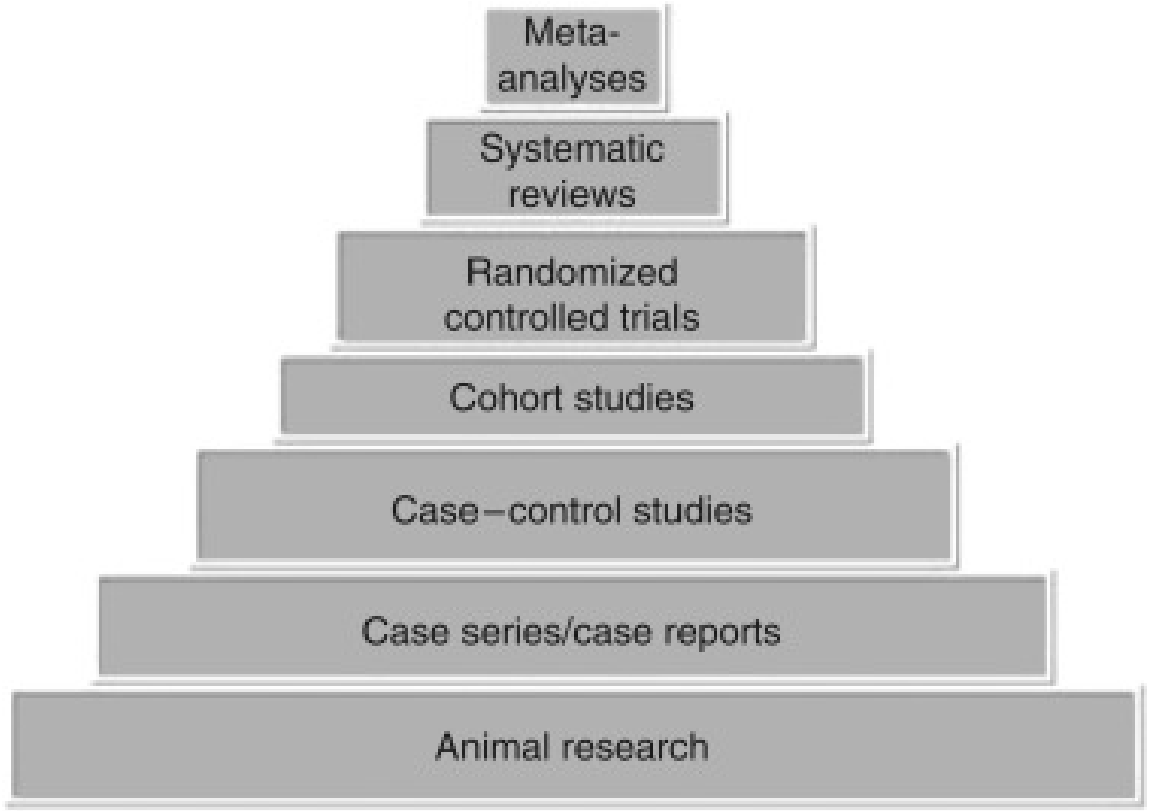
Mendelian Randomization Studies: Nature's Randomized Trials

CHIA-NI HSIUNG

DATA SCIENCE STATISTICAL COOPERATION CENTER
ACADEMIA SINICA

hierarchy of study design

Strength of Evidence



Tamar Nijsten & Robert S. Stern 2012

In observational epidemiology study

We want to find

- What is the causal effect
- the perfect approach to assess causation

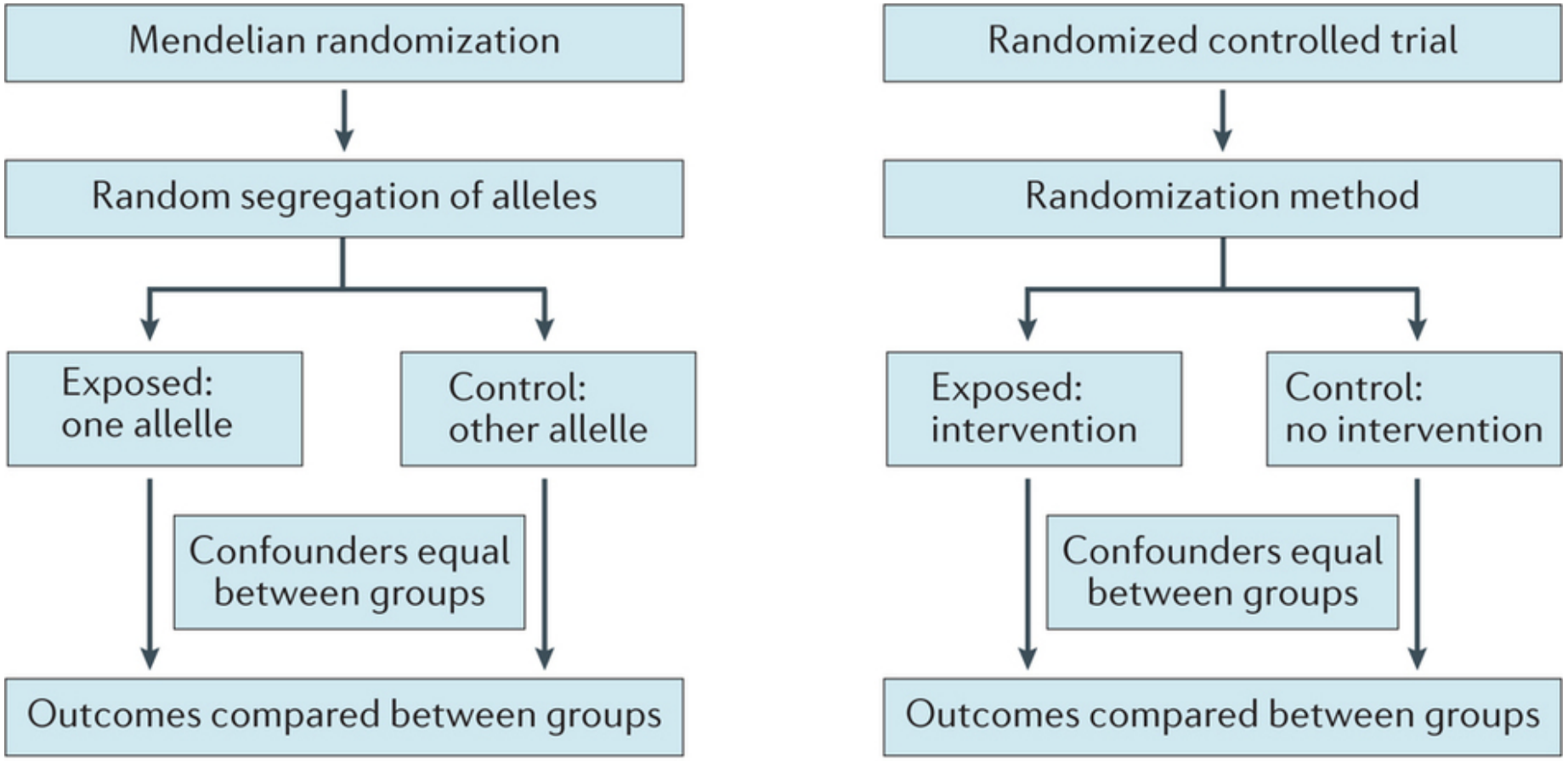
How to achieve this goal

- Randomize control trial
  - it is often not ethical or possible to carry out RCTs

What is Mendelian randomization

- Fundamental idea is that the genotypes are randomly assigned (Medel's Law)
- Mendelian randomization (MR) is a statistical technique that uses genetic variants as instrumental variables to investigate the causal relationship between an exposure and an outcome.
- simulate the randomized controlled trial in observational research.
- Approach to test for a casual effect from observational data in the presence of certain confounding factors
- Katan MB proposed this idea in genetics study in 1986

Comparison of the design of a Mendelian randomization study and a randomized controlled trial.

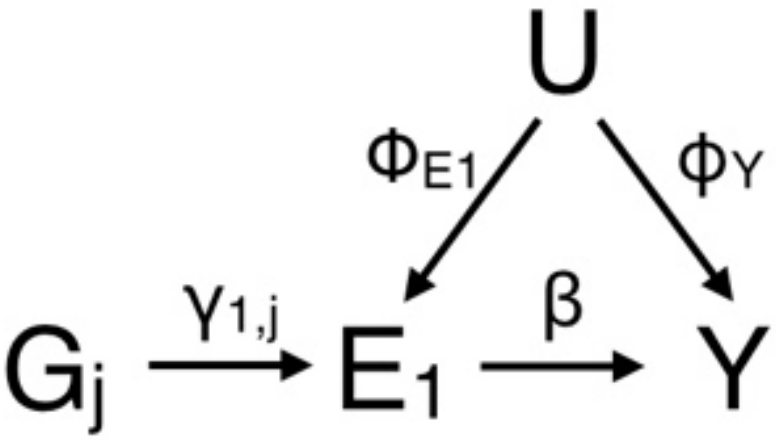


Nature Reviews | Rheumatology

Mendelian randomization (MR)

Use **SNPs (G_j)** as instrumental variables to obtain causal effect of **exposure (E)** on the **outcome (Y)**

Figure 1. Causal DAG for standard MR analysis



Estimating causal effect of the exposure on the outcome (β)

Step 1. Estimate association between G and E (γ)

$$E = \gamma_0 + \gamma_j G_j + \varepsilon_{Ej}$$

Step 2. Estimate association between G and Y (δ)

$$Y = \delta_0 + \delta_j G_j + \varepsilon_{Yj}$$

Step 3. Estimate causal effect of E on Y (β)

$$\hat{\beta}_j = \frac{\hat{\delta}_j}{\hat{\gamma}_j}$$

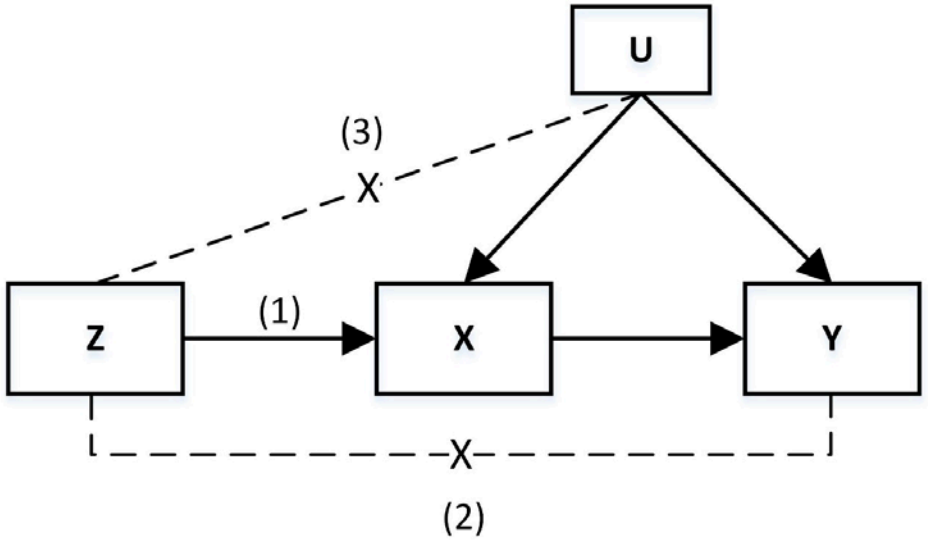
Instrument/Instrumental variable (IV)

- A variable used to control for confounding
  - Widely used in econometrics and social science research and now increasingly used in epidemiological studies
- It is a variable associated with the treatment (or exposure). In other words, it affects whether or not the treatment is received.
- It affects the outcome only through the treatment and it is independent of confounders.
- The randomization assignment in randomized controlled trials (RCT) is an example of an ideal instrument.
- Using IV identifies the causal average effect of the treatment on the outcome independent of the unobserved sources of variability.

Instrument strength

- F statistics A measure of instrument strength and can be used to judge the extent of weak instrument bias
- F statistics > 10, strong instrument

(Lawlor et al. 2008)



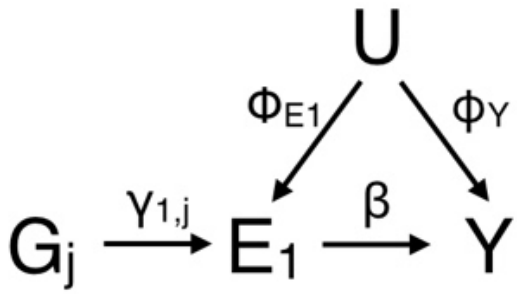
Assumptions required in MR

1. The genetic marker is associated with the exposure $\boldsymbol{\gamma_j \neq 0}$
2. The genetic marker is independent of all confounders of the exposure-outcome relationship (U)

   No effect from G_j to U

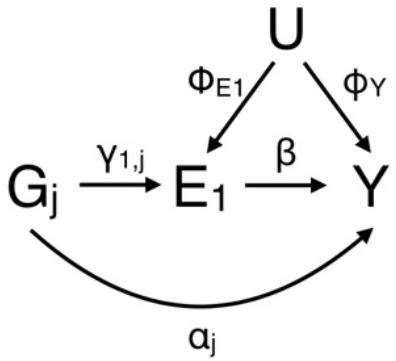
3. **[exclusion restriction]** The genetic marker is independent of the outcome given the exposure (E) and all confounders of the exposure-outcome association (U)

   No effect from G_j to Y outside of $G_j \rightarrow E \rightarrow Y$

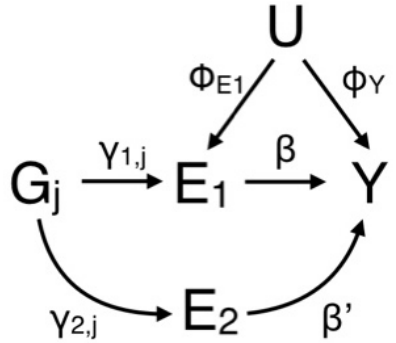


Violation of exclusion restriction assumption

Direct pleiotropic effect

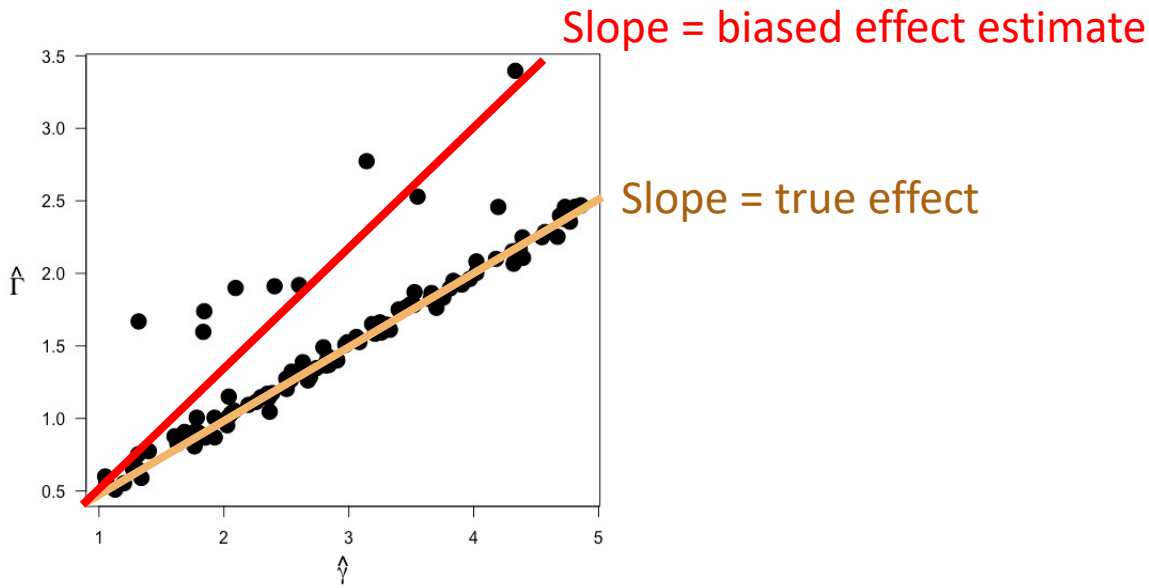


Mediated pleiotropic effect



Violation of exclusion restriction assumption in multivariable MR

An illustration of MR analysis where a subset of SNPs with pleiotropic effects



Multiple genetic variants

- In most circumstances, a single genetic variant individually typically explains only a very small proportion of the variation in a risk factor; referred as "weak instruments", particularly in small sample sizes.
- To overcome this, investigators have developed methods that use multiple genetic variants that collectively explain more of the variation in a risk factor than a single variant and thus have more statistical power.

Hemani et al., 2018

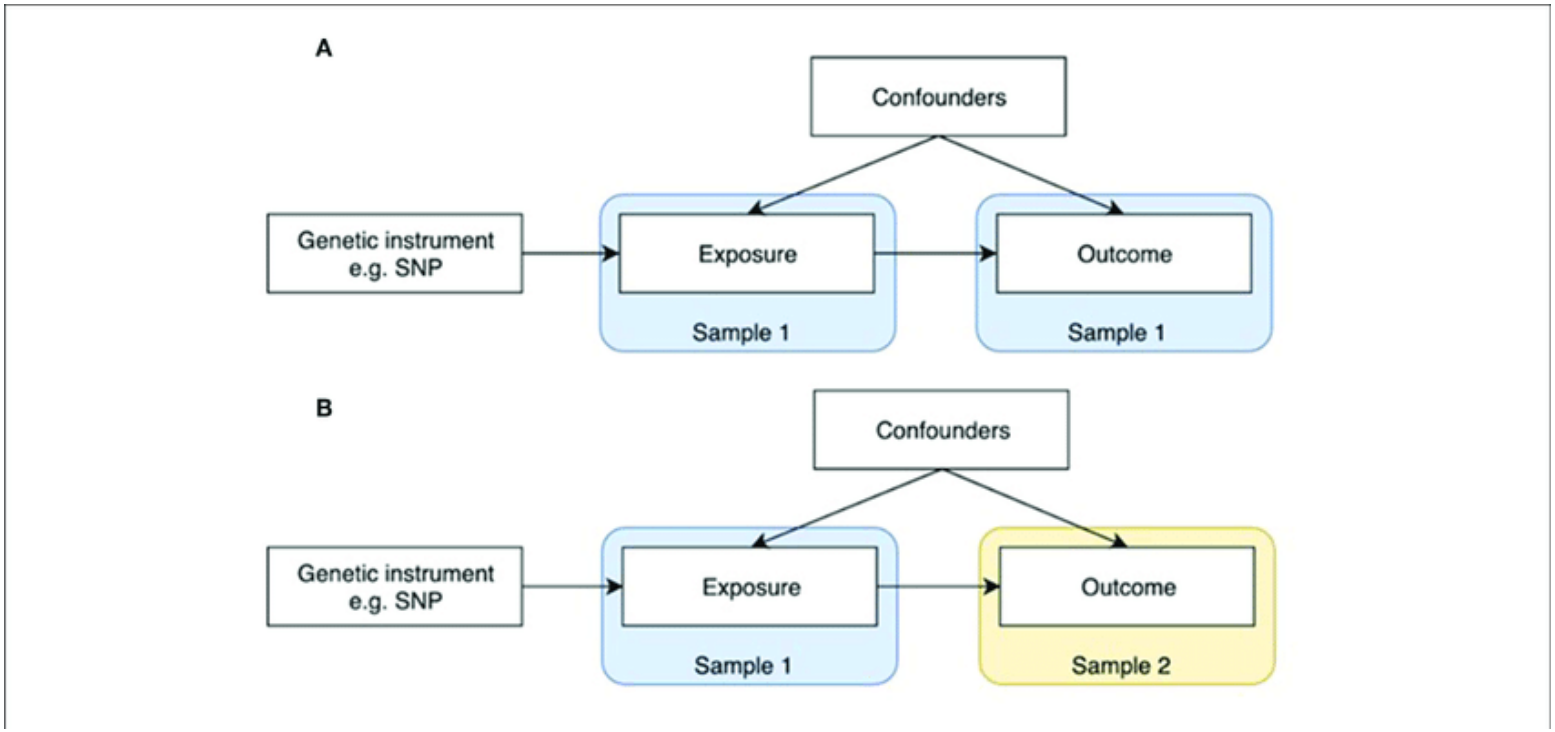
MR using multiple instruments

For a given exposure-outcome pair, MR can be done with **multiple (independent) SNPs** and then aggregated for a more precise estimate

- Individual level data
  - Polygenic score as a single instrument
- Summary statistics
  - Multivariable MR: meta-analysis results from multiple instruments

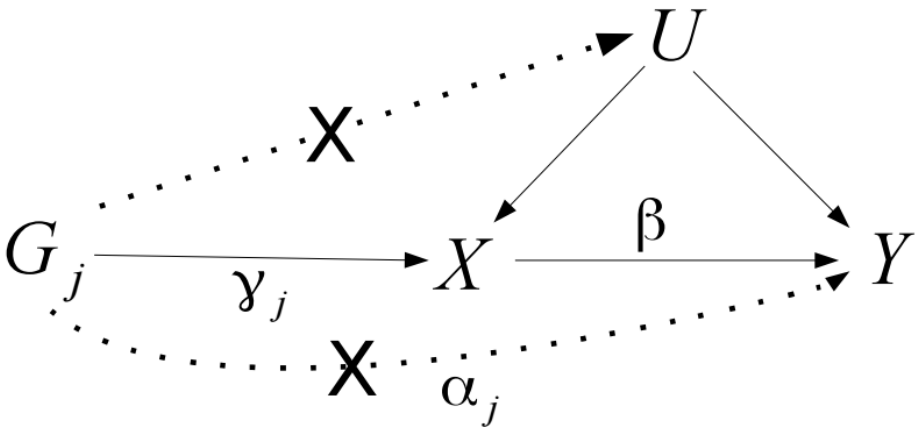
Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods (Burgess et al. 2015)

Two sample MR



Zeng., et al Frontiers of Epidemiology 2019

Two Sample MR:
Multiple Variants



Causal estimate using IVW from summarized data:

$$\frac{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Yj}^{-2} \hat{\beta}_j}{\sum_{j=1}^{J} \hat{\gamma}_j^2 \sigma_{Yj}^{-2}}. \quad = \beta.$$

(Approximates TSLS)

where $\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$ is the ratio method estimate for variant j, and σ_{Yj} is the standard error in the regression of the outcome on the jth genetic variant, assumed to be known.

Steps to perform two-sample MR

1. Identify genetic instrumental variables (IV)
2. Obtain SNP-exposure associations from data source 1
3. Obtain SNP-outcome associations from data source 2
4. Harmonize SNP effects on exposure and outcome
5. Generate MR estimates
6. Perform sensitivity analyses

1. Identify genetic instrumental variables

- Genetic IV are characterized as SNPs that reliably associate with the exposure.

Genetic IV selection

- *Statistical significance*
  - Genetic IV should be obtained from well-conducted GWAS, typically involving their detection in a discovery sample at a GWAS threshold of statistical significance (e.g. $p<5x10^{-8}$) followed by replication in an independent sample.

1. Identify genetic instrumental variables

Genetic IV selection (cont.)

- *Independence*
  - Genetic IV should be independent, i.e., not in linkage disequilibrium (LD).
  - LD is the correlation between nearby variants such that the alleles at neighboring polymorphisms (observed on the same chromosome) are associated within a population more often than if they were unlinked.
  - Set LD threshold at, e.g., R^2 =0.001 or R^2=0.1 (LD clumping)

- *Biological link with the exposure*

2. Obtain SNP-exposure associations from data source 1

- Data to be extracted for each SNP are..
  - Reference allele (e.g. G)
  - Effect allele (e.g. A)
  - Effect sizes (β_x) and standard errors (σ_x) of effect alleles on the exposure.
- Other data are..
  - Sample size, reference allele and effect allele frequency.

3. Obtain SNP-outcome associations from data source 2

- As with the exposure data, the outcome data must contain at a minimum the effect alleles, the reference alleles, the effect sizes (β_y) and their standard errors (σ_y) of the effect alleles on the outcome.

LD proxies

- If a particular SNP is not present in the outcome dataset, it is possible to use SNPs that are LD proxies instead, i.e., use SNPs that are in strong linkage disequilibrium with the missing SNP.
  - E.g. minimum R^2 is 0.6 or 0.8.

4. Harmonize SNP effects on exposure and outcome

- Genetic associations with exposures and outcomes are typically reported per additional copy of a particular allele. Hence, when combining summarized data on genetic associations, it is important to ensure that genetic associations are expressed per additional copy of the same allele.
- This is particularly important as not all publicly-available data resources are consistent about reporting strand information correctly.
- To generate a summary set for each SNP, we need its effect and standard error on the exposure and the outcome **corresponding to the same effect alleles.**

5 estimate MR

Multiple instruments:

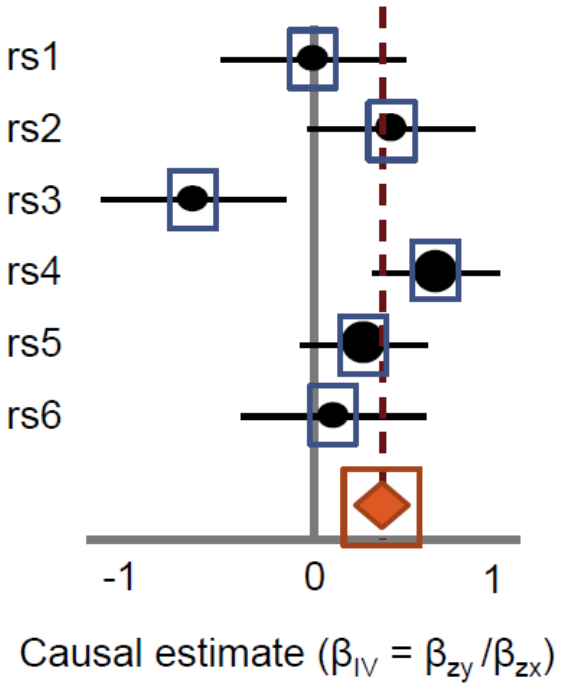
Inverse variance weighted (IVW) method

- Traditional MR method which uses a meta-analysis approach to combine the Wald ratio estimates of the causal effect obtained from different SNPs.
- IVW estimates are equivalent to a weighted linear regression of SNP-outcome associations on SNP-exposure associations with the intercept constrained to zero

  - $\widehat{\Gamma}_j$: genotype-disease associations (SEs: σ_{Yj})
  - $\widehat{\gamma}_j$: genotype-phenotype associations (SEs: σ_{Xj})
  - With L instruments
  - and instrument specific ratio estimates: $\widehat{\beta}_j = \widehat{\Gamma}_j / \widehat{\gamma}_j$

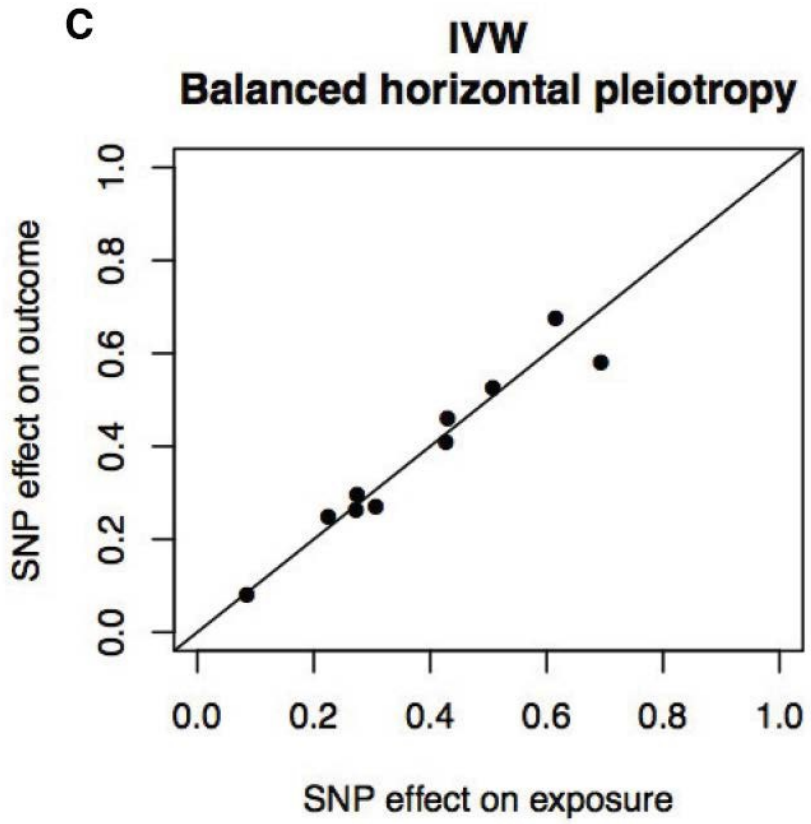
$$\widehat{\beta}_{\text{IVW}} = \frac{\sum_{j=1}^{L} w_j \widehat{\beta}_j}{\sum_{j=1}^{L} w_j}, \; w_j = \frac{\widehat{\gamma}_j^2}{\sigma_{Yj}^2}$$

IVW estimate similar to IVW meta-analysis



Borges MC. Mendelian Randomization. [PowerPoint presentation]. MRC Integrative Epidemiology Unit University of Bristol.

Inverse variance weighted (IVW) method



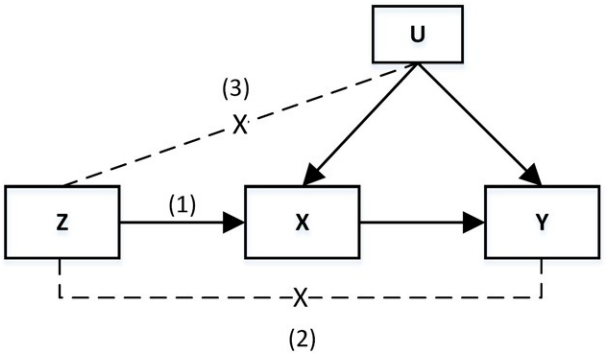
Hemani et al., 2018

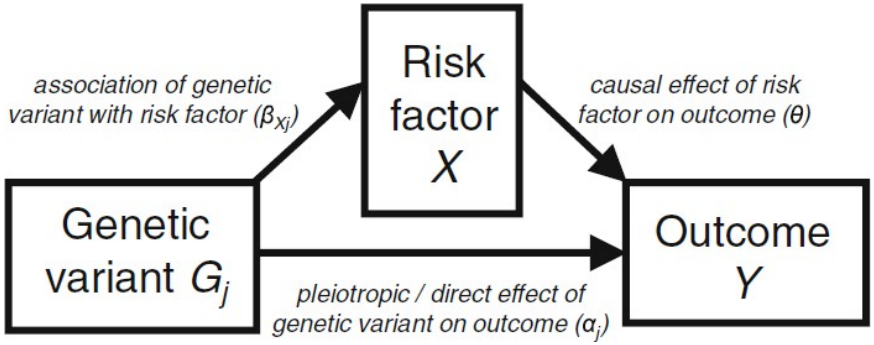
Inverse variance weighted (IVW) method

- The IVW method is the most efficient estimate of the causal effect when all genetic variants are valid instruments.
- IVW estimates can be biased in cases where one or more variants exhibit horizontal pleiotropy (invalid instruments).

Horizontal pleiotropy

- A genetic variant affects the outcome through pathways that are not mediated via the exposure



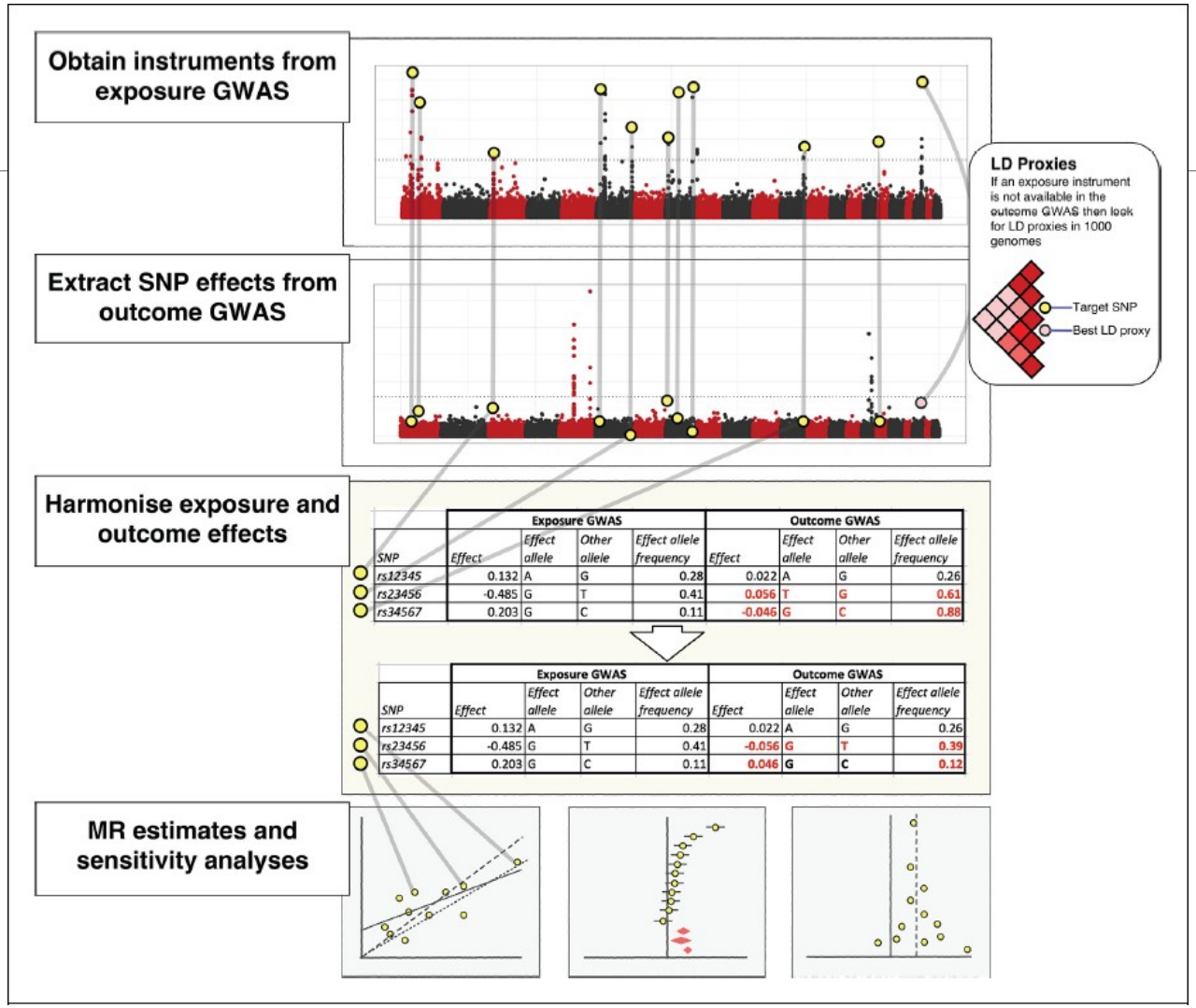


1.
2. **Obtain instruments from exposure GWAS**

3. **Extract SNP effects from outcome GWAS**

4. **Harmonise exposure and outcome effects**

5.
6. **MR estimates and sensitivity analyses**



	Exposure GWAS				Outcome GWAS			
SNP	Effect	Effect allele	Other allele	Effect allele frequency	Effect	Effect allele	Other allele	Effect allele frequency
rs12345	0.132	A	G	0.28	0.022	A	G	0.26
rs23456	-0.485	G	T	0.41	0.056	T	G	0.61
rs34567	0.203	G	C	0.11	-0.046	G	C	0.88

	Exposure GWAS				Outcome GWAS			
SNP	Effect	Effect allele	Other allele	Effect allele frequency	Effect	Effect allele	Other allele	Effect allele frequency
rs12345	0.132	A	G	0.28	0.022	A	G	0.26
rs23456	-0.485	G	T	0.41	-0.056	G	T	0.39
rs34567	0.203	G	C	0.11	0.046	G	C	0.12

Hemani et al., 2018

Multiple instruments –notice

LD assessment

- More result in confounding
- We can use plink clumped independent SNPs

Pleiotropy assessment

- MR-Egger regression
  - Egger regression is used to examine publication bias
  - intercept distinct from the origin provides evidence for pleiotropic effects

Population stratification assessment

- Exposure and outcome should be from the same race

Pleiotropy

- One gene can affect many (even seemingly unrelated) phenotypes
- Mendelian Randomisation makes the assumption of no pleiotropy
- In this case, this means that we know the genotype is only influencing the phenotype via the considered exposure
- I.e. ApoE2 only affects serum cholesterol levels, and cannot affect cancer risk by other, unobserved means.
- This is a big assumption, prior knowledge is necessary.
- If possible, using multiple, independent SNPs (instruments) helps to alleviate this issue (as if they are all consistent then it is unlikely that they all have other pathways causing the same change) - but note they must not be in Linkage Disequilibrium!

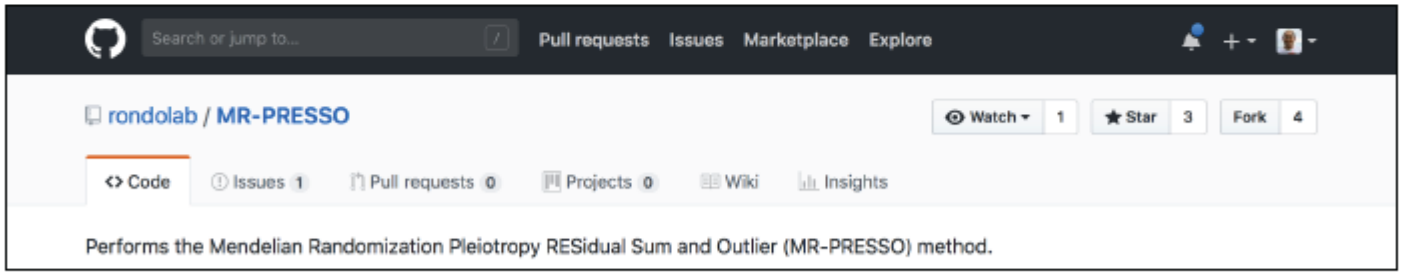
Horizontal pleiotropy

Detecting and controlling for pleiotropic bias in MR

- Detecting for pleiotropic bias
  - Average pleiotropic bias
    - MR-Egger regression
  - SNPs with pleiotropic bias as outliers
    - Heterogeneity test (modified Q and modified Q')
    - MR-PRESSO
- Controlling for pleiotropic bias
  - Average pleiotropic bias
    - MR-Egger regression
  - With known mediated pleiotropic bias
    - Multivariable MR
  - SNPs with pleiotropic bias as outliers
    - MR-PRESSO
    - Median-based MR estimator
    - Mode-based MR estimator

Mendelian Randomization Pleiotropy RESidual Sum and Outlier approach: MR-PRESSO

- MR-PRESSO detects and corrects for pleiotropic bias in 2-sample MR
  - Correcting for pleiotropic bias in MR while preserving the statistical power of IVW meta-analysis



https://github.com/rondolab/MR-PRESSO

Verbanck*, Chen*, Neale$, Do$. 2018.

MR-egger concept

- In Mendelian Randomization when multiple genetic variants are being used as IVs, Egger regression can:
  - Identify the presence of 'directional' pleiotropy (biasing the IV estimate)
  - provide a less biased causal estimate (in the presence of pleiotropy)

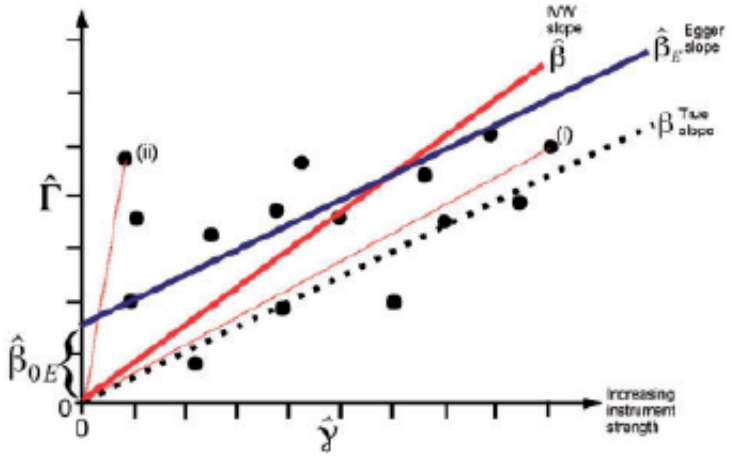
However, MR Egger lacks power

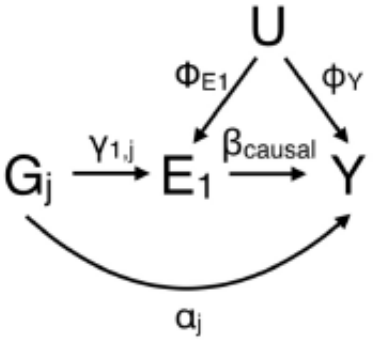
Detecting average pleiotropic bias: MR-Egger regression

- MR-Egger regression under InSIDE condition*

$$\hat{\Gamma}_{1,j} = \beta_0 + \beta_{causal}\hat{\gamma}_{1,j} + \varepsilon_j$$

*InSIDE: Instrument Strength Independent of Direct Effect
$(\gamma_{1,j} \perp \alpha_j)$



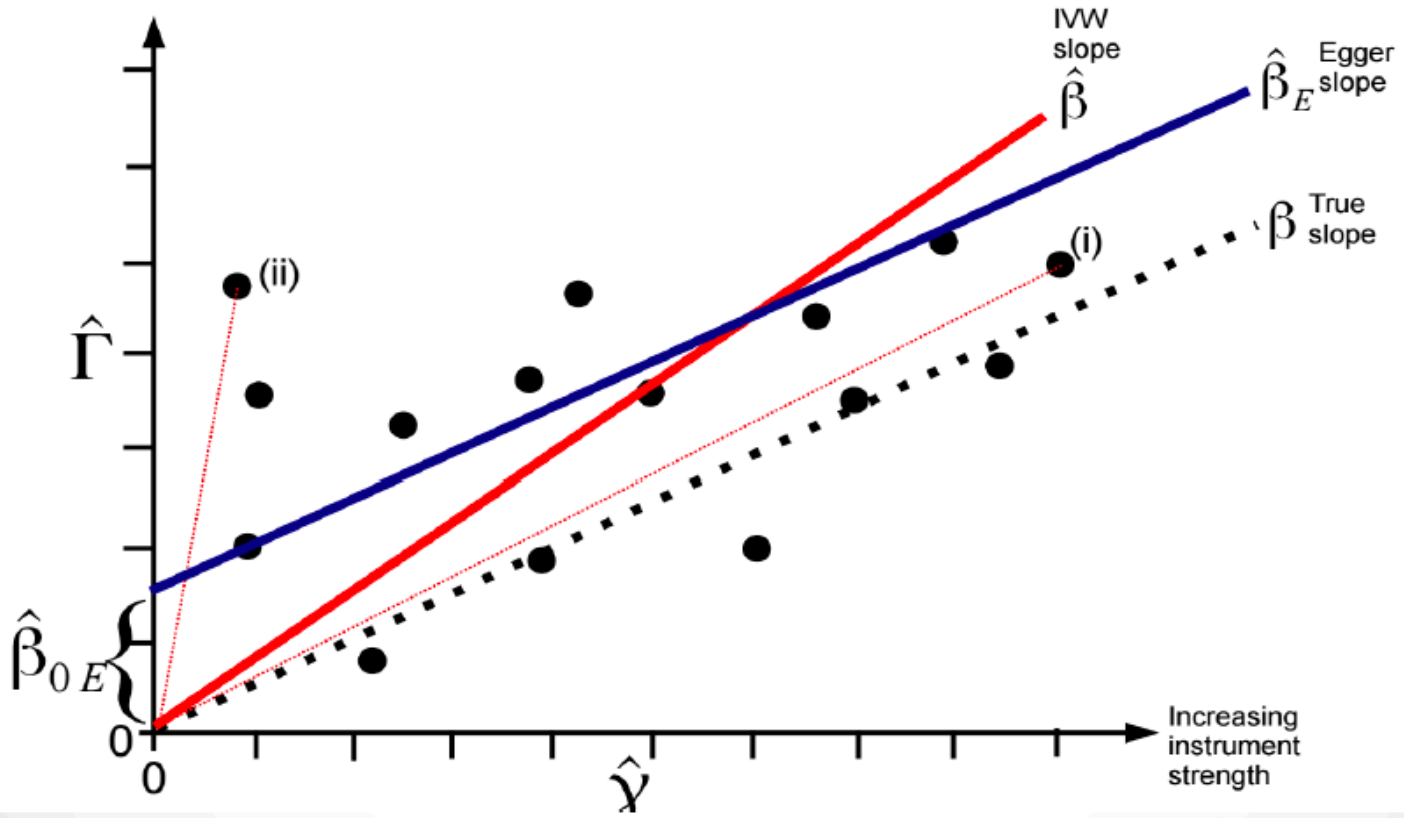


Bowden et al. 2015. International Journal of Epidemiology.

Egger regression:

$$\hat{\Gamma}_j = \beta_{0E} + \beta_E \hat{\gamma}_j .$$

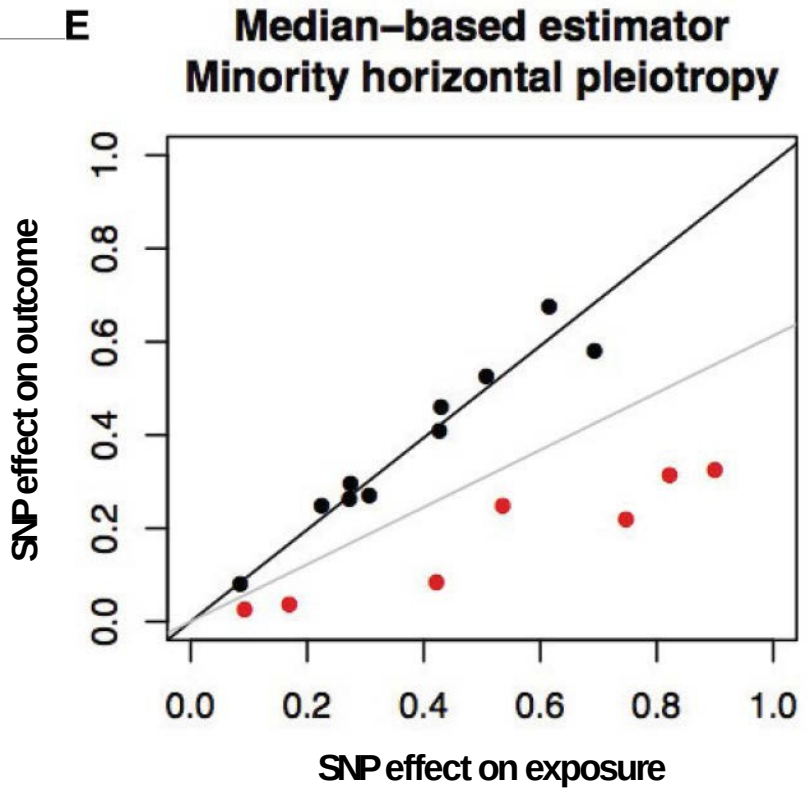
Intercept not constrained to zero



Egger's test assesses whether the intercept term is significantly different from zero. The estimated values of the intercept can be interpreted as the average pleiotropic effect across all genetic variants. An intercept term different from zero indicates directional pleiotropy

Median-based estimator

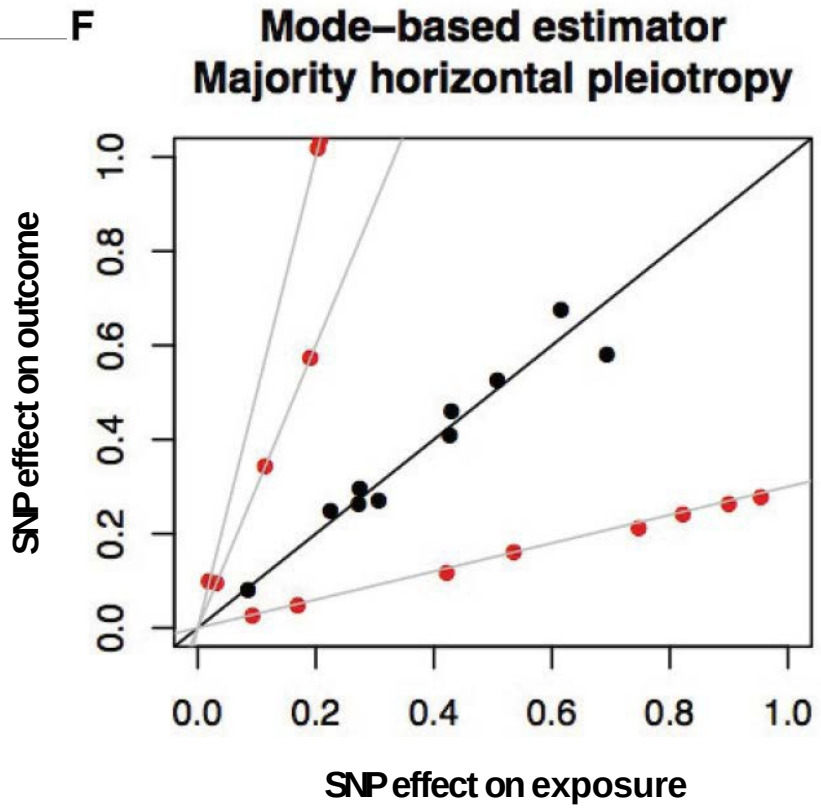
- The median-based estimator provides an unbiased causal estimate when the majority of SNPs are valid instruments.
- It takes the median (or weighted median) of all IV causal estimates.
- This estimator is consistent when at least 50% of the instrumental variables are valid.



Hemani et al., 2018

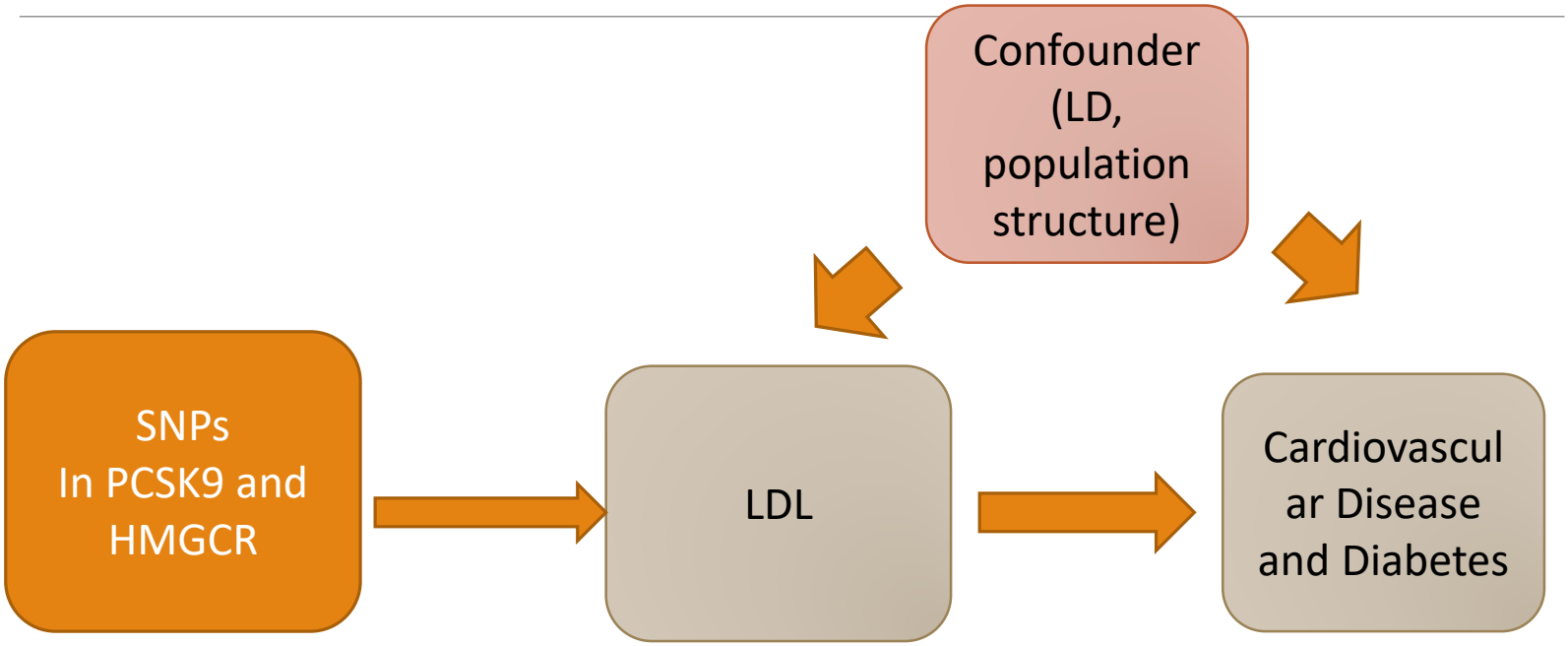
Mode-based estimator

- The mode-based estimator clusters the SNPs into groups based on similarity of causal effects, and returns the causal effect estimate based on the cluster that has the largest number of SNPs
- It gives an unbiased causal effect if the SNPs within the largest cluster are valid instruments.



Hemani et al., 2018

Two sample MR design PCSK9 & HMGCR



Nejm 2016

Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes

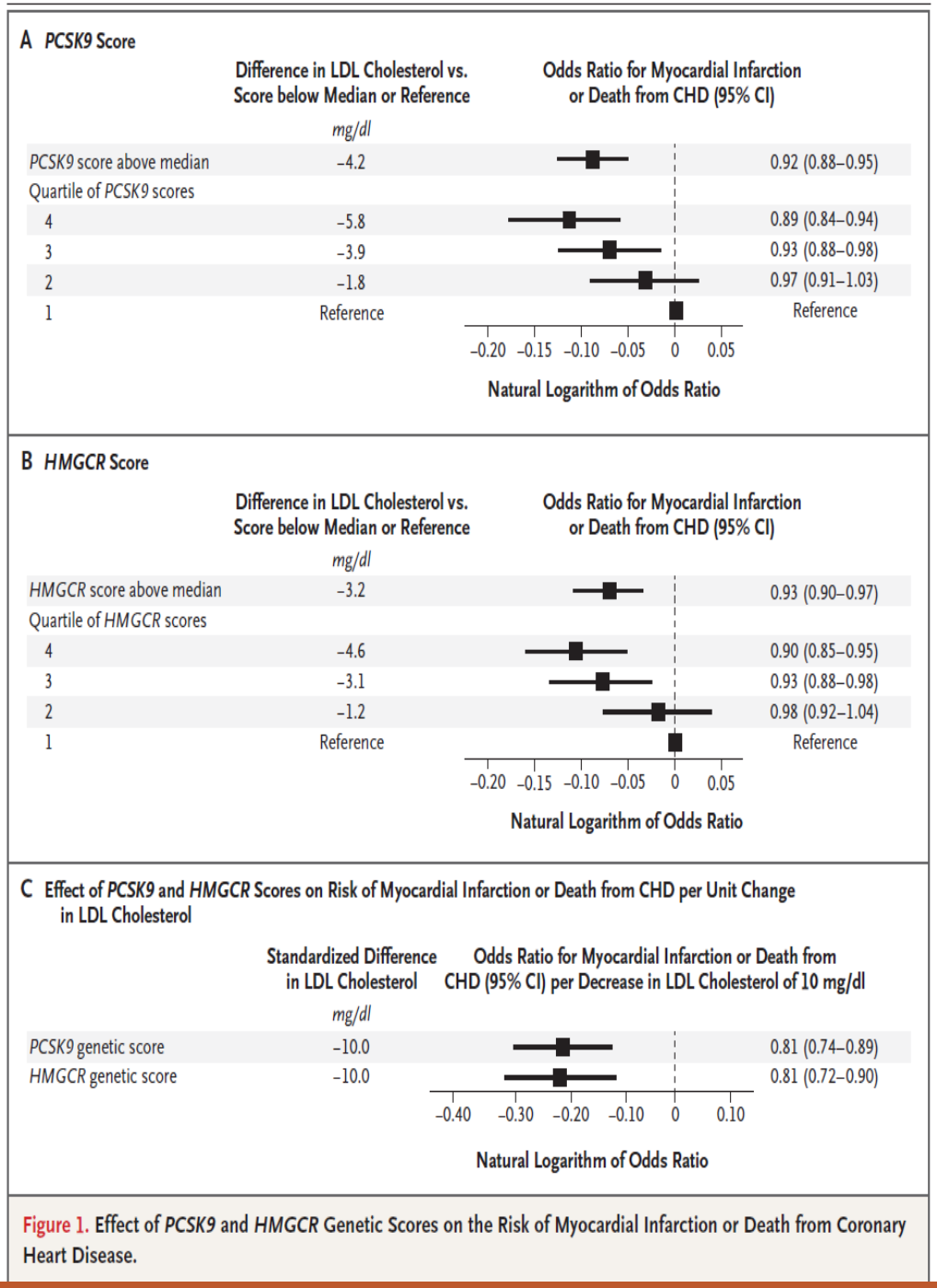


Figure 1. Effect of *PCSK9* and *HMGCR* Genetic Scores on the Risk of Myocardial Infarction or Death from Coronary Heart Disease.

- PCSK9 and HCGMR reduce serum LDL_c level
- PCSK9 are evaluated clinical trail for treatment CVD
- Global Lipids Genetics consortium choose Ivs ($P<5*10^{-8}$)

Nejm 2016

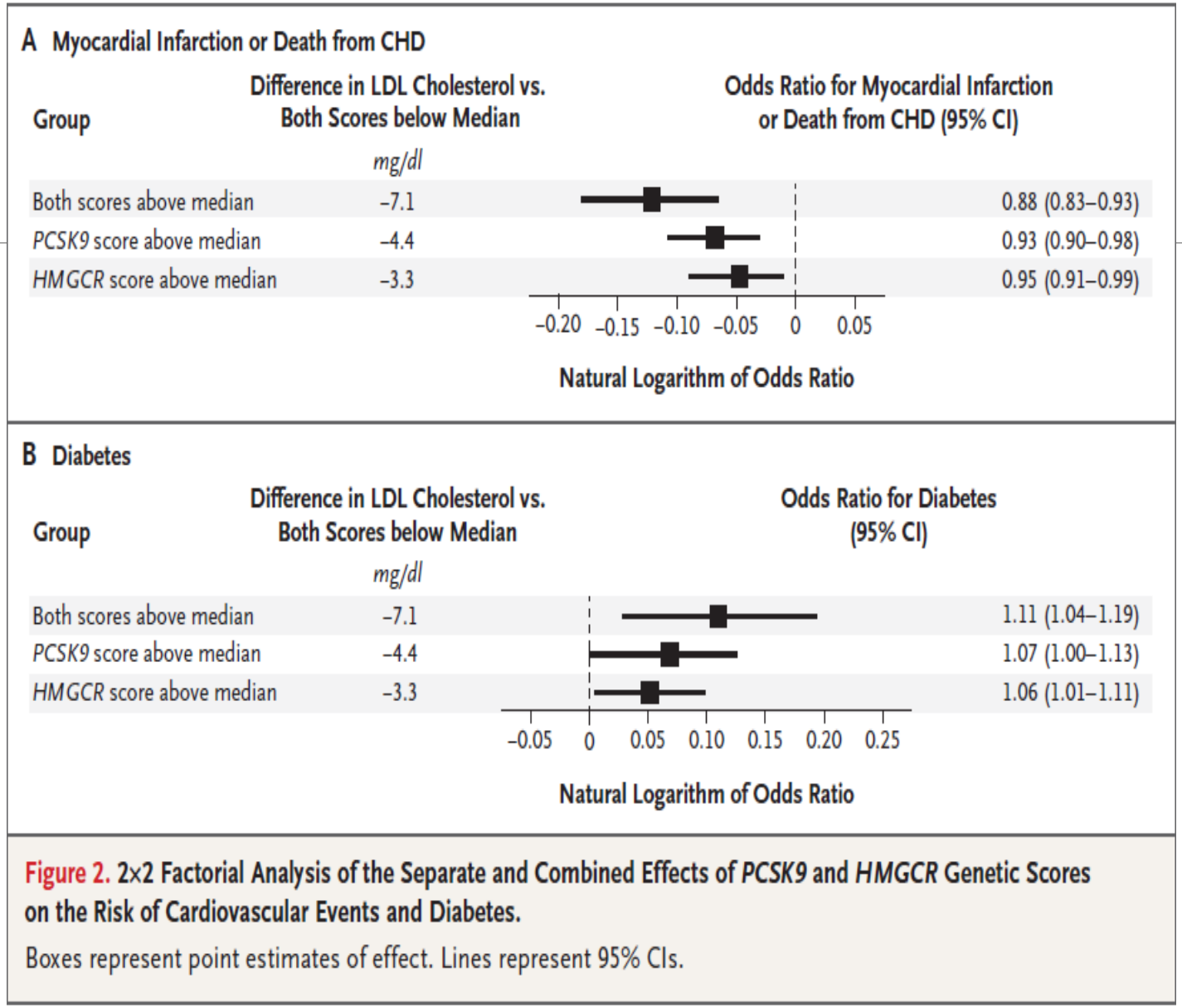


Figure 2. 2×2 Factorial Analysis of the Separate and Combined Effects of *PCSK9* and *HMGCR* Genetic Scores on the Risk of Cardiovascular Events and Diabetes.

Boxes represent point estimates of effect. Lines represent 95% CIs.

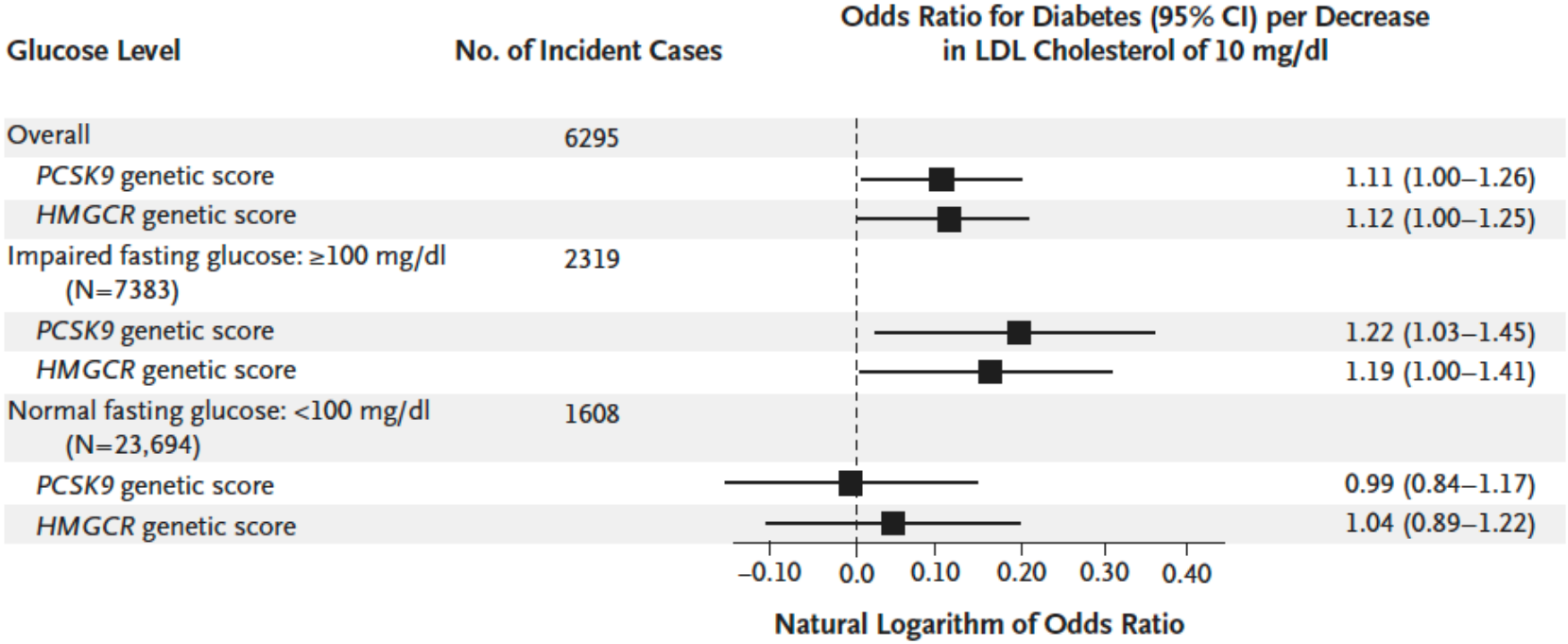
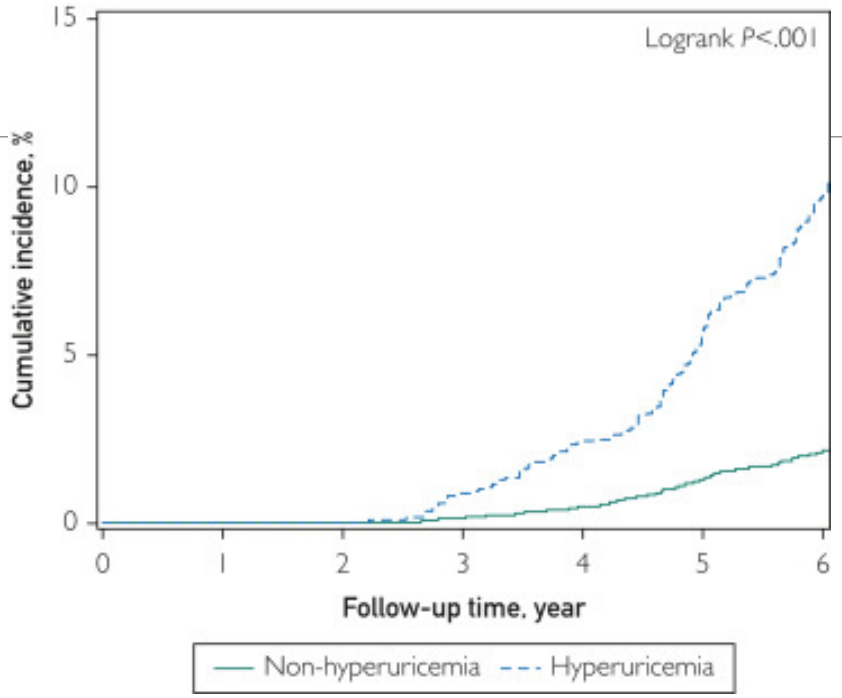


Figure 4. Effect of *PCSK9* and *HMGCR* Scores on the Risk of Incident Diabetes.

A total of 6295 incident cases of diabetes occurred during follow-up in the prospective cohort studies. After the exclusion of participants with prevalent diabetes, baseline fasting plasma glucose levels were available for 31,077 participants. The main analysis included all the participants after the exclusion of 4340 participants with prevalent diabetes; the subgroup analysis that was stratified according to fasting plasma glucose level included the 31,077 participants without prevalent diabetes for whom baseline fasting plasma glucose levels were available. Boxes represent point estimates of effect. Lines represent 95% CIs.

Serum Urate and CKD



21 SNPs associated with serum urate from GWAS catalog in EAS population

LD, confounding,

pleiotropy effect

CKD

Serum Urate and CKD

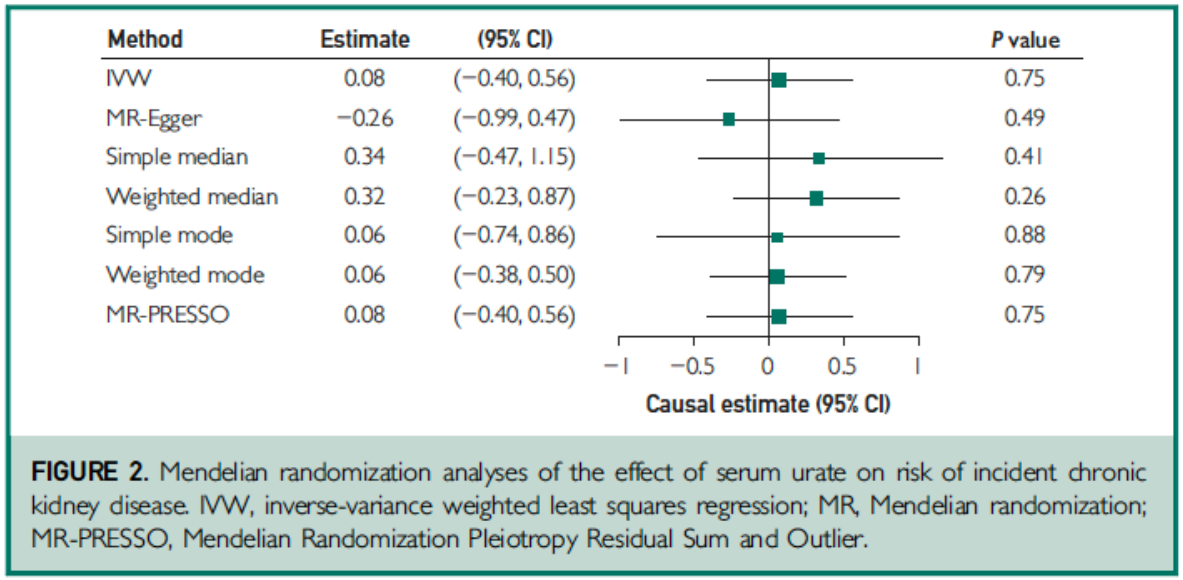


FIGURE 2. Mendelian randomization analyses of the effect of serum urate on risk of incident chronic kidney disease. IVW, inverse-variance weighted least squares regression; MR, Mendelian randomization; MR-PRESSO, Mendelian Randomization Pleiotropy Residual Sum and Outlier.

Table 3Genetic Risk Scores of Serum Urate and Risk of Incident Chronic Kidney Disease

[a], [b]

	HR (95% CI)[c]	*P*
Weighted GRS	1.03 (0.72-1.46)	.89
SLC2A9 (rs3733588)	1.09 (0.93-1.28)	.28

Mayo Clin Proc. n April 2023;98(4):513-52

Mendelian randomization software

MendelianRandomization in R package

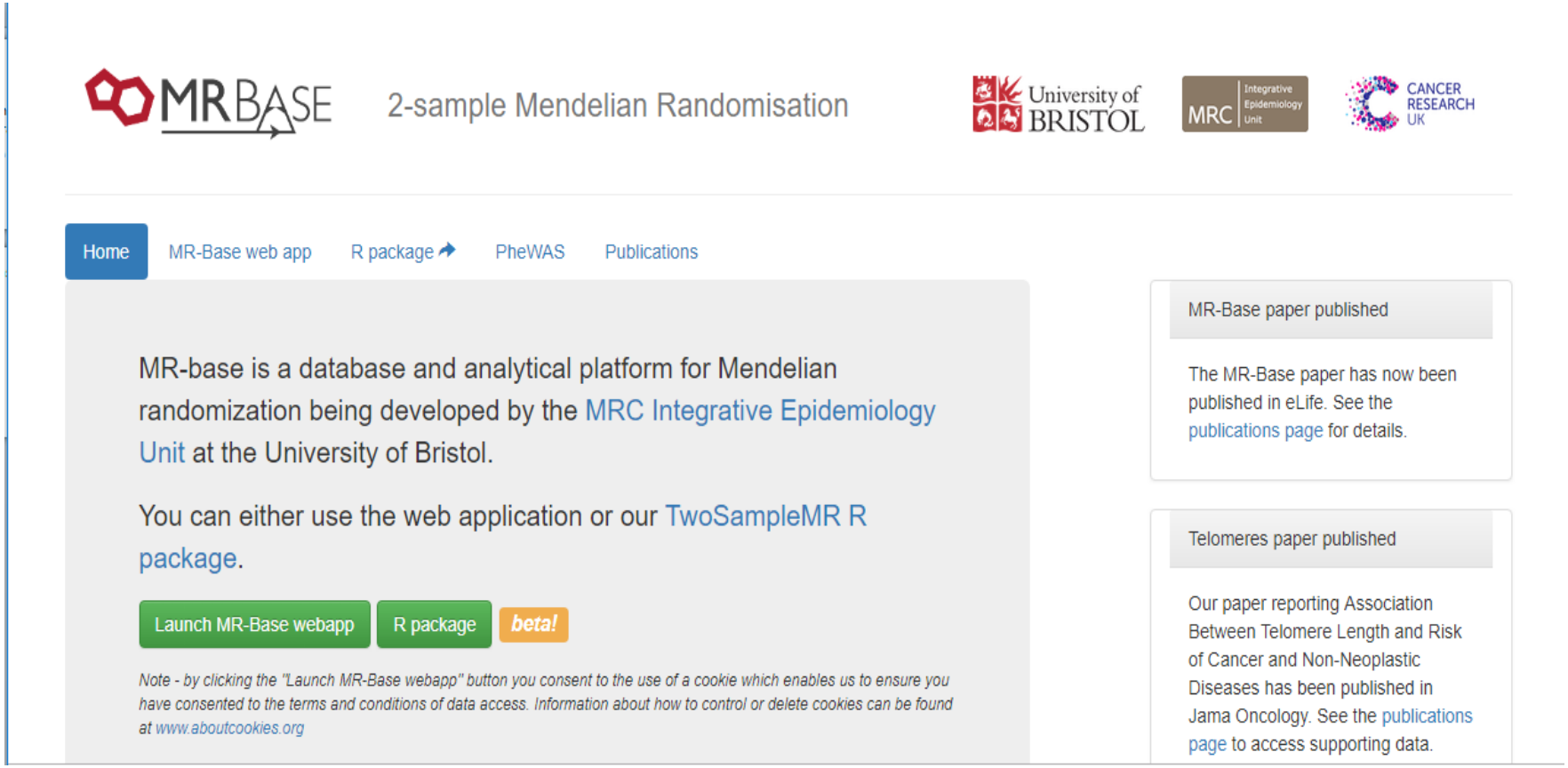
- Encodes several methods for performing Mendelian randomization analyses with summarized data. Summarized data on genetic associations with the exposure and with the outcome can be obtained from large consortia. These data can be used for obtaining causal estimates using instrumental variable methods

Two stage least square regression

- Using plink choose the instrument variants
- GRS
- SAS/R/STAT

MR base website

www.mrbase.org



Gibran Hemani, Jie Zheng, Kaitlin H Wade, Charles Laurin, Benjamin Elsworth, Stephen Burgess, Jack Bowden, Ryan Langdon, Vanessa Tan, James Yarmolinsky, Hashem A. $ *The MR-Base platform supports systematic causal inference across the human phenome.* eLife 2018. doi: https://doi.org/10.7554/eLife.34408

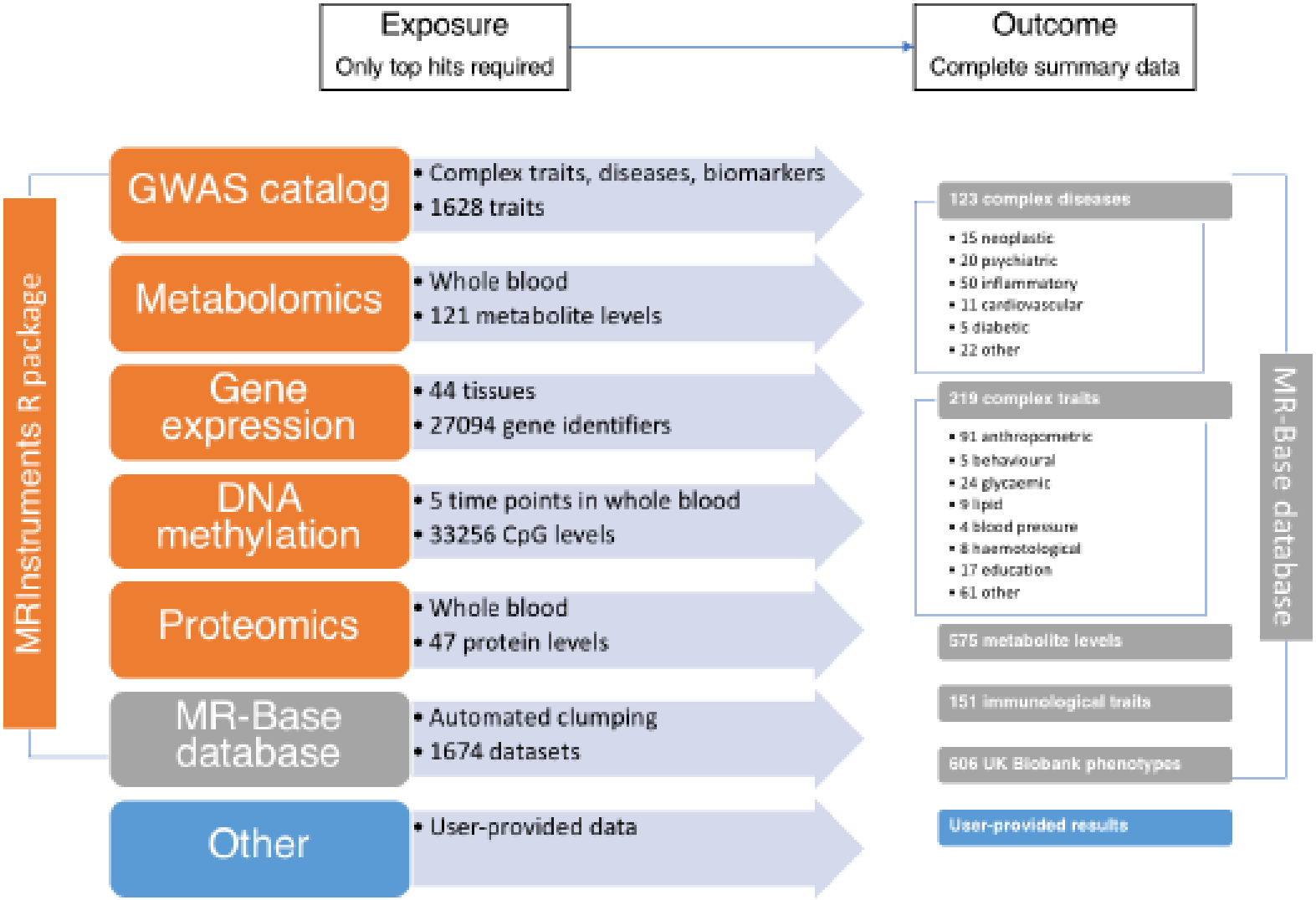
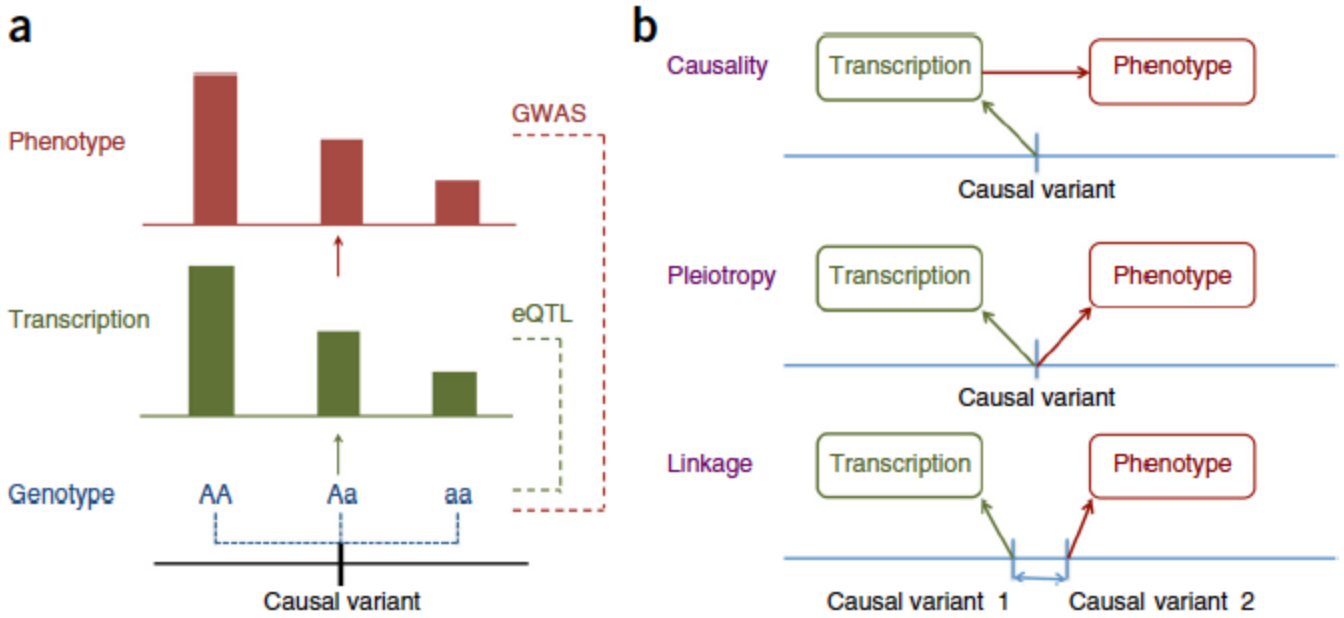


Figure 3. The data available through MR-Base and the possible exposure-outcome analyses that can be performed. Exposure traits can very broadly defined and may include molecular traits like gene expression, DNA-methylation, metabolites and proteins, as well as more complex traits, including cholesterol, body mass index, smoking and education. Further details on the traits with complete summary data can be found in *Supplementary file 1A*. The numbers reflect MR-Base in December 2017 and are updated on a regular basis.

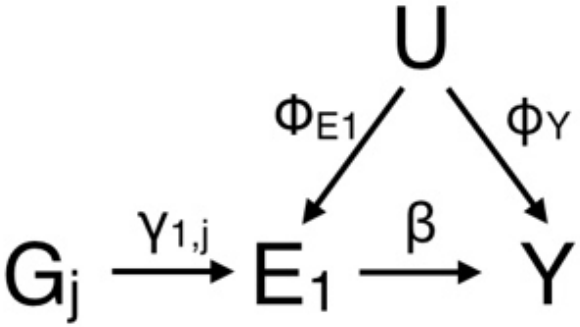
DOI: https://doi.org/10.7554/eLife.34408.005

MR applied to transcriptome-wide association study (TWAS): SMR



Zhu et al. 2016. Nature Genetics

Two step MR



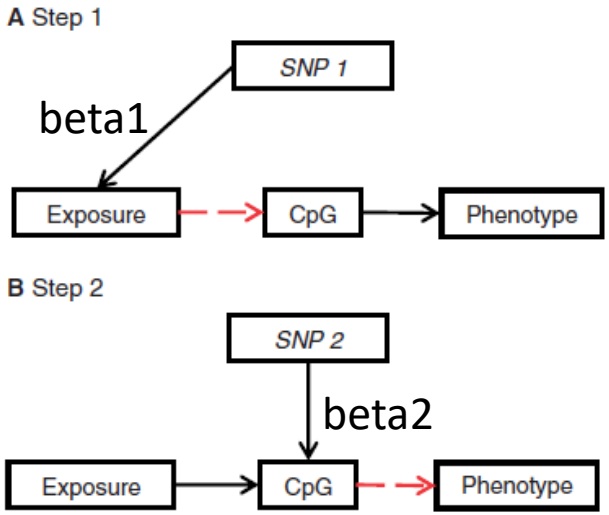
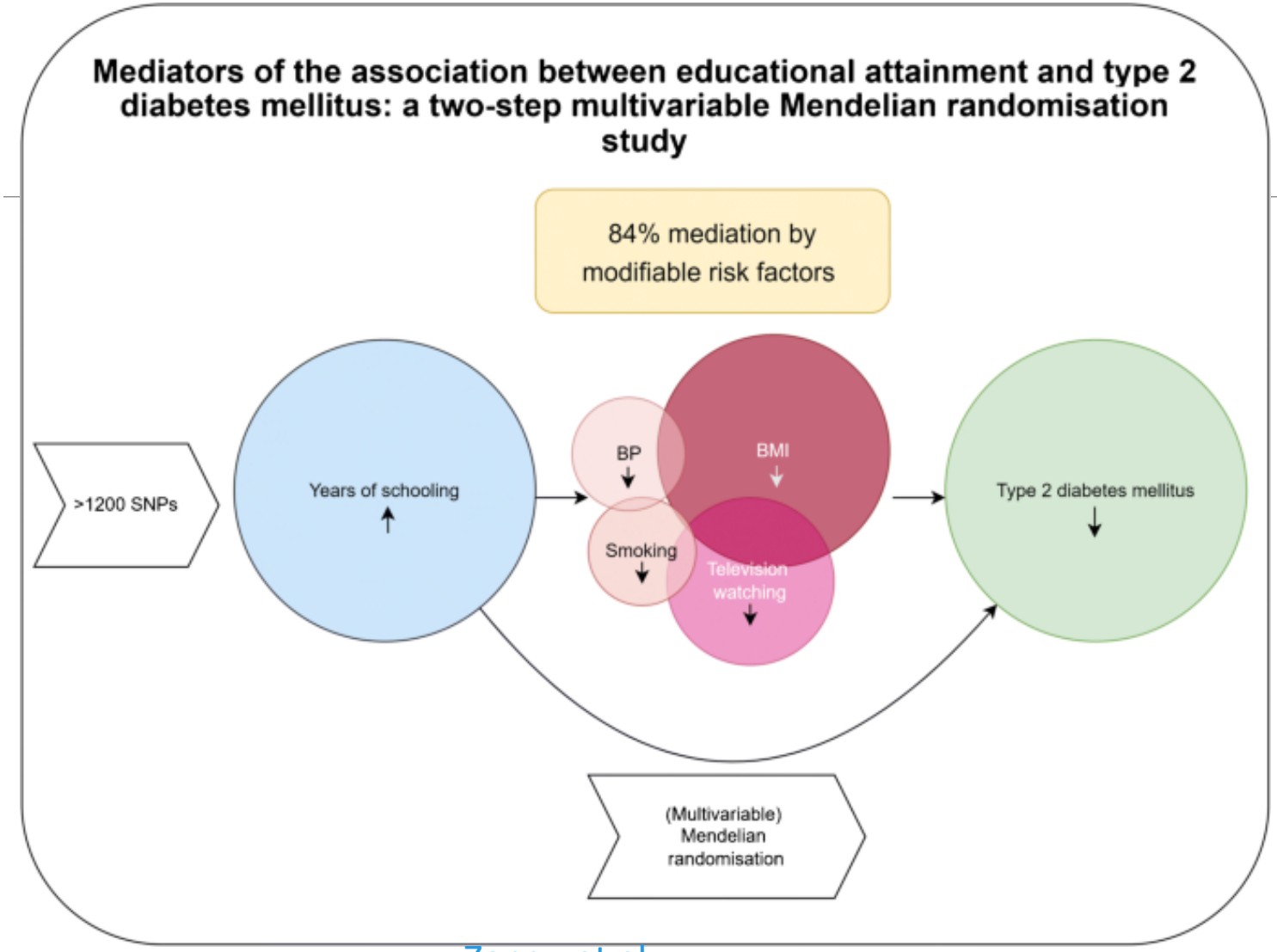


Figure 4 Two-step epigenetic Mendelian randomization: applying the principle of Mendelian randomization to DNA methylation as an intermediate phenotype. Genetic variants can be used as instrumental variables in a two-step framework to establish whether DNA methylation is on the causal pathway between exposure and disease. An overview of the two-step framework of this approach is shown. (A) First, an SNP is used to proxy for the environmentally modifiable exposure of interest and (B) secondly, a different SNP is used to proxy for DNA methylation levels

- Beta,beta 1, and beta 2 all significant
  - beta 1 * beta 2 (indirect /mediator)
  - beta-(beta 1 *beta 2) (director)
- Beta 0 no-sig + both beta 1 and beta 2 significant
  - Mediator is contributed all effect form exposure to outcome
- Beta 0 significant + beta1 or beta2 significant
  - mediator in not true

Example for two step MR



Zeng., et al
Diabetologia **volume 65**, pages1364–1374 (2022)

Summary

- IVW MR the most powerful option, but assumes the absence of horizontal genetic pleiotropy
- MR Egger, Weighted Median and Modal based estimators relax the strict requirement of no horizontal pleiotropy, but at the cost of decreased statistical power
- Crucial to perform sensitivity analyses and obtain metrics regarding the likely reliability of the MR estimates