}\gamma_j$ seems to be approximately normal when $\hat{\beta}$ is obtained as above.

A real data example: Effect of BMI on SBP

- 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 residuals.



Adjusting the score of overdispersion parameter

- A reasonable model is most $\alpha_j \sim \mathrm{N}(0, \tau^2)$.
- Statistical estimation of τ^2 is non-trivial due to the Neyman-Scott phenomenon.

Neyman-Scott problem (a simplified scenario)

Suppose we observe independent pairs

$$X_{ij} \sim \mathrm{N}(\gamma_i, \tau^2), \ i = 1, \ldots, n, \ j = 1, 2.$$

The goal is to estimate τ^2, but the MLE is inconsistent:

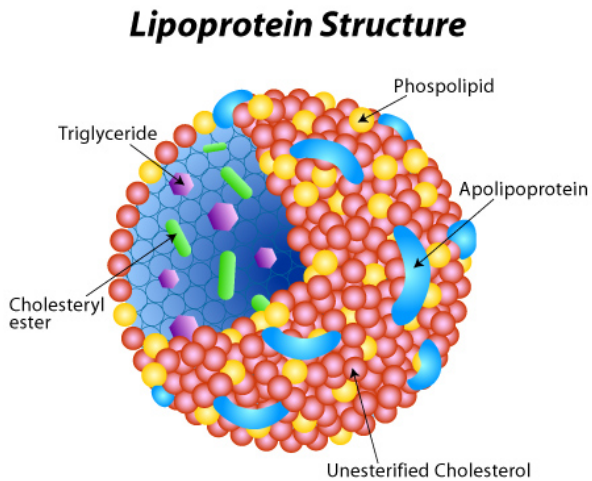
$$\hat{\tau}^2 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} (X_{ij} - \bar{X}_{i\cdot})^2 = \frac{1}{4n} \sum_{i=1}^{n} (X_{i1} - X_{i2})^2 \xrightarrow{p} \tau^2/2.$$

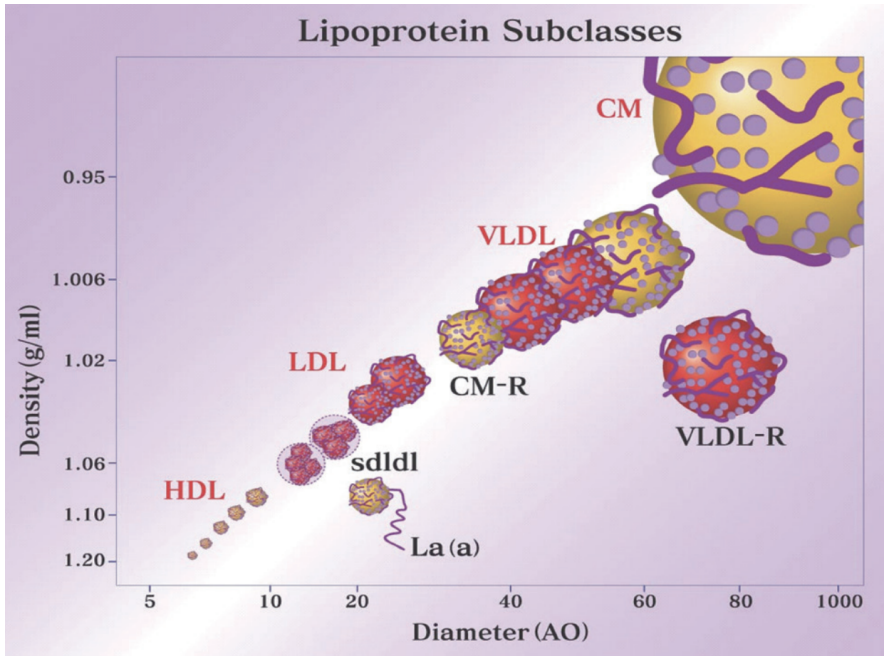
An easy fix in this case is to use $2\hat{\tau}^2$. However the inconsistency of MLE is common in many other problems involving a large number nuisance parameters, and the fix is usually complicated.

- There is a relatively simple fix in the MR problem (details omitted).

Outline

Some background about blood lipids





Left: Lipoprotein particles transport fat molecules in our body.[9]

Right: They can be categorized based on density and size.[10]

[9] `https://www.labce.com/spg659279_lipoprotein_particles.aspx`.

[10] Nakajima, K. "Remnant Lipoproteins: A Subfraction of Plasma Triglyceride-Rich Lipoproteins Associated with Postprandial Hyperlipidemia." *Clinical & Experimental Thrombosis and Hemostasis* 1.2 (2014): 45-53.

Blood test for lipid profile

Traditional lipid traits

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
					01
Lipids					01
Cholesterol, Total	**210**	**High**	mg/dL	100-199	01
Triglycerides	**236**	**High**	mg/dL	0-149	01
HDL Cholesterol	**36**	**Low**	mg/dL	>39	01
According to ATP-III Guidelines, HDL-C >59 mg/dL is considered a negative risk factor for CHD.					
LDL Cholesterol Calc	**127**	**High**	mg/dL	0-99	01
Comment					01
If initial LDL-cholesterol result is >100 mg/dL, assess for risk factors.					
T. Chol/HDL Ratio	**5.8**	**High**	ratio units	0.0-5.0	01
Estimated CHD Risk	**1.2**	**High**	times avg.	0.0-1.0	01

Three main traits: **LDL-C** ("bad" cholesterol), **HDL-C** ("good" cholesterol), **total triglycerides**.

Advanced lipoprotein testing

A "new" technology—Nuclear Magnetic Resonance (NMR)—can now measure **subclass traits** such as

S-LDL-P Concentration of small LDL particles.

M-HDL-C Total cholesterol in medium HDL particles.

Blood test for lipid profile

Traditional lipid traits

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
					01
Lipids					01
Cholesterol, Total	**210**	**High**	mg/dL	100-199	01
Triglycerides	**236**	**High**	mg/dL	0-149	01
HDL Cholesterol	**36**	**Low**	mg/dL	>39	01
According to ATP-III Guidelines, HDL-C >59 mg/dL is considered a negative risk factor for CHD.					
LDL Cholesterol Calc	**127**	**High**	mg/dL	0-99	01
Comment					01
If initial LDL-cholesterol result is >100 mg/dL, assess for risk factors.					
T. Chol/HDL Ratio	**5.8**	**High**	ratio units	0.0-5.0	01
Estimated CHD Risk	**1.2**	**High**	times avg.	0.0-1.0	01

Three main traits: **LDL-C** ("bad" cholesterol), **HDL-C** ("good" cholesterol), **total triglycerides**.

Advanced lipoprotein testing

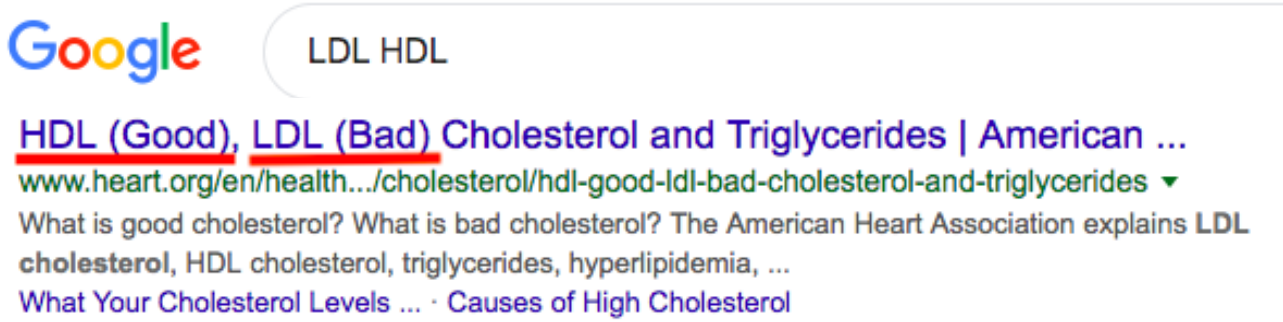
A "new" technology—Nuclear Magnetic Resonance (NMR)—can now measure **subclass traits** such as

S-LDL-P Concentration of small LDL particles.

M-HDL-C Total cholesterol in medium HDL particles.

Lipid hypothesis

Conventional wisdom (from observational studies)



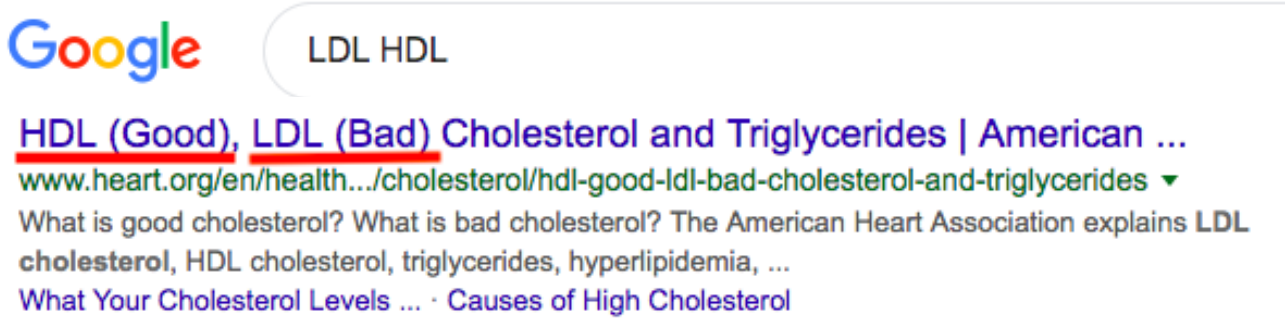
State-of-the-art perspective

In a plenary award lecture in ASHG 2018[11], I saw this slide:

[11] http://www.ashg.org/2018meeting/listing/NumberedSessions.shtml#sess3

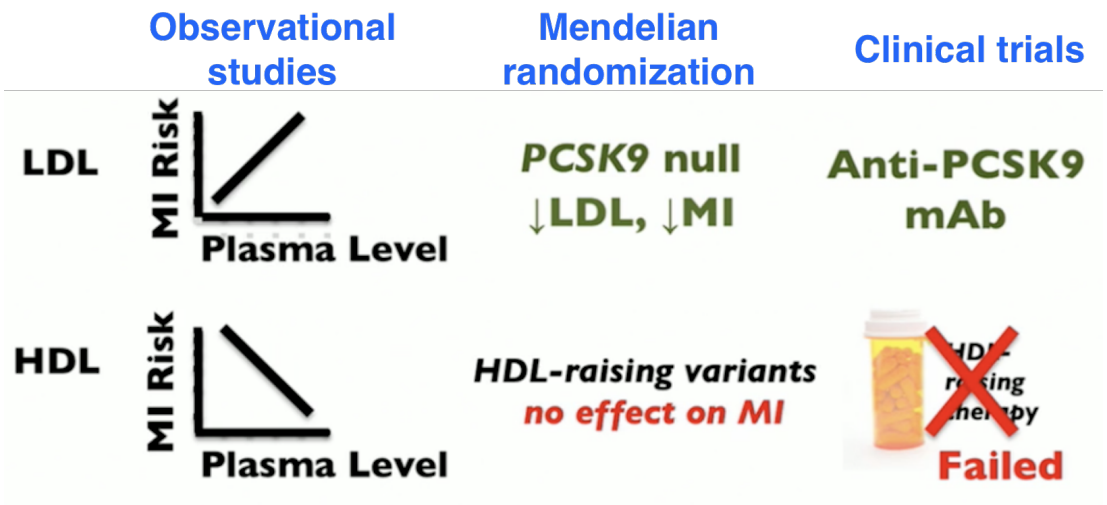
Lipid hypothesis

Conventional wisdom (from observational studies)



State-of-the-art perspective

In a plenary award lecture in ASHG 2018[11], I saw this slide:

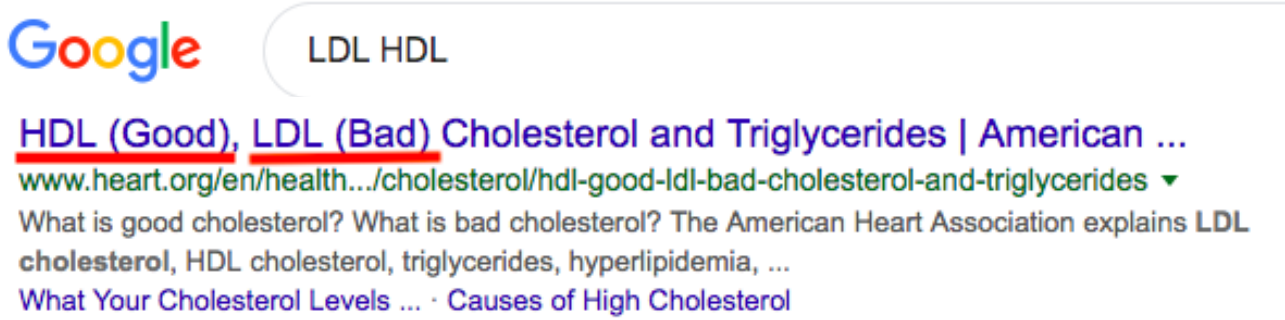


(MI = Heart attack)

[11] http://www.ashg.org/2018meeting/listing/NumberedSessions.shtml#sess3

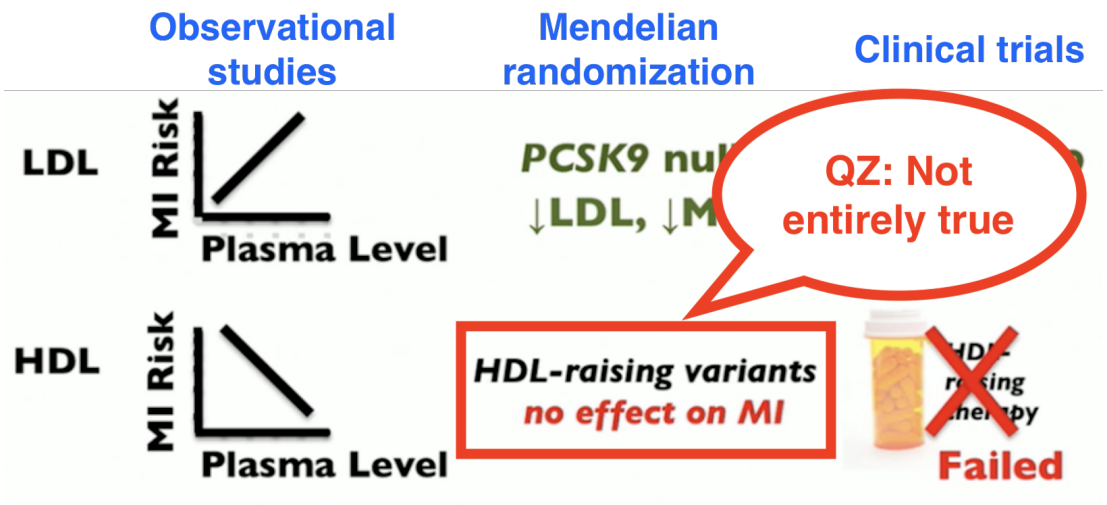
Lipid hypothesis

Conventional wisdom (from observational studies)



State-of-the-art perspective

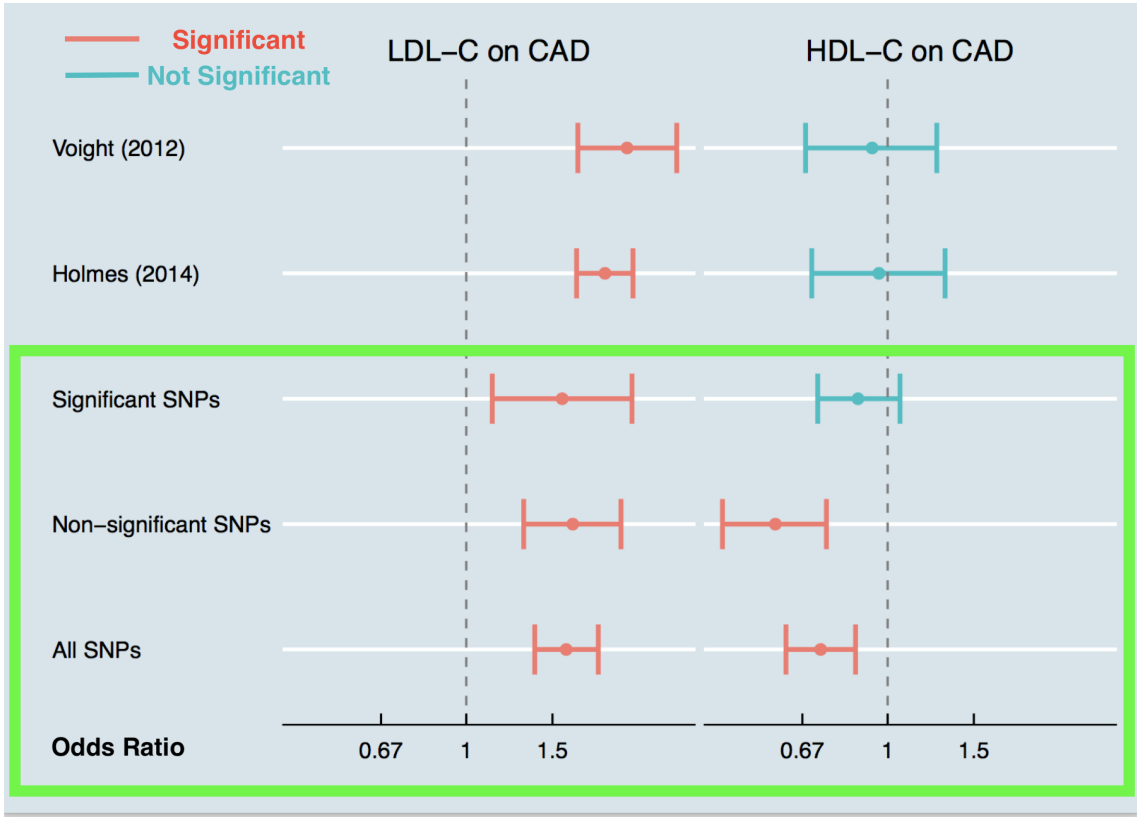
In a plenary award lecture in ASHG 2018[11], I saw this slide:



(MI = Heart attack)

[11] http://www.ashg.org/2018meeting/listing/NumberedSessions.shtml#sess3

New MR results for LDL-C and HDL-C[12]

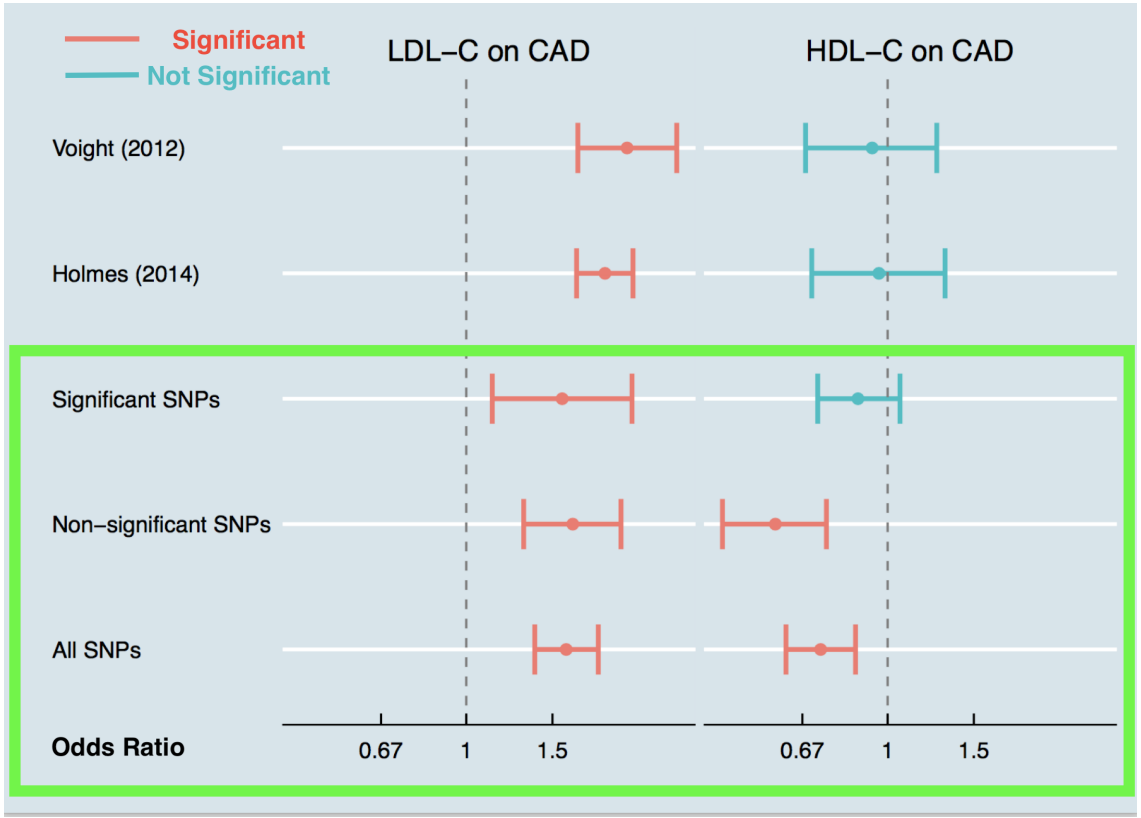


- For HDL-C, log odds ratio $\hat{\beta} = -0.245$ $(\mathbf{SE = 0.035})$ is highly significant.
- **It is too hasty to dismiss the HDL hypothesis!**

However, strong and weak IVs don't agree on HDL-C.
(Unlike BMI and LDL-C)

[12] Zhao, Q. et al. (2018). arXiv:1804.07371 (to appear in *IJE*)

New MR results for LDL-C and HDL-C[12]



- For HDL-C, log odds ratio $\hat{\beta} = -0.245$ (SE $= 0.035$) is highly significant.
- It is too hasty to dismiss the HDL hypothesis!

However, strong and weak IVs don't agree on HDL-C.
(Unlike BMI and LDL-C)

[12] Zhao, Q. et al. (2018). arXiv:1804.07371 (to appear in *IJE*)

A more formal (falsification) test

Under our assumptions

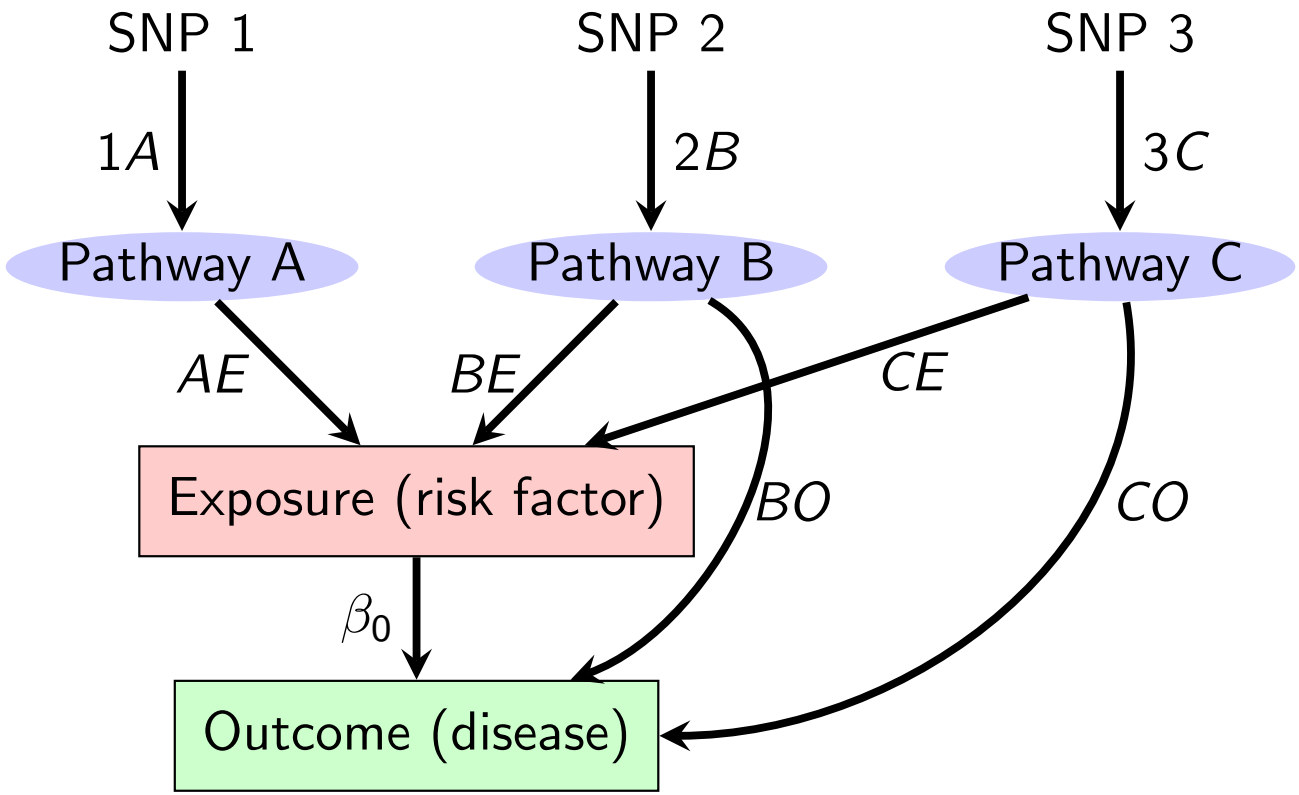
$$\frac{\hat{\Gamma}_j - \beta_0 \hat{\gamma}_j}{\sqrt{\sigma^2_{Yj} + \tau^2_0 + \beta^2_0 \sigma^2_{Xj}}} \,\Big|\, \hat{\gamma}_{j,\text{MLE}}(\beta_0, \tau^2_0) \dot{\sim} N(0,1).$$

- This can be empirically tested (e.g. using regression splines).

Example: Effect of HDL cholesterol on CAD

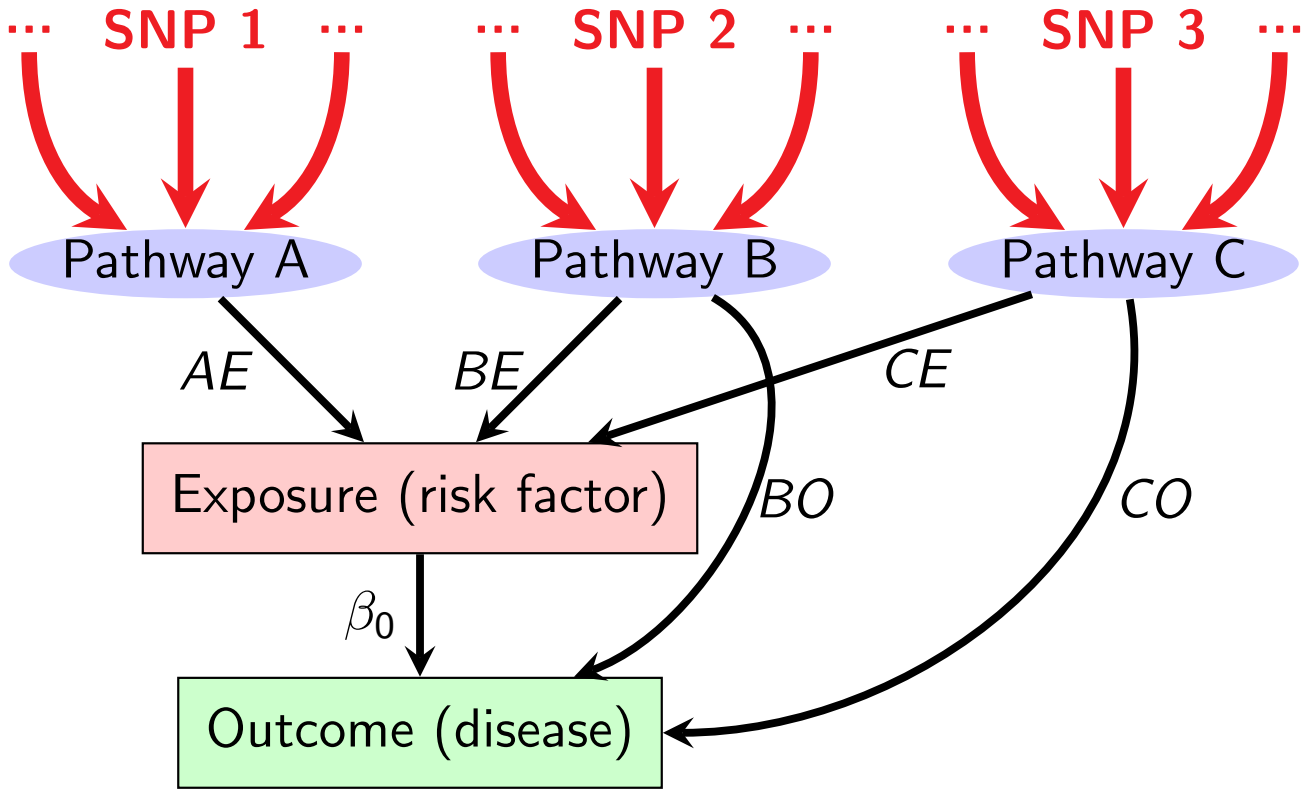
The diagnostic plot shows **evidence of heterogeneity**.

Multiple genetic pathways $\Rightarrow$ Multiple modes of β



	Exposure effect γ	Outcome effect Γ	**Ratio**
SNP 1	$1A \cdot AE$	$1A \cdot AE \cdot \beta_0$	β_0
SNP 2	$2B \cdot BE$	$2B \cdot BE \cdot \beta_0 + 2B \cdot BO$	$\beta_0 + (BO/BE)$
SNP 3	$3C \cdot CE$	$3C \cdot CE \cdot \beta_0 + 3C \cdot CO$	$\beta_0 + (CO/CE)$

Multiple genetic pathways $\Rightarrow$ Multiple modes of β



	Exposure effect γ	Outcome effect Γ	**Ratio**
SNP 1	$1A \cdot AE$	$1A \cdot AE \cdot \beta_0$	$\boldsymbol{\beta_0}$
SNP 2	$2B \cdot BE$	$2B \cdot BE \cdot \beta_0 + 2B \cdot BO$	$\boldsymbol{\beta_0 + (BO/BE)}$
SNP 3	$3C \cdot CE$	$3C \cdot CE \cdot \beta_0 + 3C \cdot CO$	$\boldsymbol{\beta_0 + (CO/CE)}$

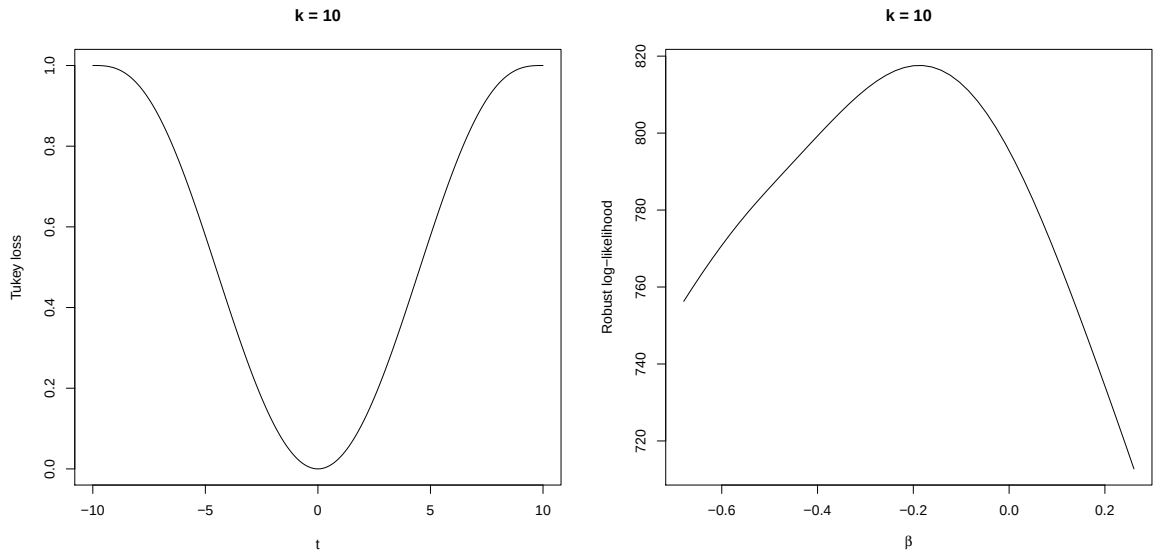
Detection via modal plot

- $l(\beta) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\sigma^2_{Yj} + \beta^2\sigma^2_{Xj}}$ penalizes too much on "outliers".
- We can plot "robust" log-likelihood and search for **multiple modes**:

$$l_\rho(\beta) = -\sum_{j=1}^{p} \rho\left(\frac{\hat{\Gamma}_j - \beta\hat{\gamma}_j}{\sqrt{\sigma^2_{Yj} + \beta^2\sigma^2_{Xj}}}\right).$$

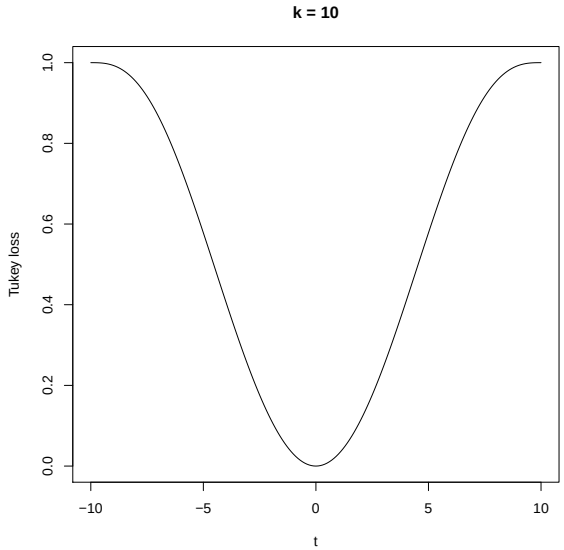
Example: Effect of HDL cholesterol on CAD

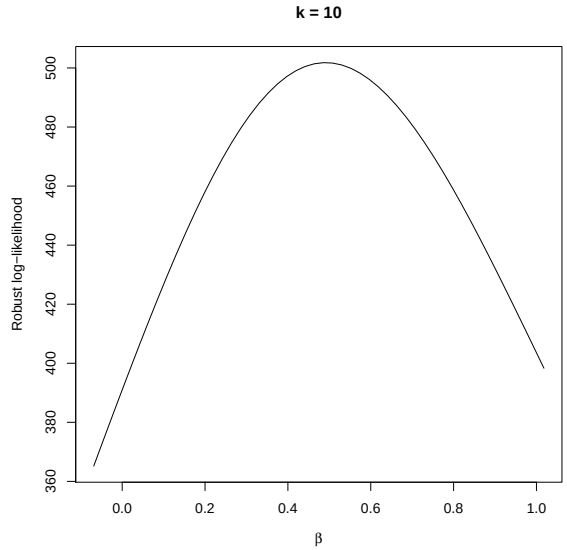
Left: loss function ρ; Right: robust log-likelihood.



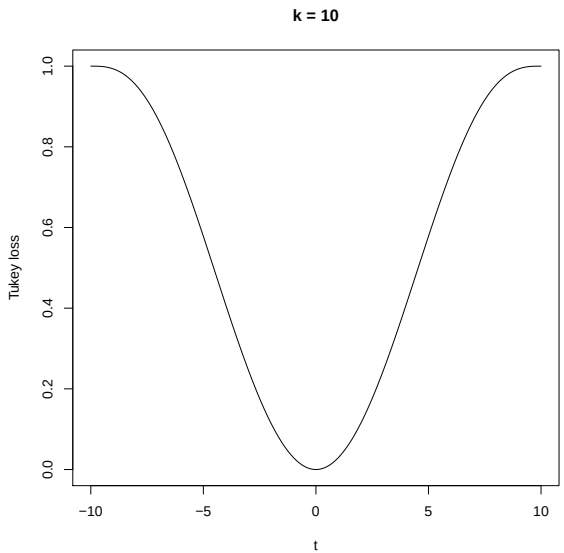
Compare with the modal plot for LDL-C

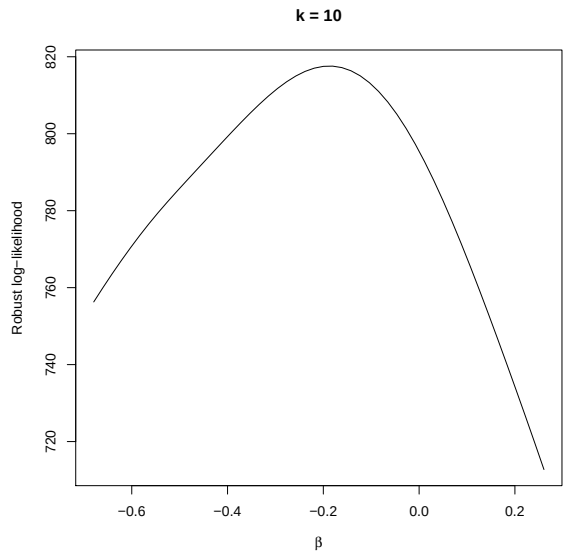
LDL-C





HDL-C





Mechanistic heterogeneity

This phenomenon can occur **even if all the IVs are valid**.

Heuristic: Local/Complier average treatment effect

- Under the monotonicity assumption

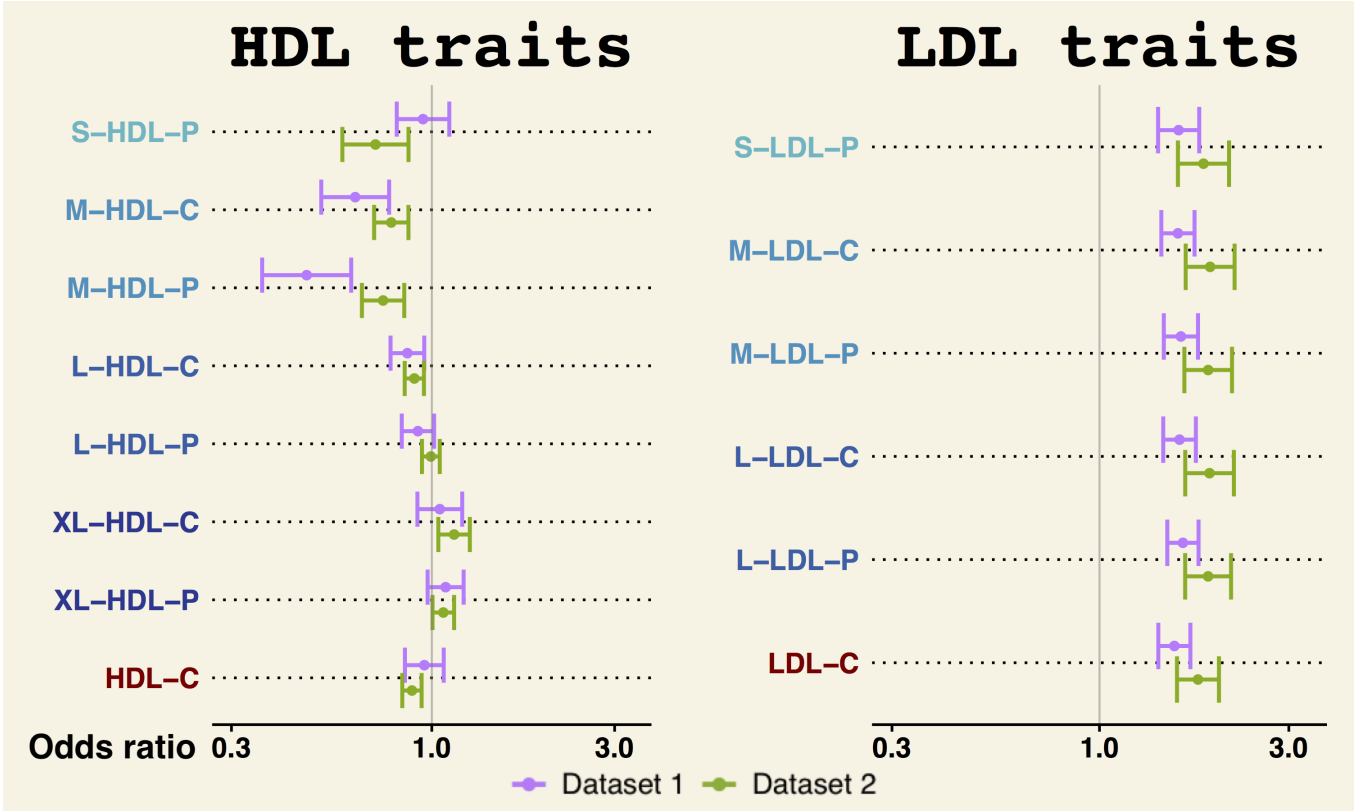
$$\mathbb{P}\big(X(z=1) \geq X(z=0)\big) = 1,$$

  Angrist, Imbens, and Rubin (1996) showed that

$$\beta_0 = \mathbb{E}\big[Y(x=1) - Y(x=0) \,|\, X(z=1) > X(z=0)\big].$$

- Multiple modes can occur if the genetic instruments affect X in multiple pathways, thus correspond to different "complier groups" $\{X(z=1) > X(z=0)\}$.

MR screening for lipoprotein subclasses



Scientific Takeaway

- The new results show **clear heterogeneity among HDL subclasses**, confirming our observations above.
- **The actual causal agent may be medium (and possibly small) HDL particles**, which may be connected to the "HDL function hypothesis" being developed currently.

Final messages

IV methods and Mendelian randomization are very useful tools to infer causality and are becoming widely used in epidemiology and human genetics.

I hope this is only the beginning of your journey with IVs and MR. Here are some suggested readings:

- Chapter 4 of Angrist and Pischke's book *Mostly Harmless Econometrics: An Empiricist's Companion*.
- A tutorial of IV methods for biostatisticians in *Statistics in Medicine* (2014) by Baiocchi, Cheng and Small.
- A special lecture on MR in *International Journal of Epidemiology* (2003) by Davey Smith and Ebrahim.
- Webpage for my research on IV and MR: `http://www-stat.wharton.upenn.edu/~qyzhao/MR.html`.