Two-sample Mendelian Randomization

Thitiya Lukkunaprasit

PhD program in Clinical Epidemiology

Faculty of Medicine Ramathibodi Hospital, Mahidol University

Outline

1. Concept of Mendelian randomization (MR)
2. Instrumental variable assumptions
3. Multiple genetic variants
4. Data sources
5. Two-sample MR
   - What is two-sample MR?
   - Advantages of two-sample MR
   - Steps to perform two-sample MR
   - Limitations of two-sample MR

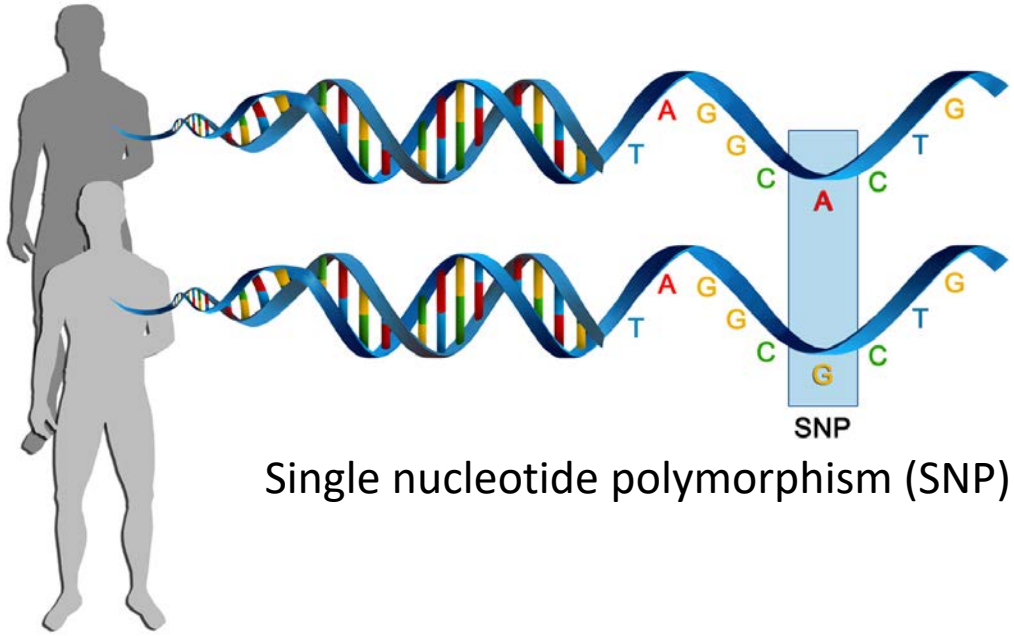
Instrument/Instrumental variable (IV)

- 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.

Mendelian randomization (MR)

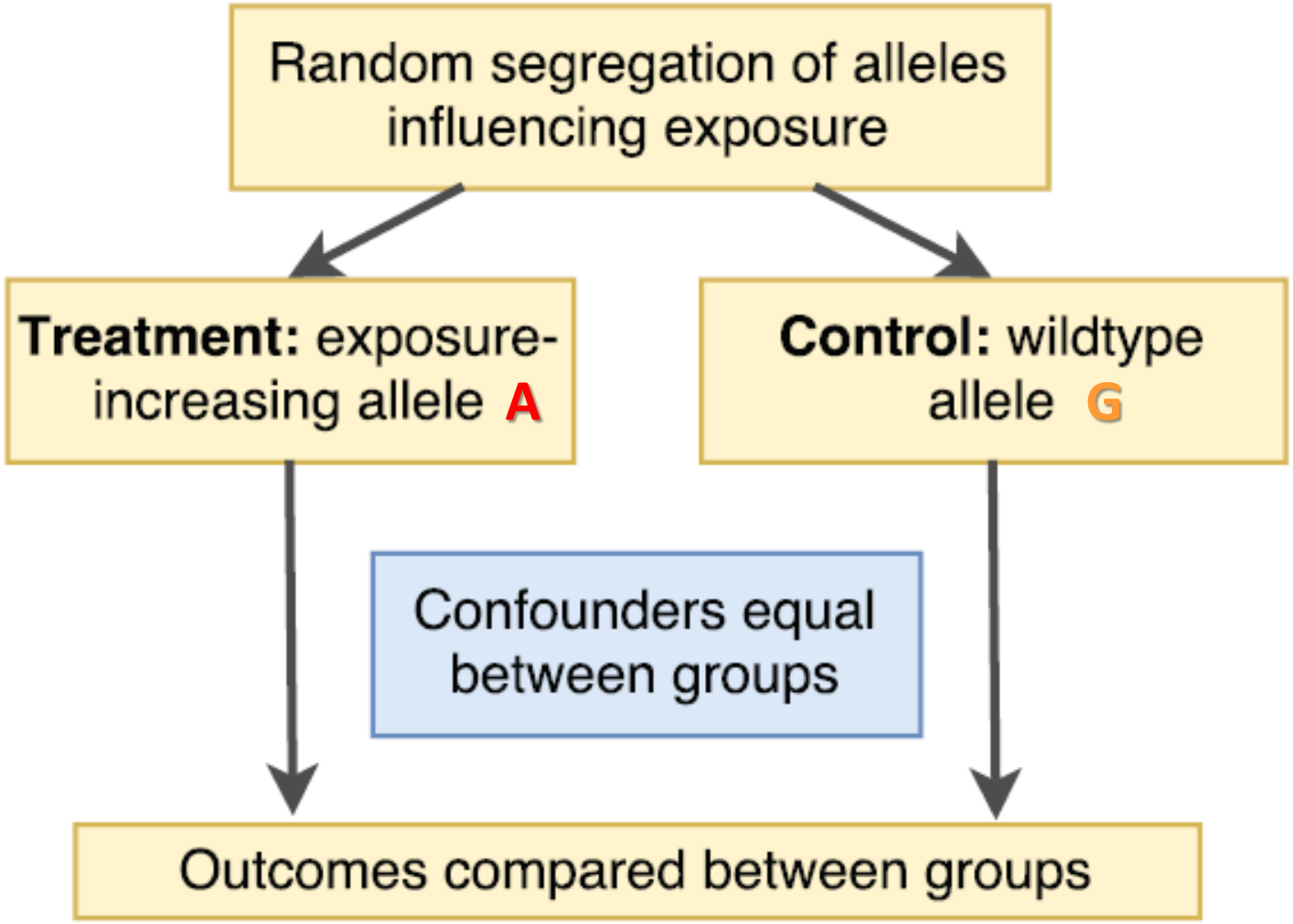
- An approach that uses genetic variants as IV to assess causal relationships between exposures/risk factors and clinical outcomes in observational data, based on the idea that alleles are randomly allocated at conception similar to RCT and thus there is no influence from confounders.
- The objective of MR analysis is a test of a causal hypothesis, and often additionally an estimate of a causal effect.

The similarity between MR and RCT



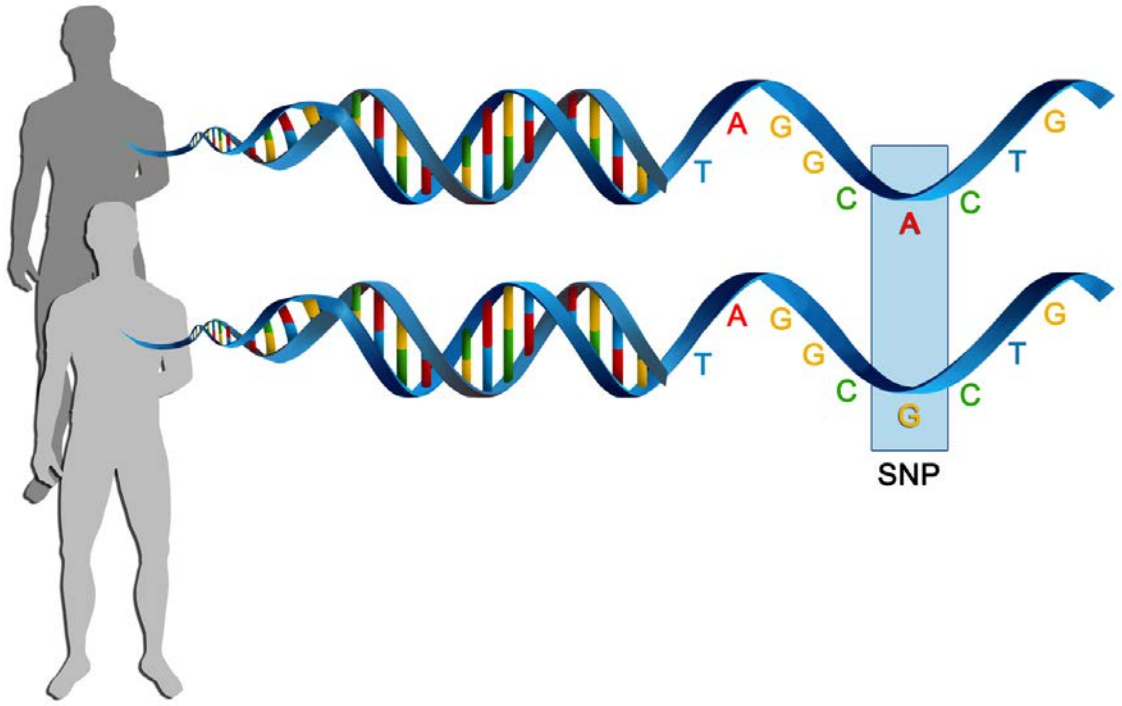
Single nucleotide polymorphism (SNP)

http://atlasofscience.org/single-nucleotide-polymorphisms-as-genomic-markers-for-high-throughput-pharmacogenomic-studies/

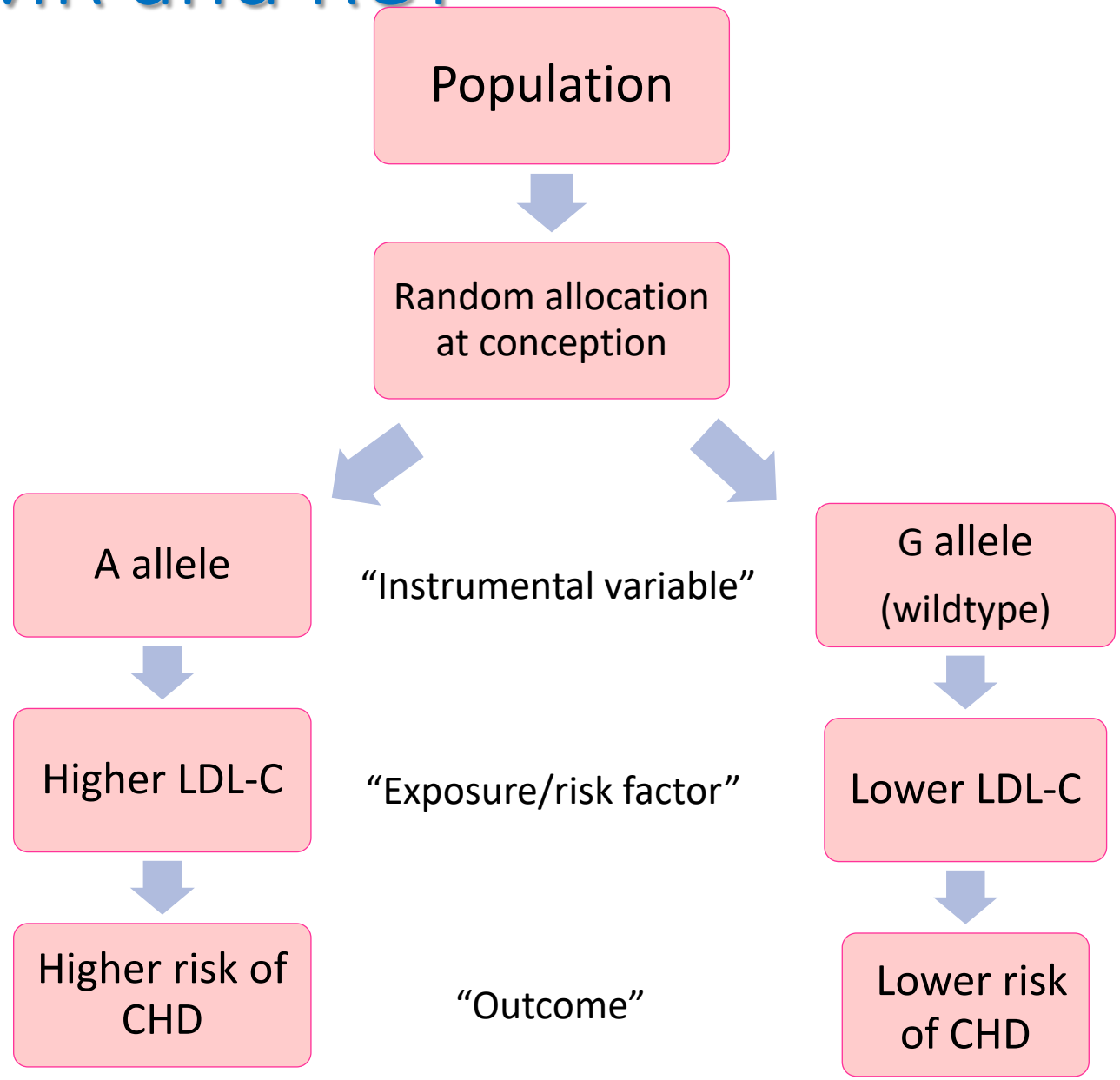


Hemani et al., 2018

The analogy between MR and RCT



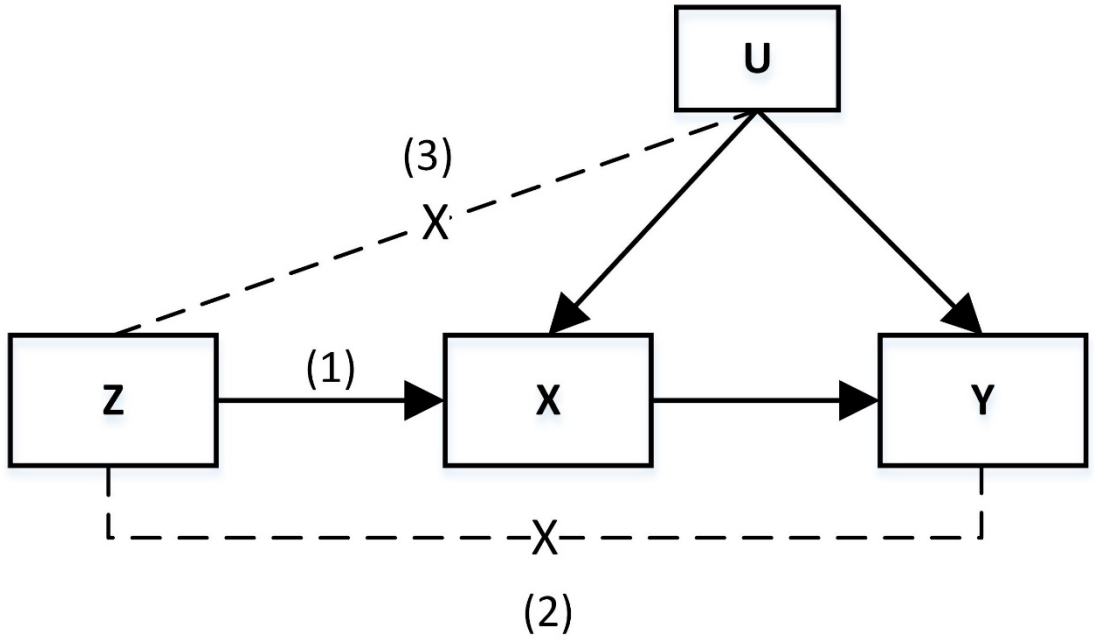
http://atlasofscience.org/single-nucleotide-polymorphisms-as-genomic-markers-for-high-throughput-pharmacogenomic-studies/



Instrumental variable (IV) assumptions

A genetic variant (usually a single nucleotide polymorphism, SNP) can be considered as an instrumental variable (IV) for a given exposure if it satisfies the IV assumptions:

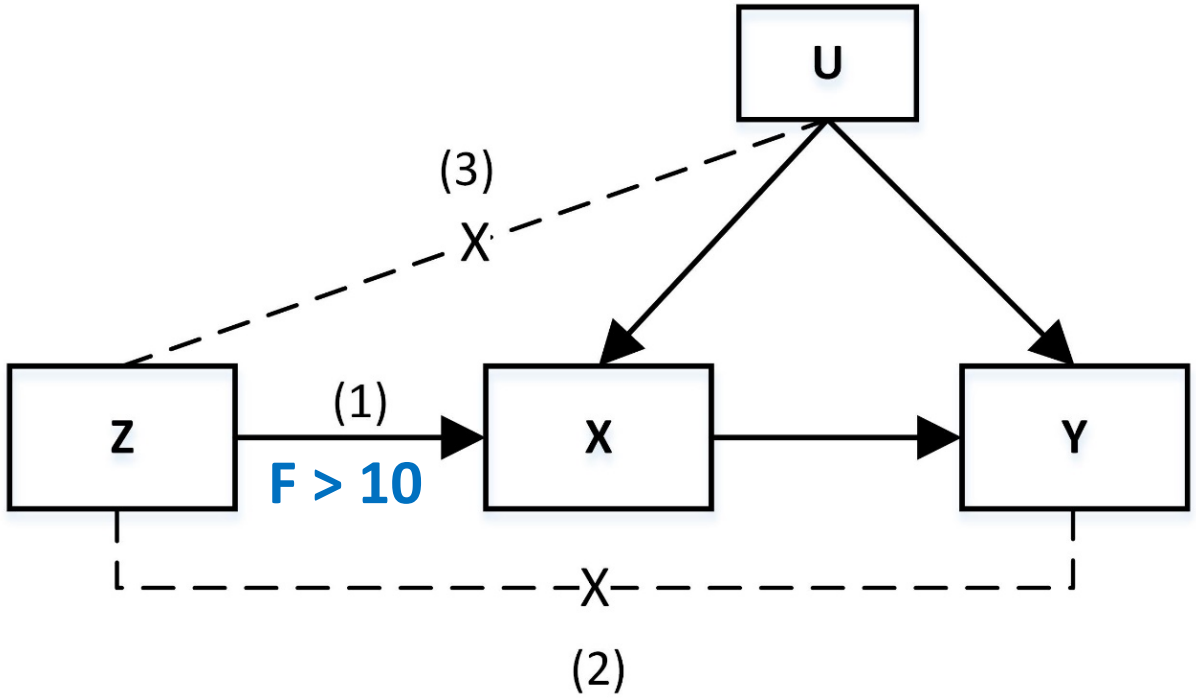
1. *The relevance assumption:* The IV (Z) is associated with the exposure/risk factor (X).
2. *The exclusion restriction:* The IV (Z) has no direct effect on the outcome (Y) except through the exposure/risk factor (X).
3. *The independence assumption:* The IV (Z) is not affected by confounders (U).



Instrument strength

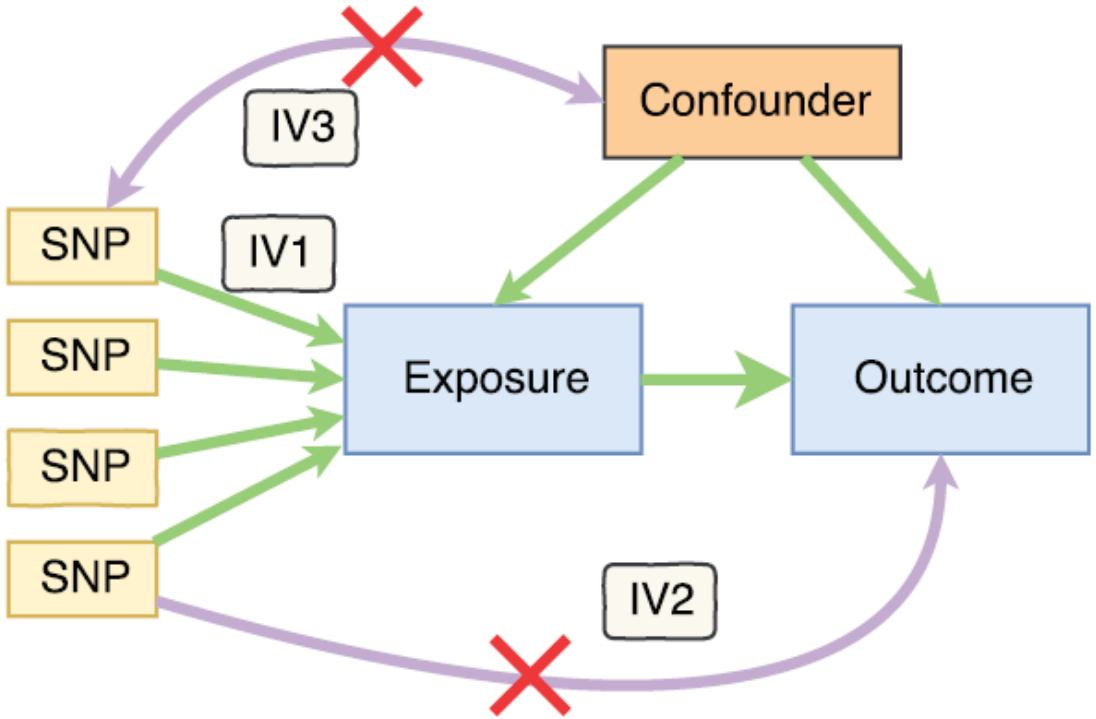
- F statistics
  - A measure of instrument strength and can be used to judge the extent of weak instrument bias
  - F statistics > 10, strong instrument

(Lawlor et al. 2008)



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

Data sources

1. One-sample MR

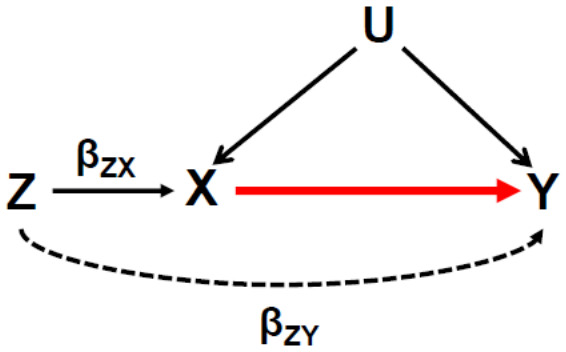
- Data from a single sample
- Genetic variants, exposure, and outcome are measured in the same individuals.
- Allows MR and conventional epidemiological findings to be compared in the same individuals
- The analysis is usually performed using individual-level data.
  - Two-stage least squares (2SLS) method

Data sources

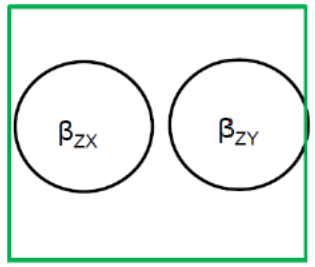
2. Two-sample MR

- Data from two samples from the same underlying population
- SNP-exposure associations (β_{ZX}) are estimated in one dataset, and SNP-outcome associations (β_{ZY}) are estimated in a second dataset.

Two-sample MR



Two-sample MR



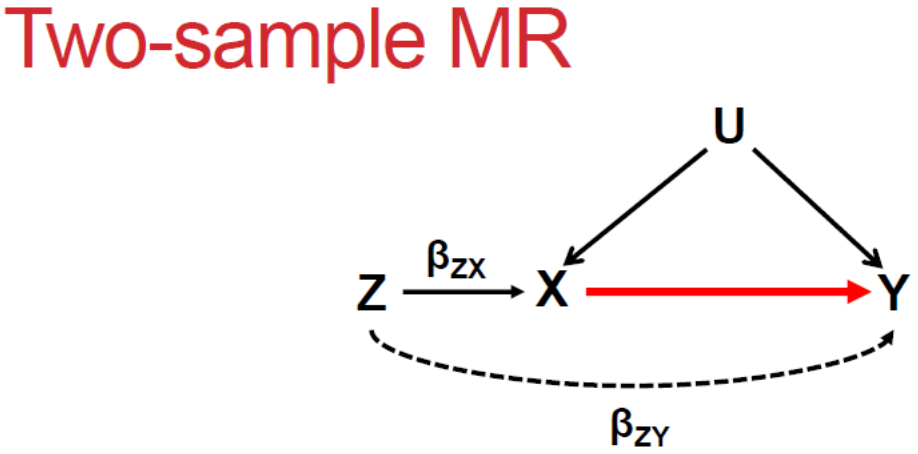
University of BRISTOL

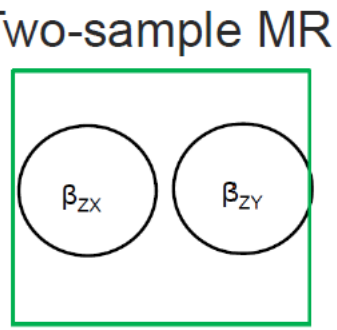
MRC Integrative Epidemiology Unit

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

Two-sample MR

- The analysis is usually performed using summary data.
- Summary data are genetic association estimates from regression of the exposure or outcome on a genetic variant.





University of BRISTOL

MRC Integrative Epidemiology Unit

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

Sources of summary data for two-sample MR

Complete summary data are currently publicly accessible for thousands of phenotypes.

- Genome-wide association studies (GWAS)
- Large GWAS meta-analysis
    - Consortium

Examples of consortia

Table 2 | Publicly available data sources for two sample Mendelian randomisation studies

Consortium name	Description	Most recent sample size
BCAC[24]	Breast cancer	256 123
CARDIoGRAMplusC4D[25]	Coronary artery disease and myocardial infarction	184 305
CKDGen[26]	Chronic kidney disease	111 666
DIAGRAM[27]	Diabetes	159 208
EAGLE[28]	Antenatal and early life and childhood phenotypes	47 541
EGG[29]	Early growth	153 781
GIANT[30]	Height, BMI, and other adiposity traits	693 529
GLGC[31]	Global lipids genetics consortium	331 368
ISGC[32]	Stroke	84 961
MAGIC[33]	Glucose and insulin related traits	224 459
PGC[34 35]	Psychiatric genetics, alcohol and tobacco, and other related traits	>500 000
SSGAC[36]	Educational attainment and wellbeing	293 723

Davies et al., 2017

Databases of GWAS results

Table 3 | Databases of genome-wide association study results

Data source	Description	Number of traits	Integrated with statistics package?
MR-Base	A curated database of genome-wide association study results with integrated R package for MR[23]	Over 1000	Yes
PhenoScanner	A curated database of genome-wide association study results with integrated R package for MR[37]	Over 500	Yes
GWAS catalog	Searchable database of genome-wide association study results[38]	Over 24 000	No

Davies et al., 2017

MR BASE 2-sample Mendelian Randomisation

University of BRISTOL | MRC Integrative Epidemiology Unit | CANCER RESEARCH UK

MR-base is a database and analytical platform for Mendelian randomization being developed by the MRC Integrative Epidemiology Unit at the University of Bristol.

You can either use the web application or our TwoSampleMR R package.

Data are also available through the MRC IEU OpenGWAS database.

Launch MR-Base webapp | R package | OpenGWAS database

Note - by clicking the "Launch MR-Base webapp" button you consent to the use of a cookie which enables us to ensure you have consented to the terms and conditions of data access. Information about how to control or delete cookies can be found at www.aboutcookies.org

MR-Base paper published

The MR-Base paper has now been published in eLife. See the publications page for details.

Telomeres paper published

Our paper reporting Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases has been published in Jama Oncology. See the publications page to access supporting data.

Advantages of two-sample MR

- Neither the exposure/risk factor nor the outcome needs to be measured in all studies, which is particularly useful if they are difficult or expensive to measure.
- It allows the summary results from GWAS to be used, which can be very large and thus increase the power.
- Summary results from several large consortia are publicly available for hundreds of thousands of variants.
- Transparency, reproducibility

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.

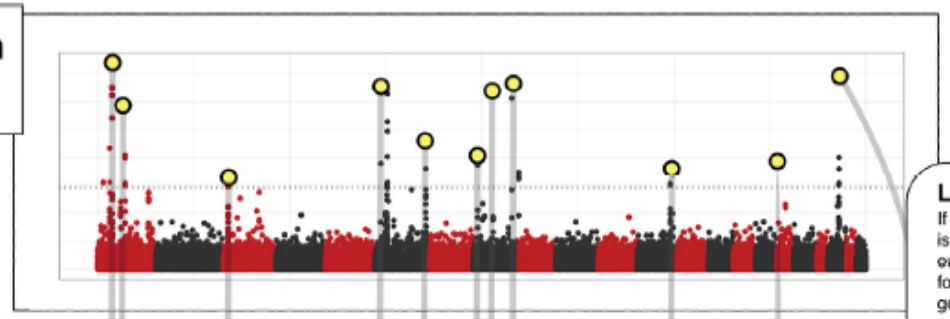
LD proxies

- 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

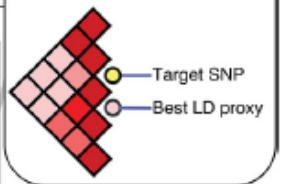
- Genetic associations with exposures and outcomes are typically reported per additional copy of a particular allele. Hence, when combining summarized data on genetic associations, it is important to ensure that genetic associations are expressed per additional copy of the same allele.
- This is particularly 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.**

1. Obtain instruments from exposure GWAS
2.
3. Extract SNP effects from outcome GWAS
4. Harmonise exposure and outcome effects
5. MR estimates and sensitivity analyses
6.



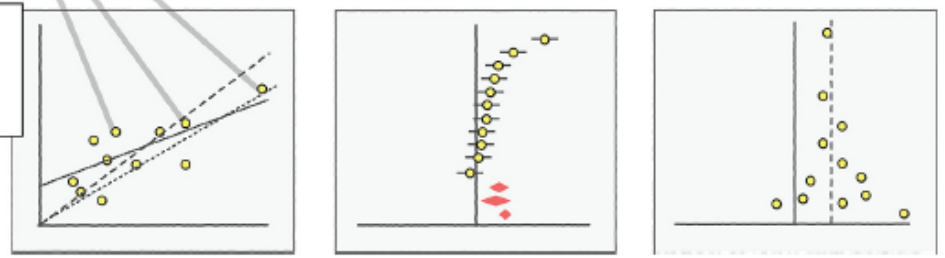
LD Proxies

If an exposure instrument is not available in the outcome GWAS then look for LD proxies in 1000 genomes



	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

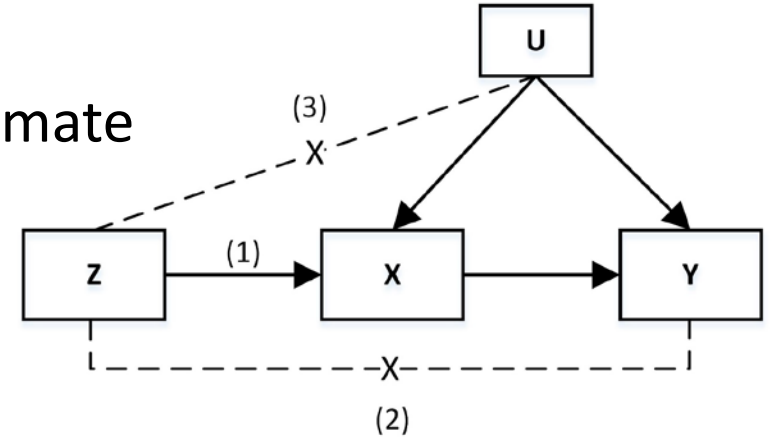


Hemani et al., 2018

5. Generate MR estimates

MR estimates: single instrument

- For a **single** instrument → Wald ratio /Ratio estimate



$$\beta = \frac{SNP - outcome\ association}{SNP - exposure\ association}$$

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{ZY}}{\hat{\beta}_{ZX}}$$

- Where both β's on the right hand side are regression coefficients

Assumption: no invalid instruments

Hypothetical example: LDL-c → CHD

Instrumental variable: randomization to *HMGCR* variant (rs12345)

Z → X: 0.05 mmol/L of LDLc per T allele

Z → Y: 0.03 log odds CHD per T allele

X → Y: 0.03/0.05=0.6 log odds CHD per 1 mmol/L of LDLc (OR=1.82)

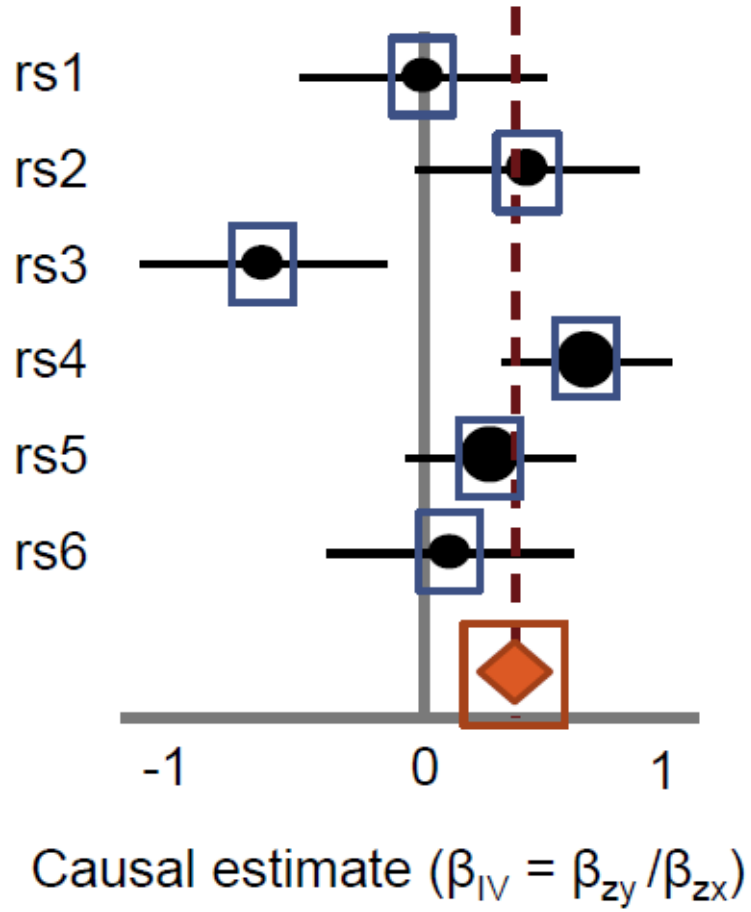
Multiple instruments: Inverse variance weighted (IVW) method

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

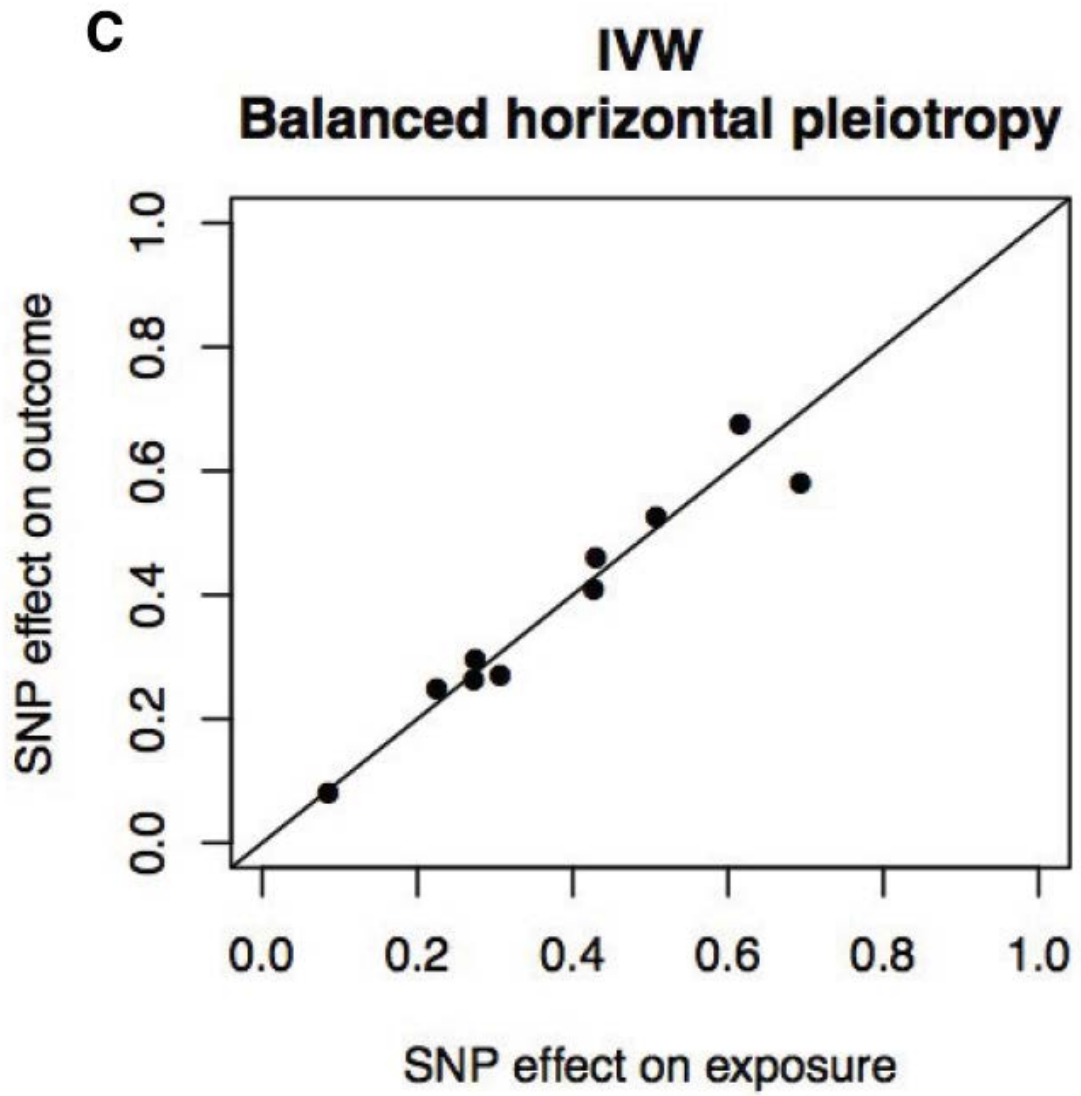
$$\widehat{\beta}_{\text{IVW}} = \frac{\sum_{j=1}^{L} w_j \widehat{\beta}_j}{\sum_{j=1}^{L} w_j}, \; w_j = \frac{\widehat{\gamma}_j^2}{\sigma_{Yj}^2}$$

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



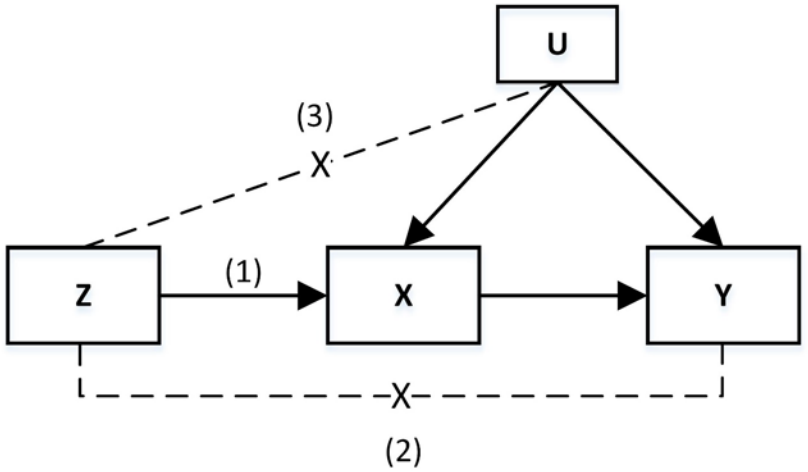
Hemani et al., 2018

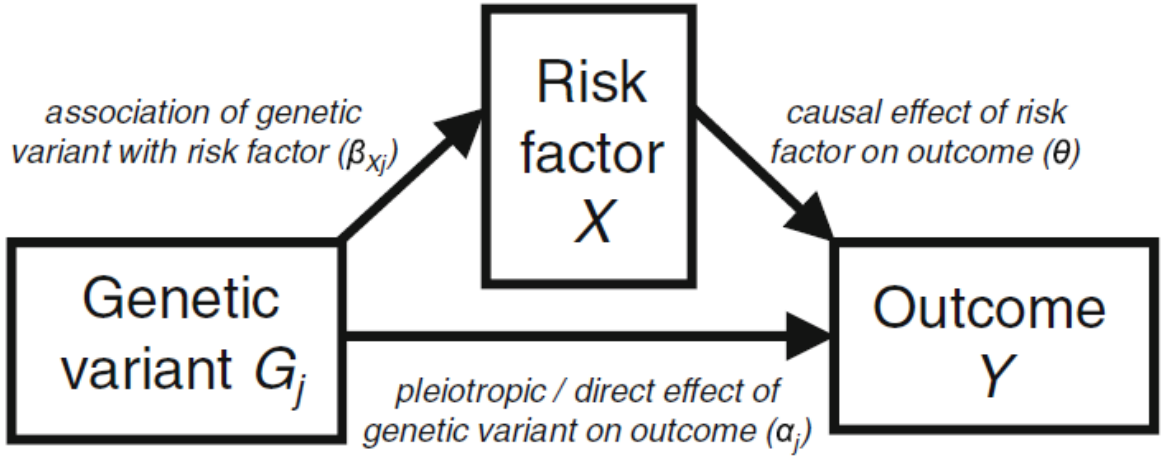
Inverse variance weighted (IVW) method

- 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





Horizontal pleiotropy

Davies et al., 2017

6. Perform sensitivity analysis

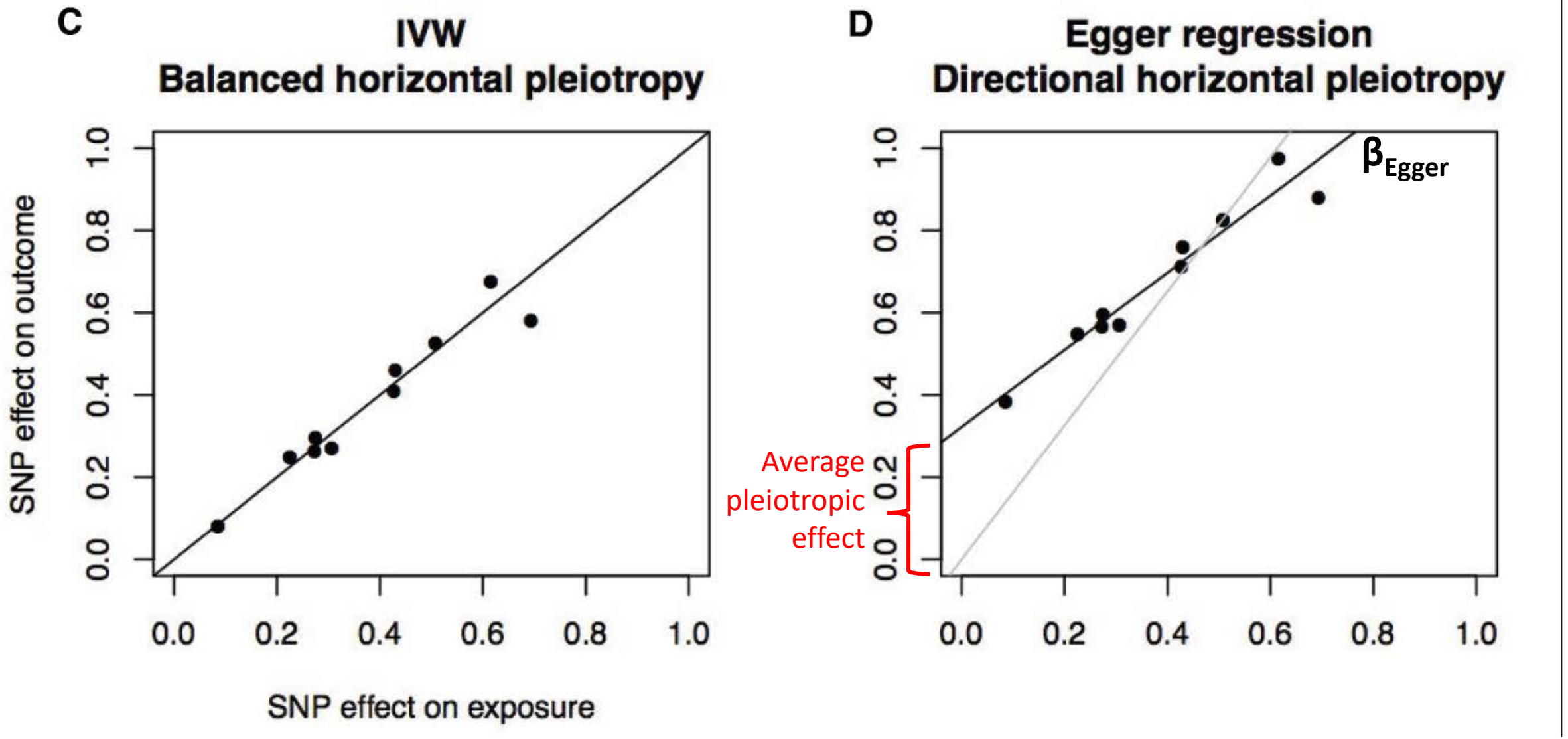
- If the genetic variants have pleiotropic effects on the outcome, IVW causal estimates will be biased.
- Use other robust analysis methods that can provide valid causal inferences under weaker assumptions than the standard IVW method.

Other MR methods used for sensitivity analysis

- MR Egger
- Median-based estimator
- Mode-based estimator
- Etc.

MR-Egger

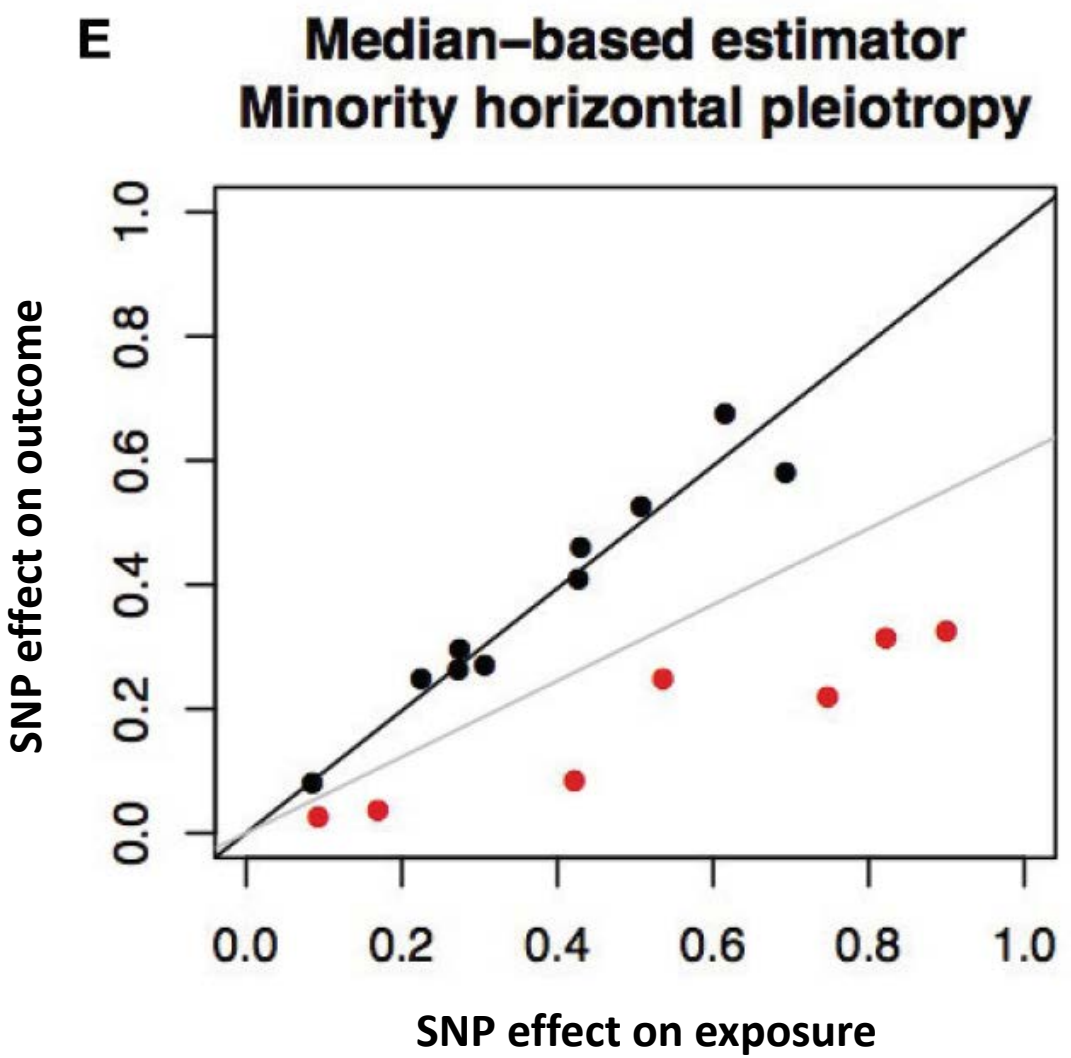
- MR-Egger gives an unbiased causal effect estimate when the genetic variants are invalid instrumental variables.
  - Under the assumption that the association of each genetic variant with the exposure is independent of the pleiotropic effect of the variant
- MR Egger allows a non-zero intercept.
- MR Egger intercept: average directional pleiotropic effect across the set of variants
- MR Egger slope: an estimate of the causal effect corrected for pleiotropy



Hemani et al., 2018

Median-based estimator

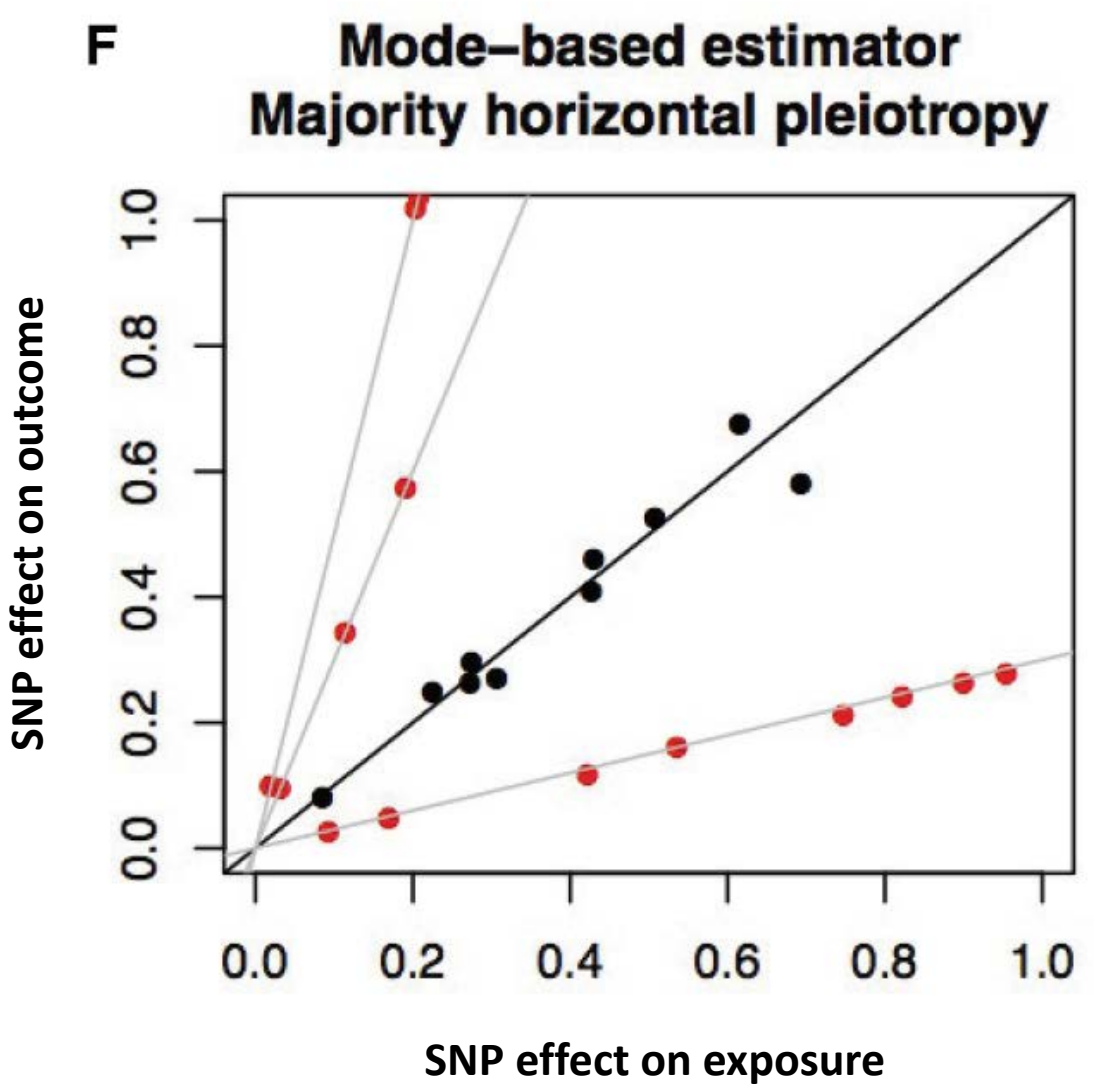
- The median-based estimator provides an unbiased causal estimate when the majority of SNPs are valid instruments.
- It takes the median (or weighted median) of all IV causal estimates.
- This estimator is consistent when at least 50% of the instrumental variables are valid.



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

Limitations of two-sample MR

- The two samples should be from the same underlying population and the same ethnic group.
  - Different minor allele frequency and linkage disequilibrium in different ethnic groups
- The two samples could differ according to population characteristics, e.g., age, sex, socio-economic background.
  - Such differences can affect the validity of causal inferences.
- Sample overlapping
  - As several large consortia have overlapping studies, participants may overlap between the datasets used to estimate the genetic associations with the exposure and outcome.
  - The bias varies linearly depending on the degree of overlap.

References

1. Hemani G, Zheng J, Elsworth B et al. The MR-Base platform supports systematic causal inference across the human phenome. eLife. 2018;7.
2. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27(8):1133-1163.
3. Davies NM, Holmes MV and Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ. 2018;362:k601.
4. Burgess S, Davey Smith G, Davies N, et al. Guidelines for performing Mendelian randomization investigations. In: Wellcome Open Res; 2019.

Thank you for your attention.