Introduction to Mendelian randomisation

Gibran Hemani

Outline

Session 1:

- Context, basic principles and assumptions of MR
- How to perform and interpret MR analyses
- What can go wrong
- Sensitivity analyses

Session 2:

- Practical

Causality

- Modification of X will lead to Y
  - X is in the past and Y is in the future
  - X might not be modifiable
  - X might be one of many factors influencing Y
  - The state of Y might be probabilistic – X only influences risk
  - X might be causal but with a small effect

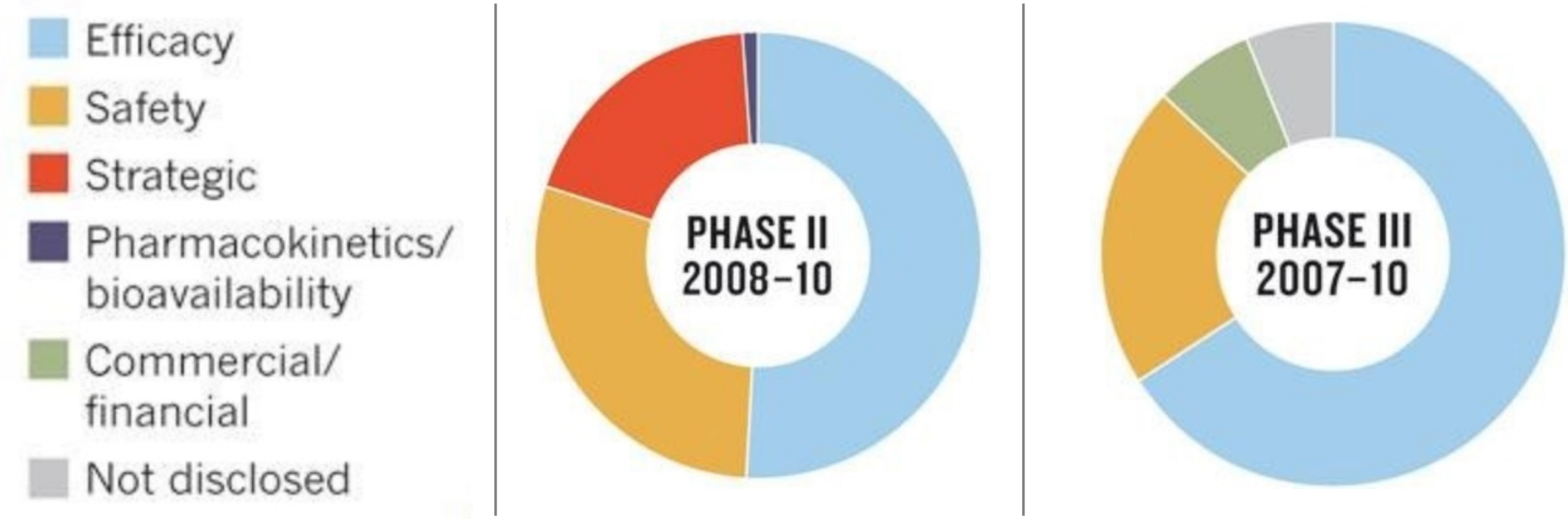
- Counterfactual
  - If we had two universes in which everything was identical except the state of X, how would Y differ?

Background

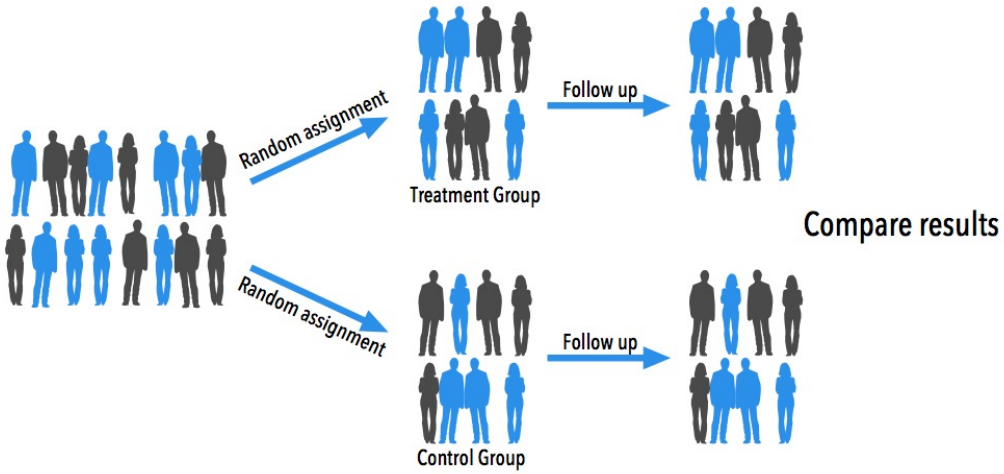
Study designs in epidemiology


Today's Random Medical News
from the New England Journal of Panic-Inducing Gobbledygook
JIM BORGMAN
EXERCISE
FATTY FOODS
SMOKING
STRESS
COFFEE
RED WINE
COMPUTER TERMINALS
DAYCARE
CAN CAUSE
HEART DISEASE
BREAST CANCER
SPONTANEOUS REMISSION
HYPO-THERMIA
GLAUCOMA
A FEELING OF WELL-BEING
SEXUAL DYSFUNCTION
DEPRESSION
IN
CHILDREN
TWINS
ARTHRITIS SUFFERERS
TWO-INCOME FAMILIES
7 OUT OF 10 WOMEN
MEN 25-40
OVER-WEIGHT SMOKERS
RATS
ACCORDING TO A REPORT RELEASED TODAY....
NEWS

Vast majority of pharmaceutical compounds fail clinical studies…



…due to the drug target not having the expected outcome



Randomised control trials

Experimental design (not observational)

- Recruit samples
- Randomly assign an intervention (X) to half
- Provide placebo to the other half

The only difference between the groups is whether or not they had the intervention (X)

If the outcome (Y) is different between groups then this must be due to the intervention

Approximating the counterfactual framework

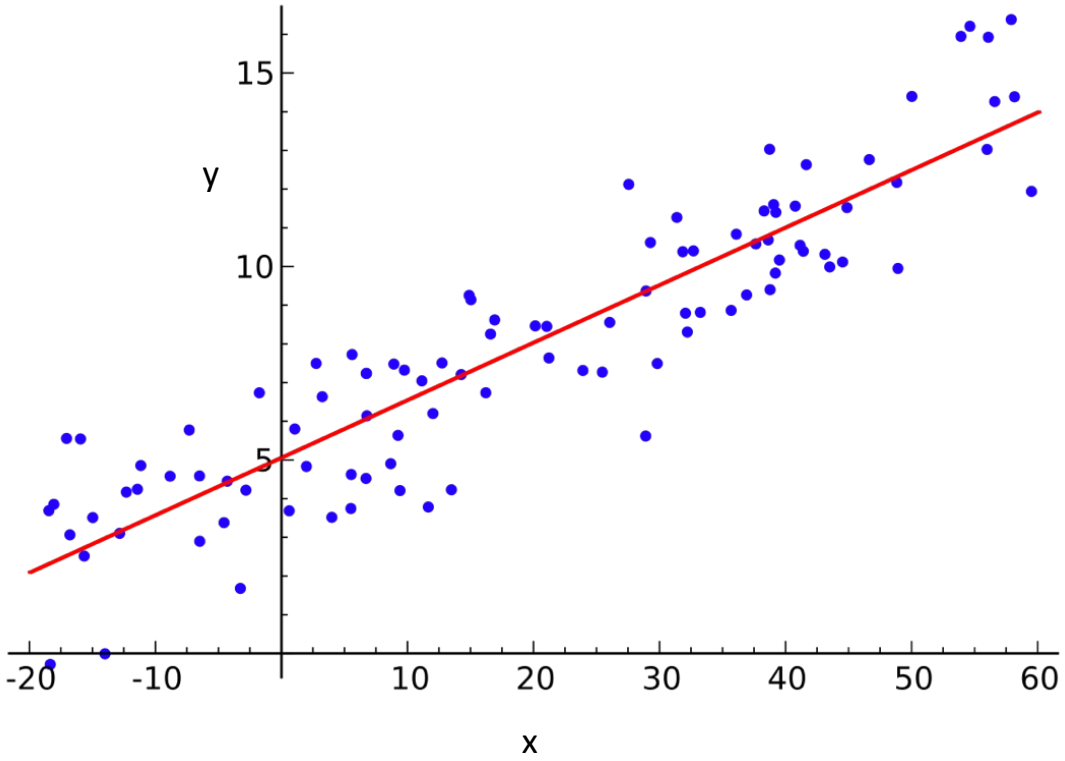
Problems with RCT

- Very expensive
- Not every intervention is ethical
- Not every intervention is practicable
- Could require very long follow up time

What data do we have?

- Many different study designs, but we will focus on **observational studies**
- “Observational” data – sampling from a population
- What are the methods for and limitations of making causal inference from observational data?

Linear regression



Slope is the association of X with Y

What can give rise to this slope?

- X causing Y
- Y causing X
- A confounder influencing X and Y

Similarly – if we try to estimate influence of X on risk of some binary outcome There could be a confounder
- Some variable associates with X, and is differential between cases and controls

Matched controls

Treated units			
Age	Gender	Months unemployed	Secondary diploma
19	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Untreated units			
Age	Gender	Months unemployed	Secondary diploma
24	1	8	1
38	0	2	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

1. Start with observational data
2. Only compare cases and controls using a subset of samples

- The measured characteristics between cases and controls are the same
- The only thing that is allowed to vary (i.e. not matched) is the hypothesized explanatory variable (X)

This is analogous to adjusting for covariates in logistic regression

Problem: there might be unknown or imperfectly measured confounding

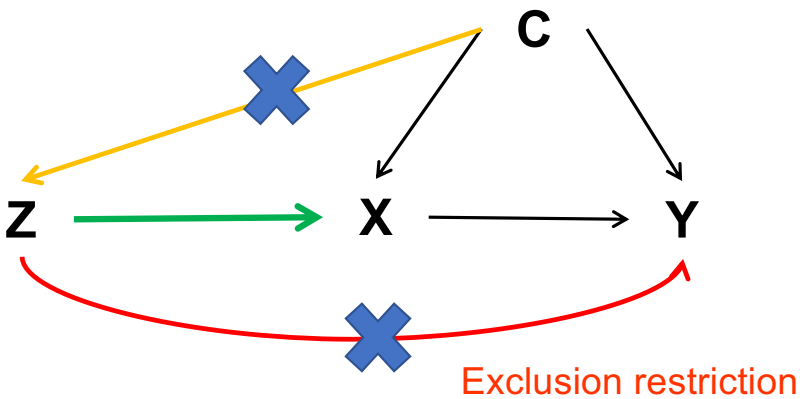
Can we better approximate the counterfactual framework using observational data?

Intro to Mendelian randomization

Instrumental variables (IVs)

- Statistical method from econometrics
- An instrumental variable (Z) is one that:

1. Associates with the exposure
2. Does not associate with confounders
3. Only influences the outcome through the exposure (exclusion restriction principle)



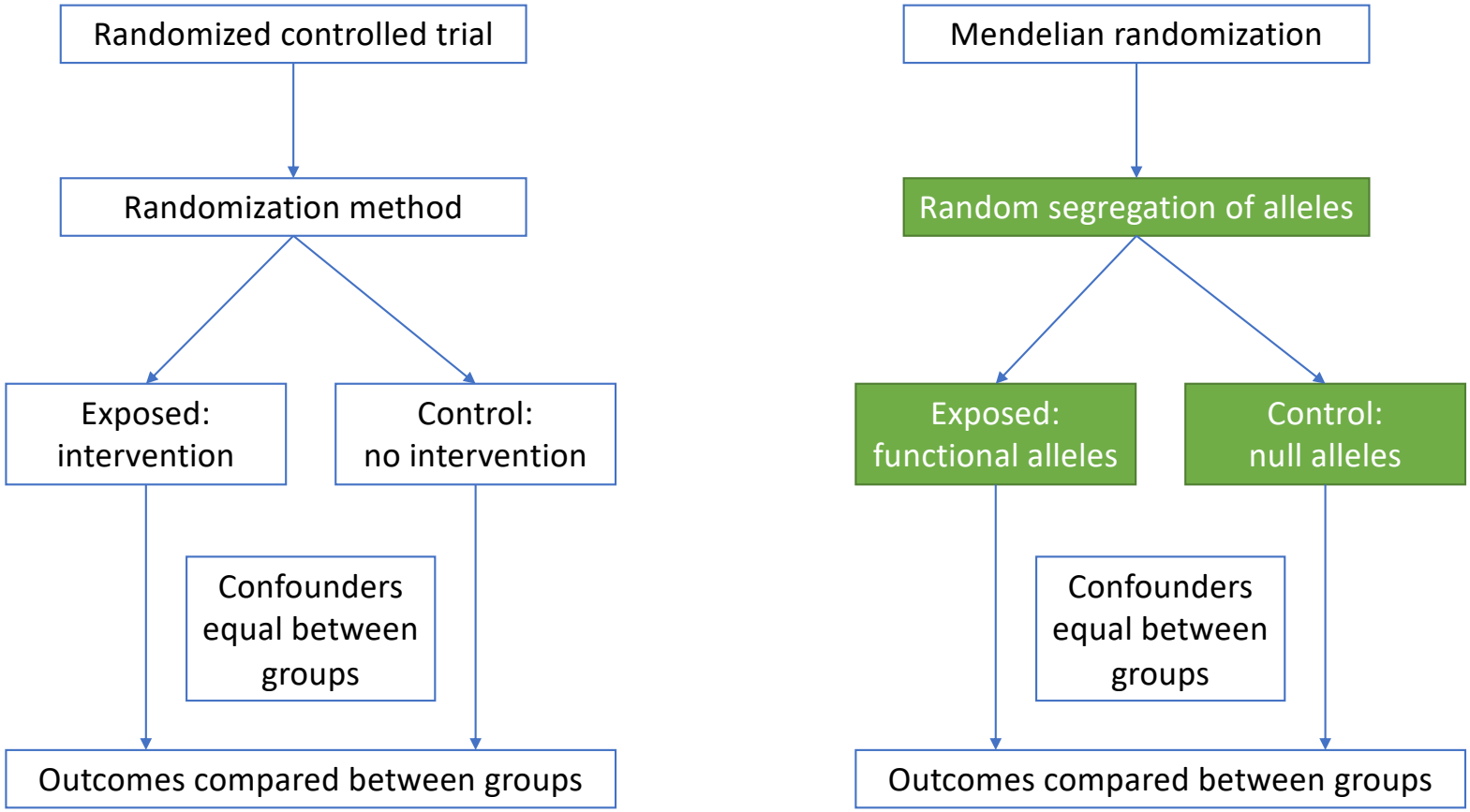
Z is not independent of X
Z is independent of C
Z is independent of Y, given X & C

Why are genetic associations special?

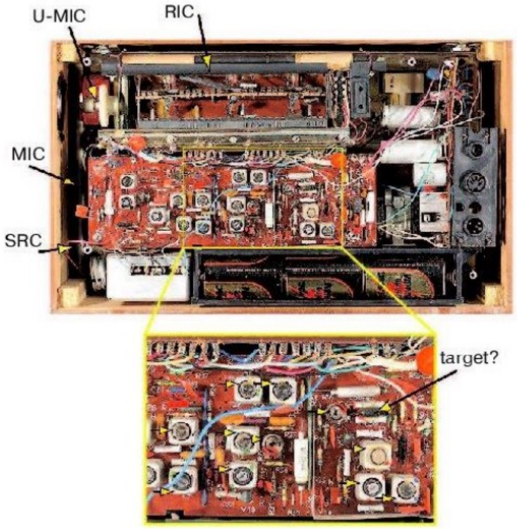
- Due to Mendel's laws
  - Law of segregation: inheritance of an allele is random and independent of environment etc
  - Law of independent assortment: genes for different traits segregate independently (assuming not in LD)

- The direction of causality is known – always from SNP to trait
- Genetic variants are **potentially** very good instrumental variables
- Using genetic variants as IVs is a special case of IV analysis, known as Mendelian randomization

Mendelian randomization analogous to RCTs



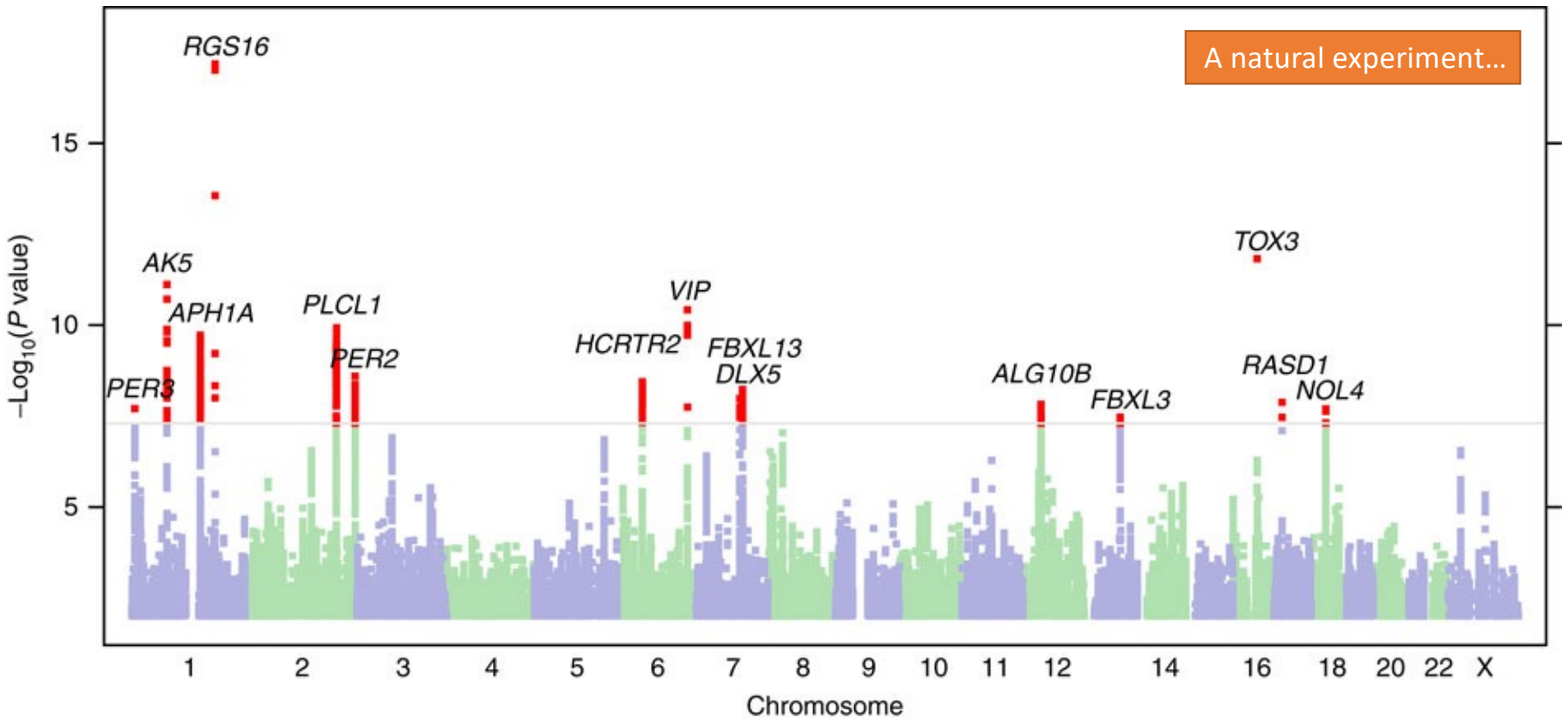
Can a biologist fix a radio?



Lazebnik 2004

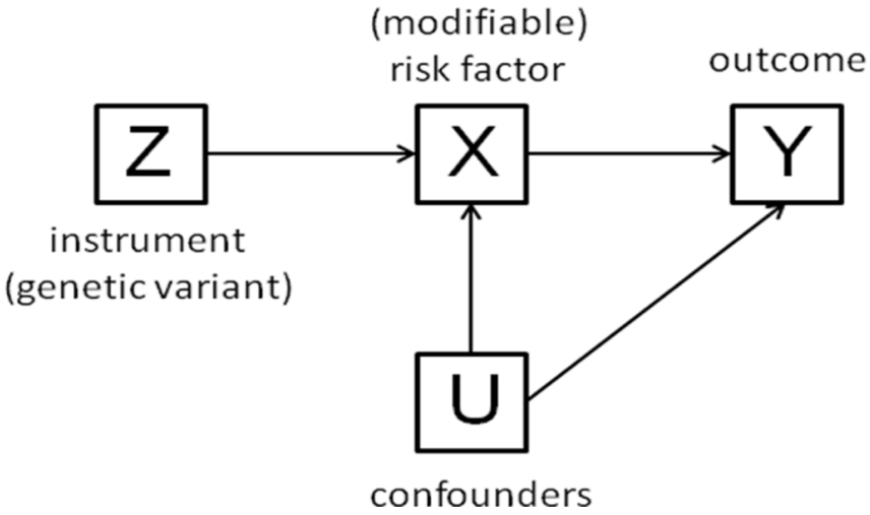
1. Obtain 1000 radios
2. Break components at random
3. Record effect on radio operations
4. Figure out what each component does

Works in model organisms but
Can't really do this to humans…


A natural experiment…
RGS16
AK5
APH1A
PER3
PLCL1
PER2
HCRTR2
VIP
FBXL13
DLX5
ALG10B
FBXL3
TOX3
RASD1
NOL4
15
10
5
−Log10(P value)
1
2
3
4
5
6
7
8
9
10
11
12
14
16
18
20
22
X
Chromosome

How to use genetic associations to infer causality

- If X causes Y then everything that causes X should also associate with Y
  - Given we have sufficient power
- If Z is a valid instrument for X then if it associates with Y then X causes Y.



Mendelian randomisation using individual level data

- Need: X, Y and instrument(s) for X
- Perform two-stage least squares

1. Create a genetic predictor for X

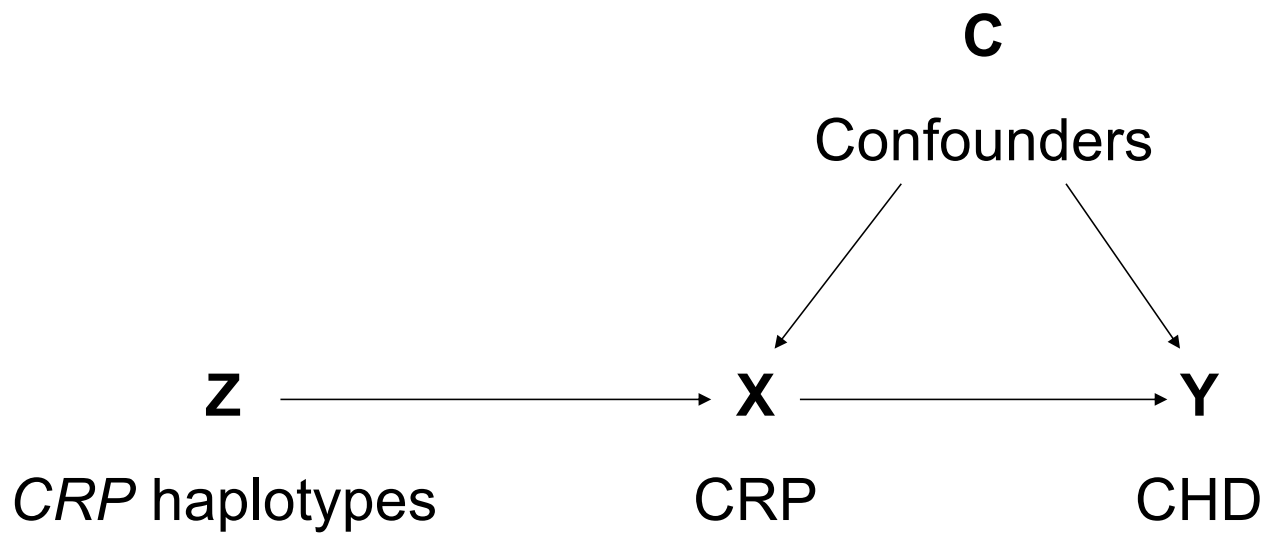
$$x_i = \gamma z_i + \varepsilon_i$$
$$\hat{x}_i = \hat{\gamma} z_i$$

2. Regress the predictor against Y

$$y_i = \beta \hat{x}_i + \epsilon_i$$

Now β is the asymptotically unbiased estimate of the effect of x on y

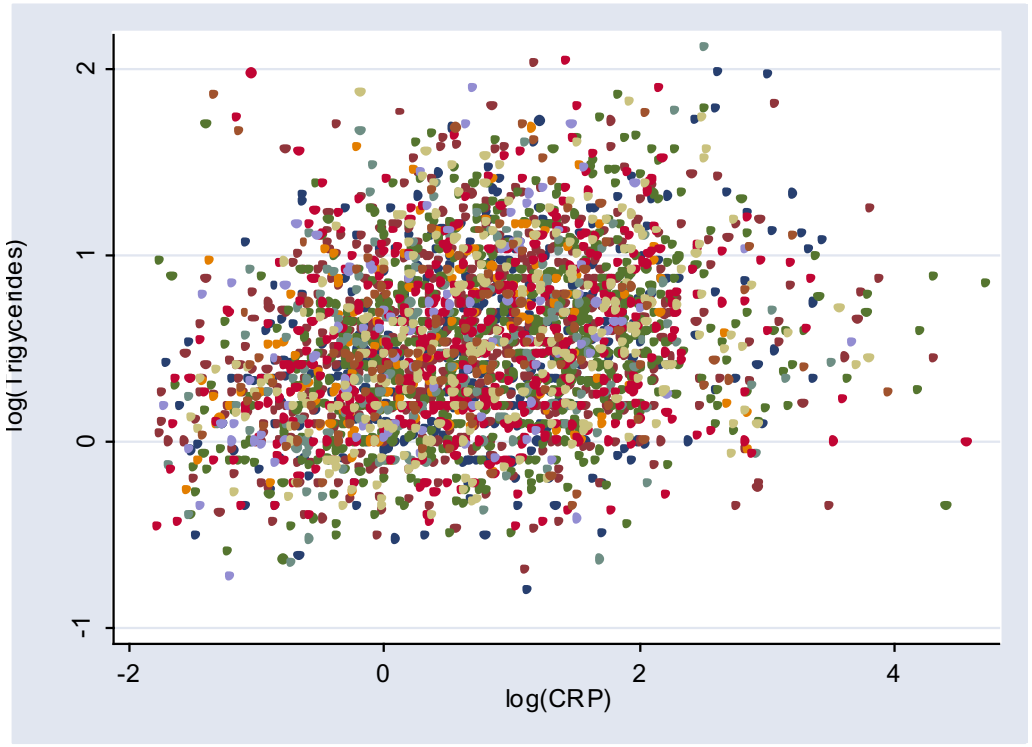
Example: genes, CRP and CHD (risk factors)



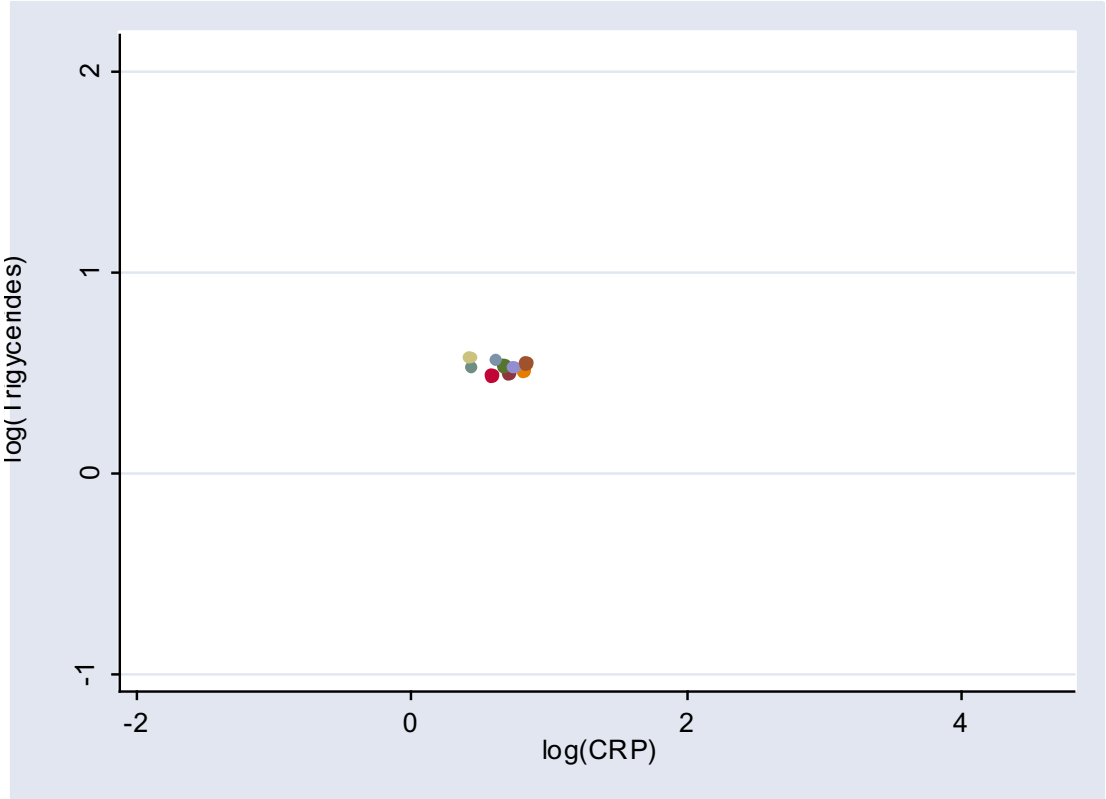
Timpson N, et al. Lancet 2005

Observational association

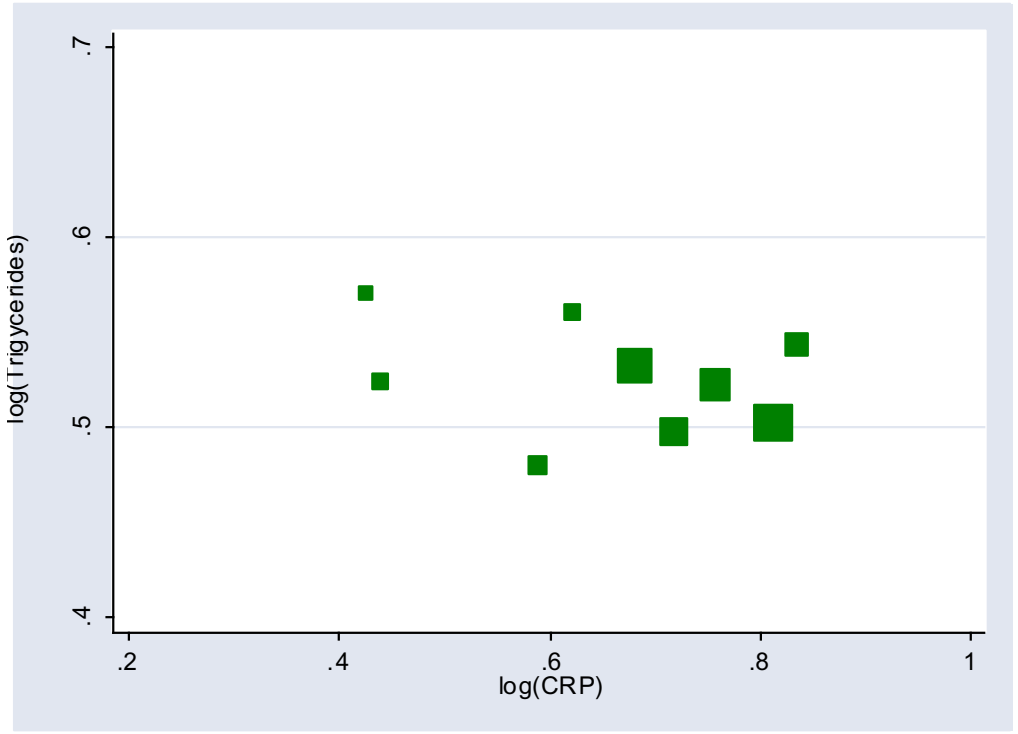
… by genotype



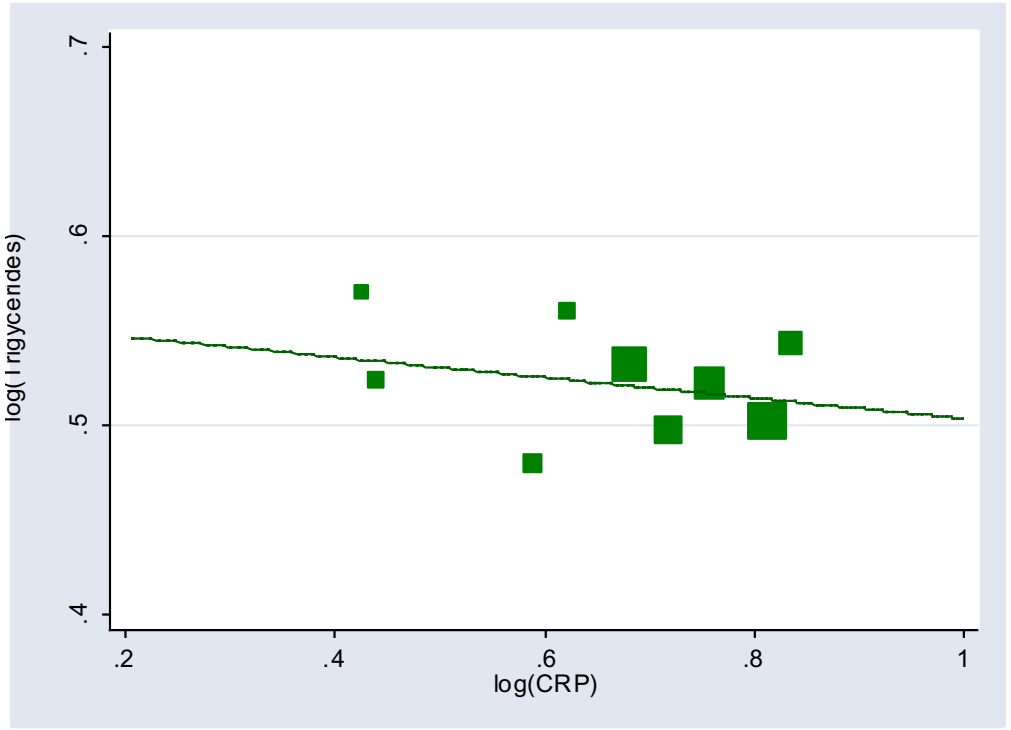
Genotype means



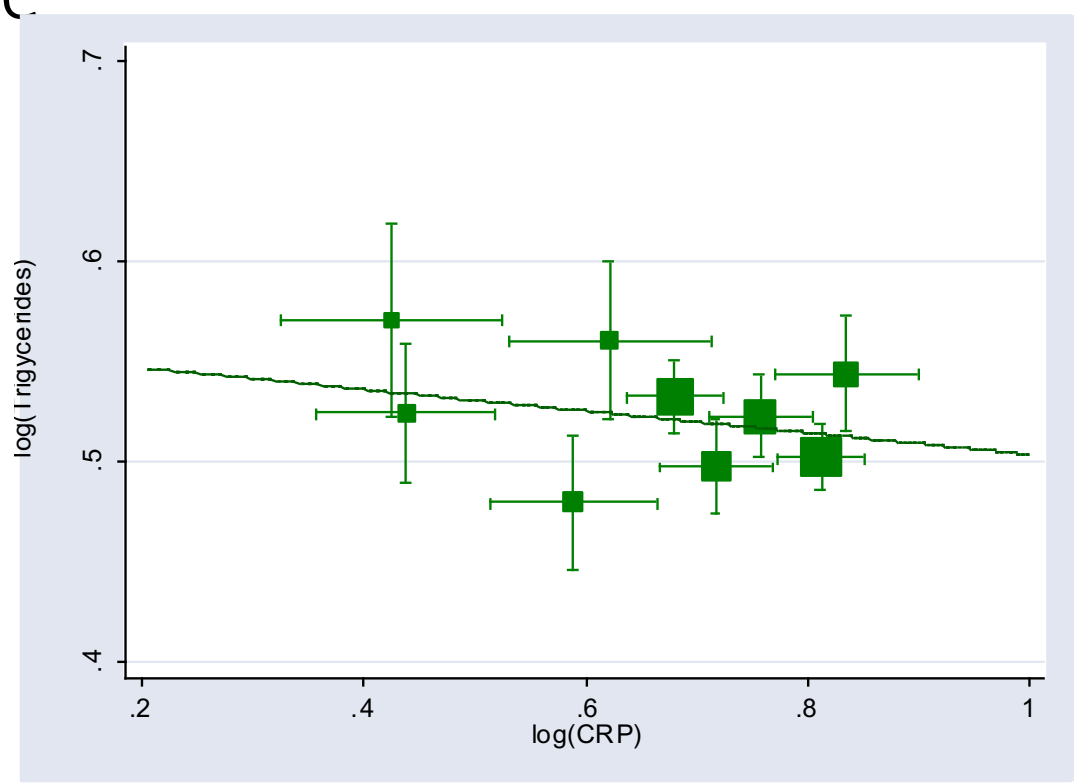
Zoom in…



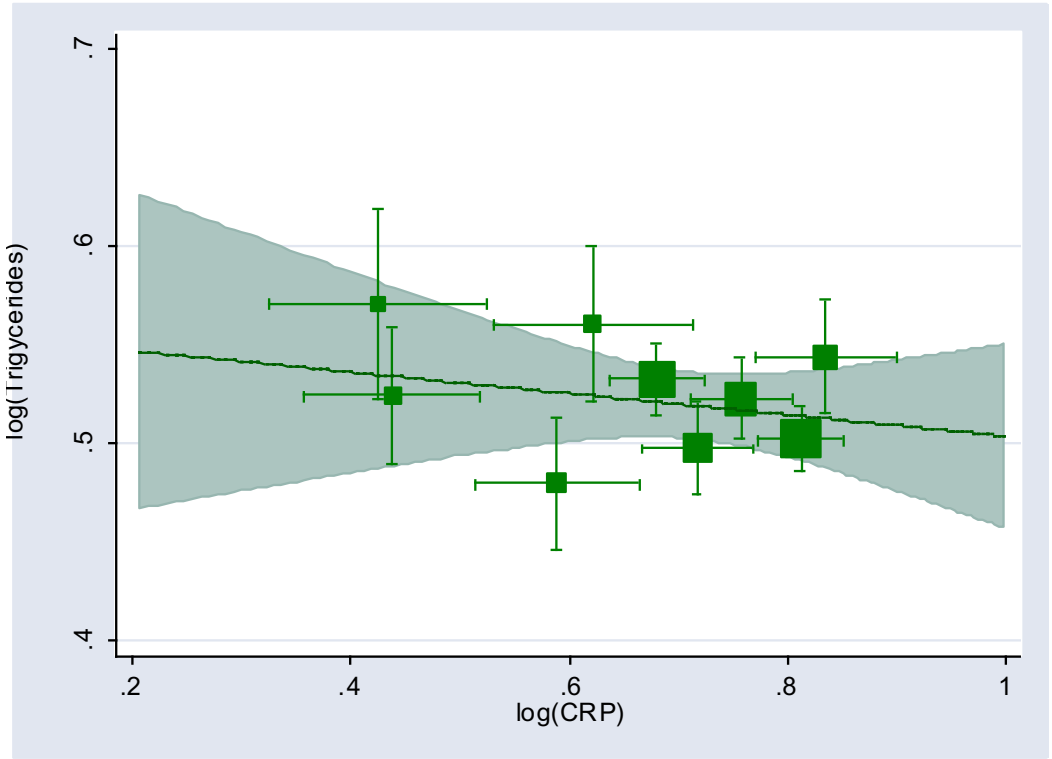
Add weighted regression line



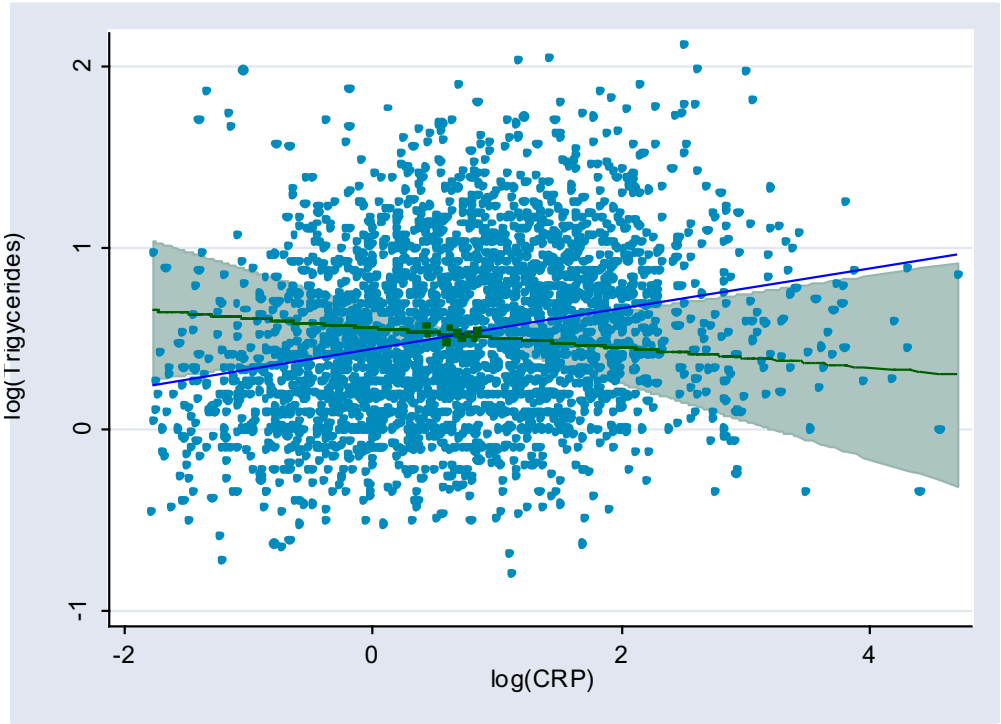
+ standard errors of means+ fitted line



+ CI for fitted line



Zoom out and put it all together!



Comparison of IV and OLS results

Ratio of values with doubling of CRP:

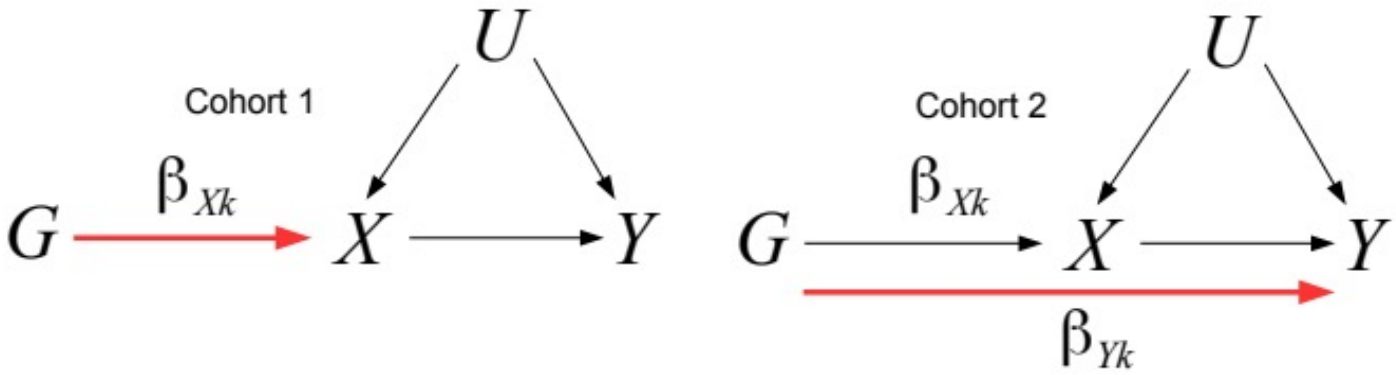
Linear regression	1.08 (1.07, 1.09)
Instrumental variables	0.99 (0.92, 1.08)

Comparison (test for endogeneity):
P = 0.031

MR can also be performed using just the results from GWAS

- Also known as two-sample MR, SMR, or MR with summary data etc
- Advantages:
  - The data is readily available, non-disclosive, free, open source
  - The exposure and outcome might not be measured in the same sample
  - The sample size of the outcome variable, key to statistical power, is not limited by requiring overlapping measures of the exposure
- Disadvantages:
  - Some extensions of MR not possible, e.g. non-linear MR, use of GxE for negative controls, various sensitivity analyses

Mendelian randomisation using summary data

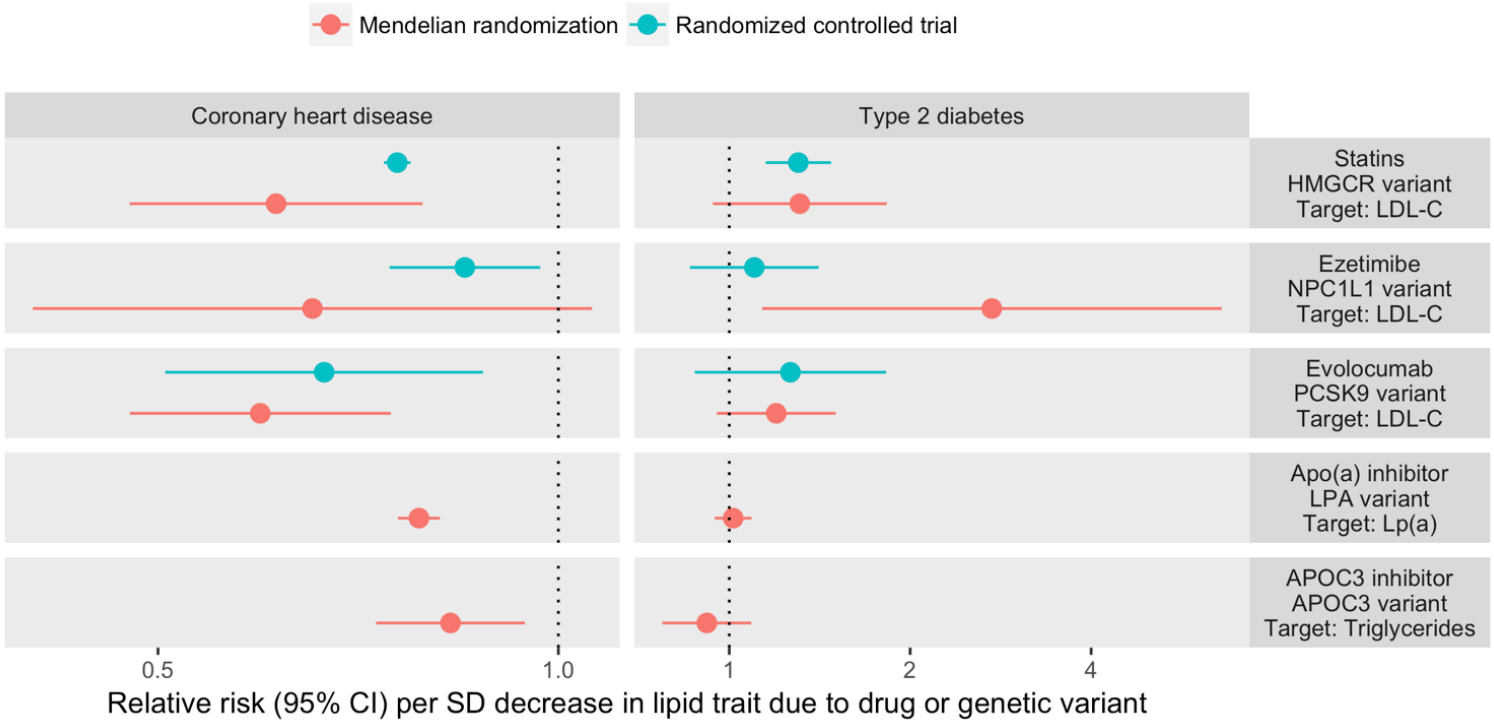


Causal estimate obtained from Wald ratio using SNP k $\hat{\beta}_k = \hat{\beta}_{Yk} / \hat{\beta}_{Xk}$

If there are multiple SNPs, obtain **overall** causal estimate by meta analysing the Wald ratios from each SNP.

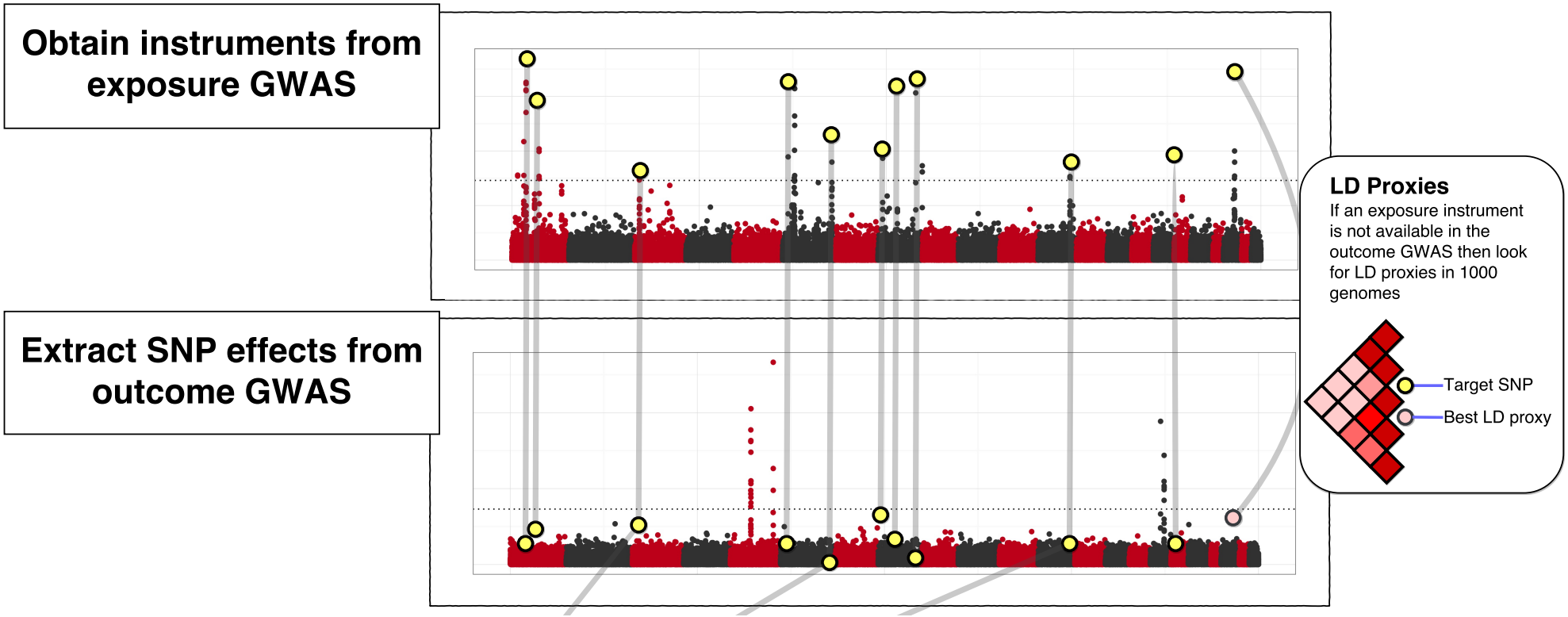
Each SNP is used like an independent trial

Predicting drug RCTs using MR



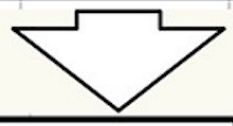
MR in practice

How to perform Mendelian randomization using only GWAS summary results

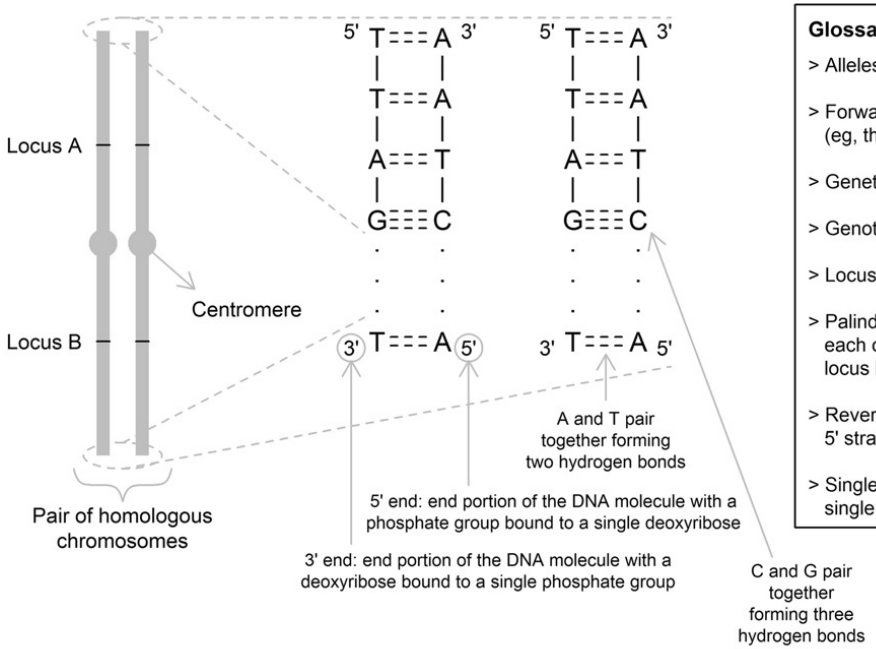


Harmonise exposure and outcome effects

	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**



Glossary

> Alleles: variant forms that a locus may present.

> Forward or positive strand: the DNA strand that is read from the 5' to the 3' end (eg, the 5' TTAG...T 3' strand in the figure).

> Genetic variant: locus with more than one allele in a population.

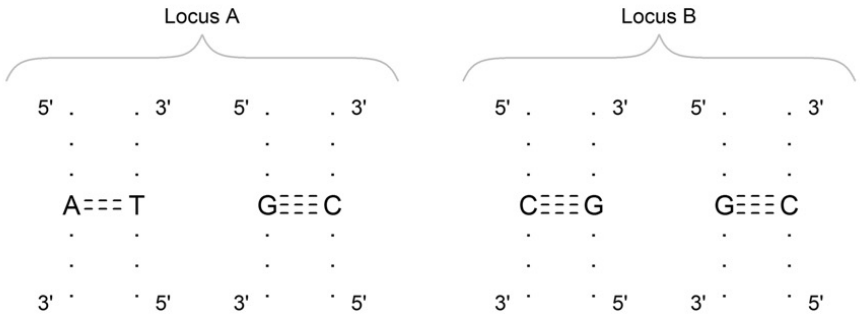
> Genotype: combination of alleles that an individual presents at a given locus.

> Locus (plural loci): a specific location in a DNA sequence.

> Palindromic SNP: SNPs whose alleles correspond to nucleotides that pair with each other in a double-stranded DNA molecule. SNPs with A/T or G/C (as in locus B below) alleles are palindromic SNPs.

> Reverse or negative strand: the DNA strand that is read from the 3' to the 5' strand (eg, the 3' AATC...A 5' strand in the figure).

> Single nucleotide polymorphism (SNP): a type of genetic variant that involves single base pair changes.



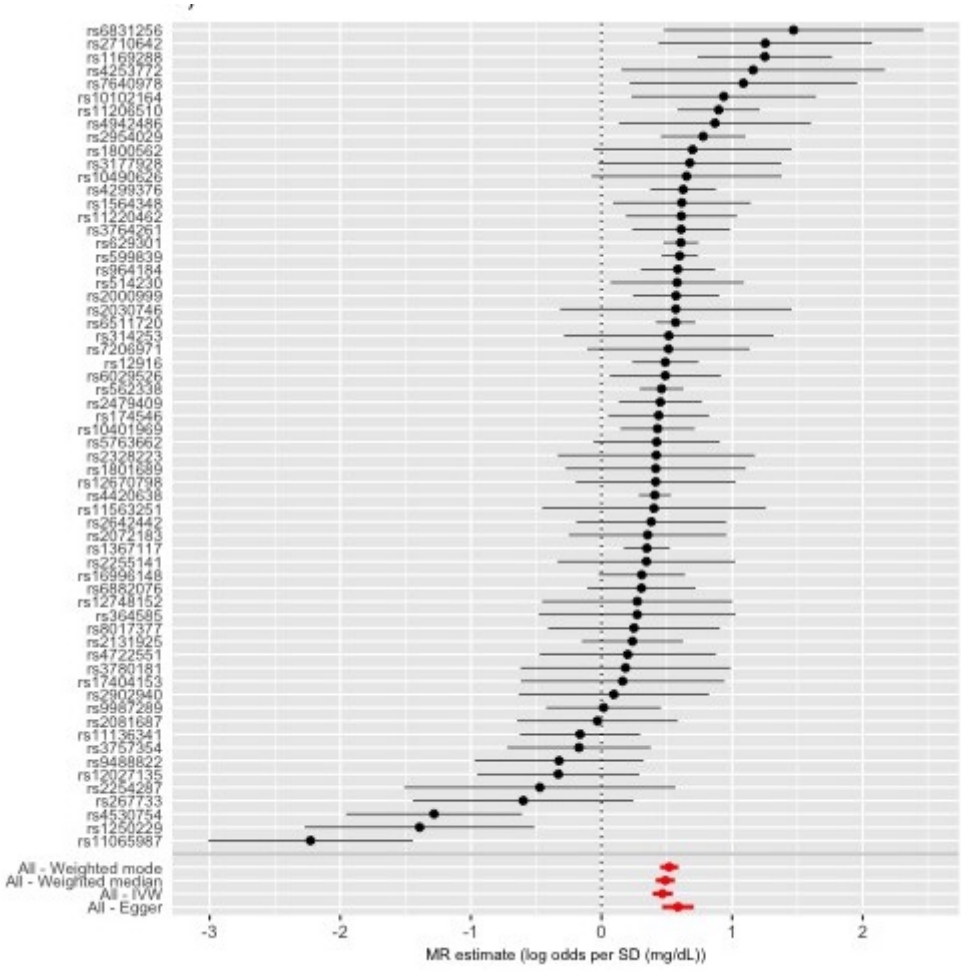
	Locus A	Locus B
Type of genetic variation	Single nucleotide polymorphism	Single nucleotide polymorphism
Alleles (5' to 3')	A and G	C and G
Alleles (3' to 5')	T and C	G and C
Genotype (5' to 3')	AG	CG
Genotype (3' to 5')	TC	GC
Palindromic variant	No	Yes

Hartwig et al 2017

Strand issue exercise

SNP	Study 1 alleles	Study 1 allele freq	Study 2 alleles	Study 2 allele freq	Verdict
rs1	A/G	0.2	A/G	0.2	
rs2	G/T	0.3	T/G	0.72	
rs3	G/C	0.65	G/C	0.62	
rs4	A/T	0.49	A/T	0.50	
rs5	A/T	0.12	A/T	0.89	
rs6	A/G	0.4	A/T	0.4	

Forest plot example for LDL cholesterol and CHD



Fixed effects meta analysis (aka inverse variance weighted, IVW)

- There is one underlying ‘true’ effect
- All deviations of sample effects from the ‘true’ effect are due to chance

Inverse variance weighted (IVW) fixed effects method

For N studies, each study *i* contributes more to the meta analysis if its standard error is lower

$$\mathrm{var}(\beta_i) = se\ (\beta_i)^2$$

$$se\ (\beta_i) = \frac{SD_i}{\sqrt{n_i}}$$

$$w_i = \frac{1}{\mathrm{var}(\beta_i)}$$

$$\beta_{pooled} = \frac{\sum_{i=1}^{N}(w_i * \beta_i)}{\sum_{i=1}^{N}(w_i)}$$

$$se_{pooled} = \sqrt{\frac{1}{\sum_{i=1}^{N}(w_i)}}$$

Calculate p-value

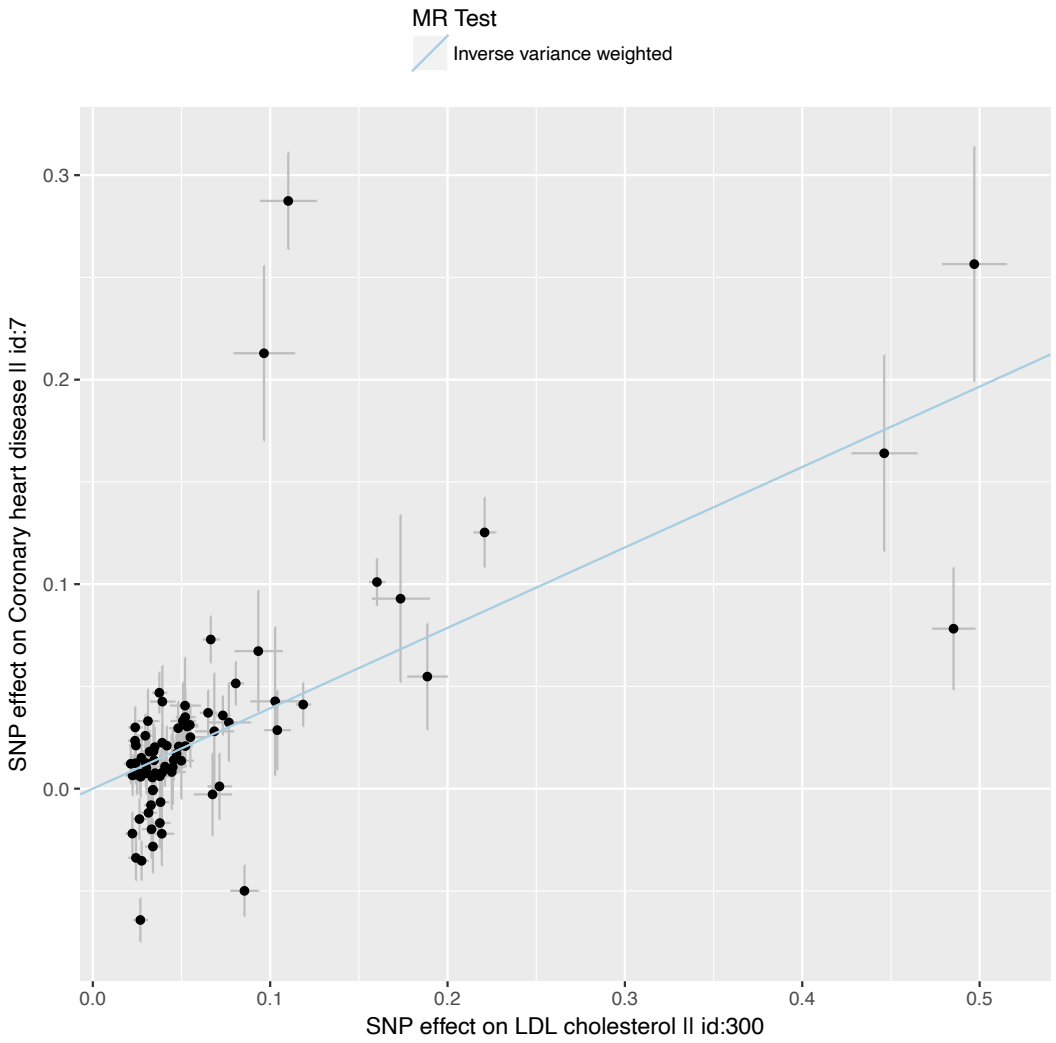
$$\chi^2_{df=1} = \frac{\beta_{pooled}^{\ 2}}{se_{pooled}^{\ 2}} = \frac{(\sum_{i=1}^{N} w_i * \beta_i)^2}{\sum_{i=1}^{N} w_i}$$

$$z = \frac{\beta_{pooled}}{se_{pooled}} = \frac{\sum_{i=1}^{N} w_i * \beta_i}{\sqrt{\sum_{i=1}^{N} w_i}}$$

Look up or compute the associated p-value

P=0.05 → χ^2=3.84
Z=1.96

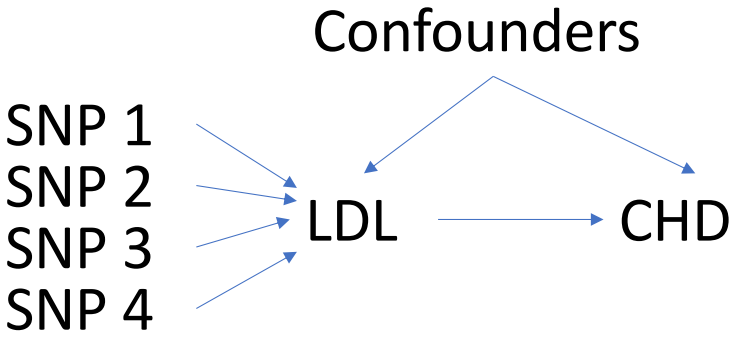
P=0.001 → χ^2=10.83
Z=3.29



IVW is equivalent to a weighted regression of SNP-exposure against SNP-outcome effects.

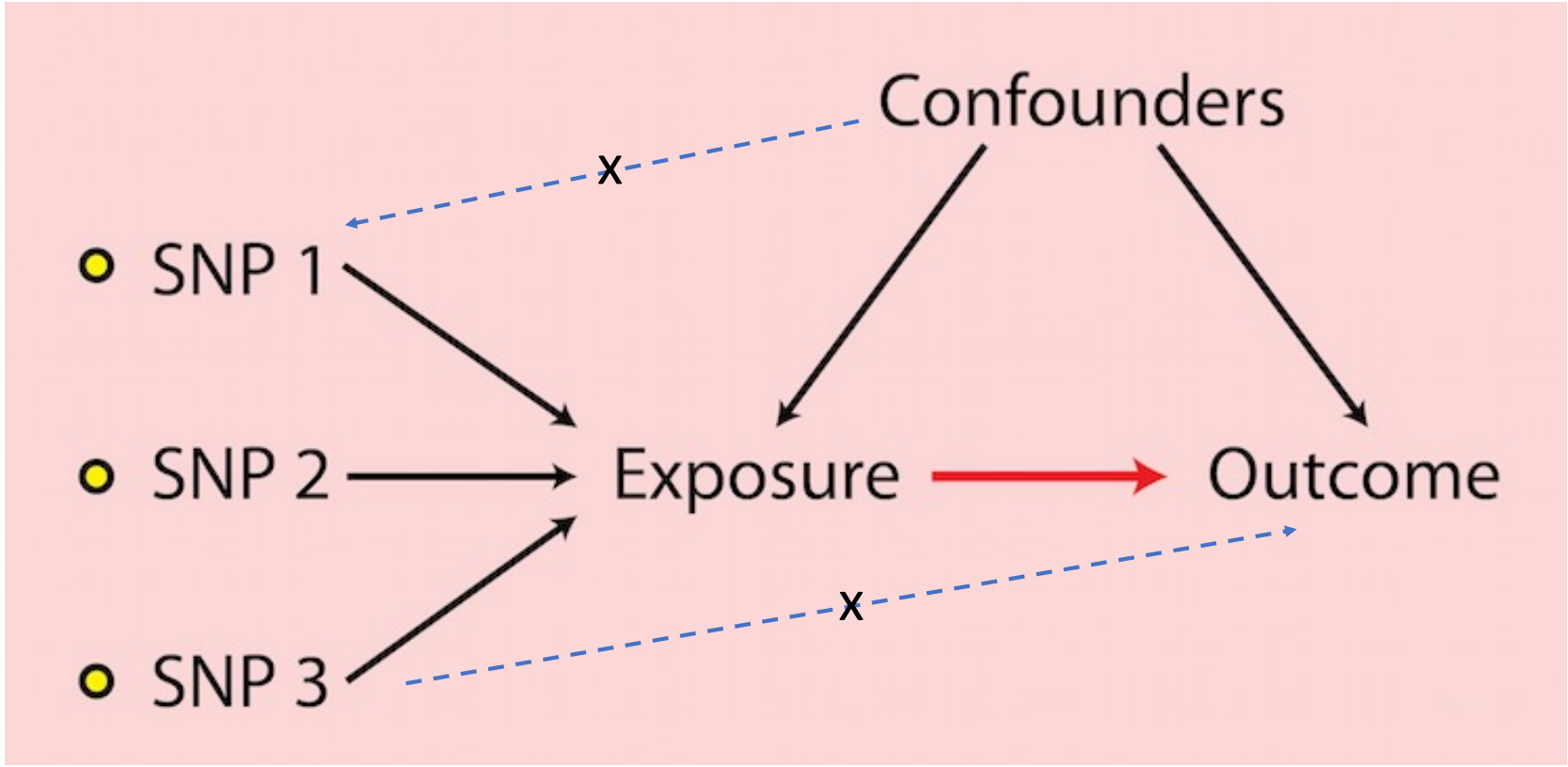
The weights are 1/se_outcome

The slope is the estimate of the causal effect



Invalid instruments

Pleiotropy, reverse causation


Confounders
x
SNP 1
SNP 2
SNP 3
Exposure
Outcome
x

Pleiotropy

One genetic variant associates with multiple traits

Vertical pleiotropy

The variant influences one trait through another trait

- This is the mechanism that enables MR

Horizontal pleiotropy

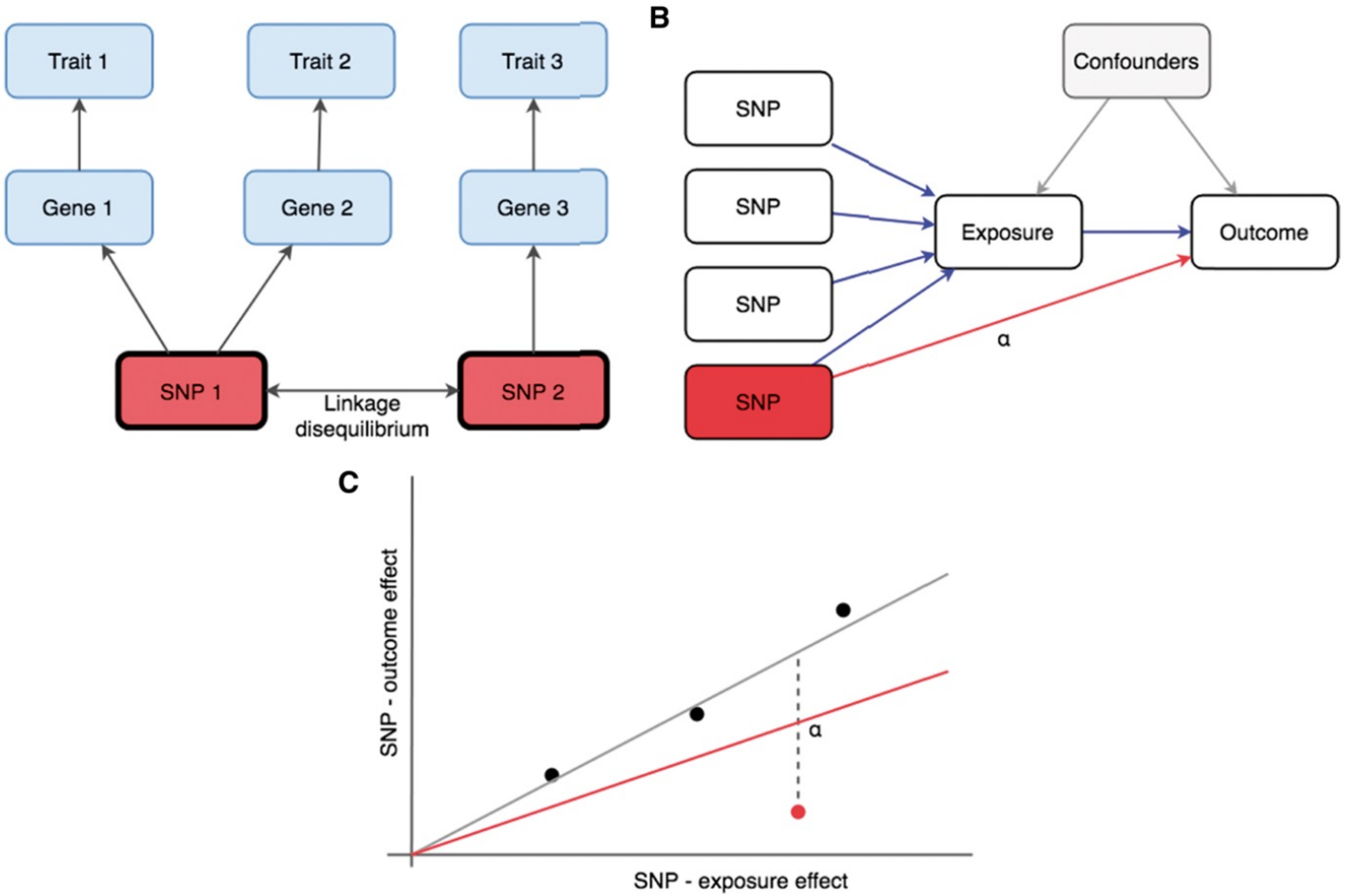
The variant influences two traits through independent pathways

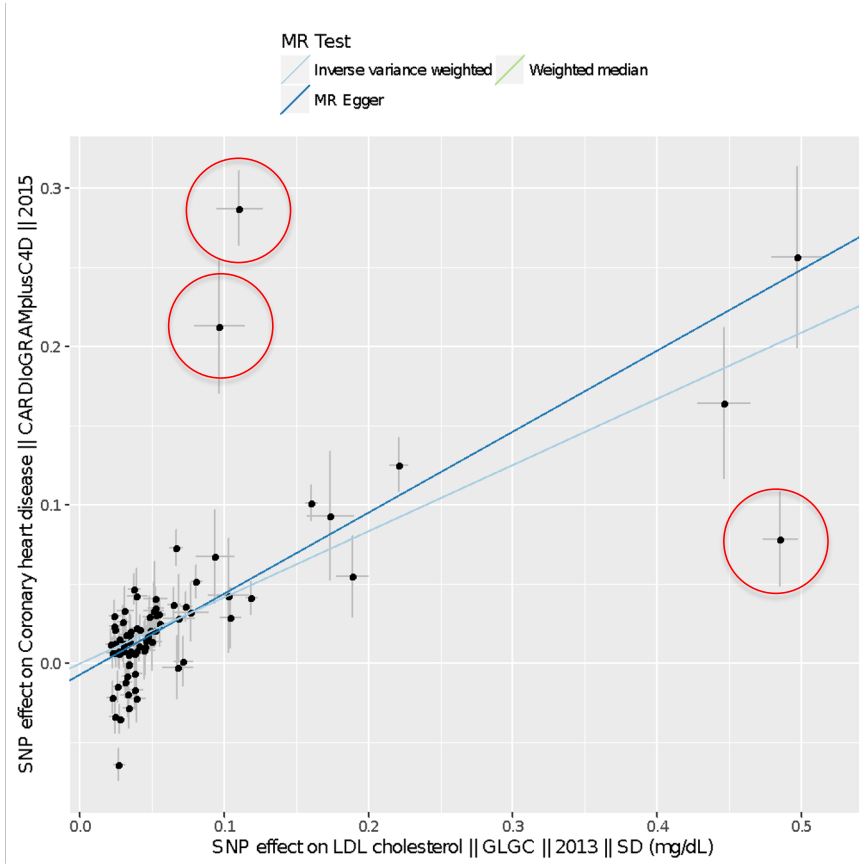
- This is a violation of MR assumptions (exclusion restriction principle)

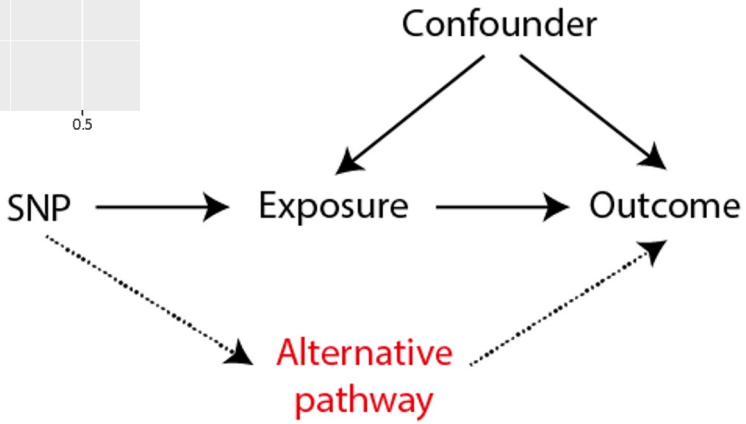
Distinct causal variants

Two causal variants in LD, one influences trait 1, the other influences trait 2

- Another means by which spurious MR results would arise




MR Test
Inverse variance weighted
Weighted median
MR Egger
SNP effect on Coronary heart disease || CARDIoGRAMplusC4D || 2015
SNP effect on LDL cholesterol || GLGC || 2013 || SD (mg/dL)
0.0
0.1
0.2
0.3
0.4
0.5


Confounder
SNP
Exposure
Outcome
Alternative
pathway

MR methods for handling horizontal pleiotropy

Many methods now exist

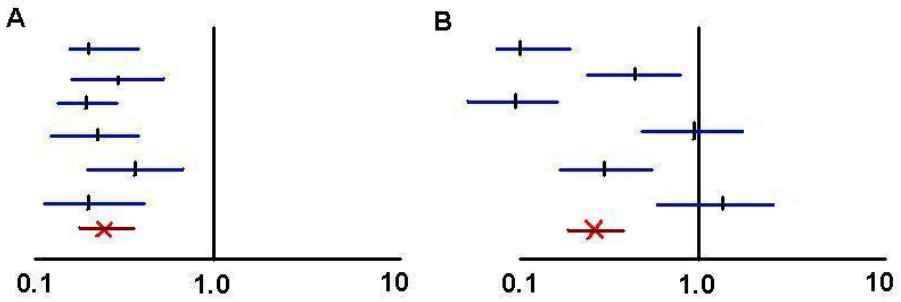
Heterogeneity

We expect that each SNP represents an independent study, and each should give an unbiased (if imprecise) estimate of the causal effect of x on y

Heterogeneity, where effect estimates are more different than expected due to standard errors, arises because at least some of the instruments are invalid

Cochran's Q statistic

$$Q = \sum_{k=1}^{K} w_k(\hat{\beta}_k - \hat{\beta}_{IVW})^2$$



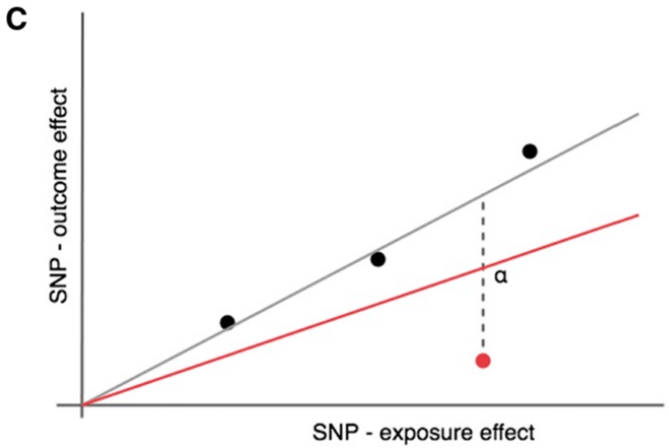
n=6 instruments
Expect Q = 5 if there is no heterogeneity
Q is chi-square distributed with n-1 degrees of freedom

Option 1: Remove outliers

- Some SNPs might contribute to the majority of the heterogeneity
- If we assume these are the invalid instruments then the IVW estimate excluding them should be less biased

However – beware of:

- Cherry picking – remove outliers will artificially provide a more precise estimate
- What if the outlier is the only valid instrument, and all the others are invalid?
  - E.g. cis-variants for gene expression, DNA methylation, protein levels. CRP levels are best instrumented by variants within the *CRP* gene region. Most other variants that come up in CRP GWAS are upstream effects related to inflammation

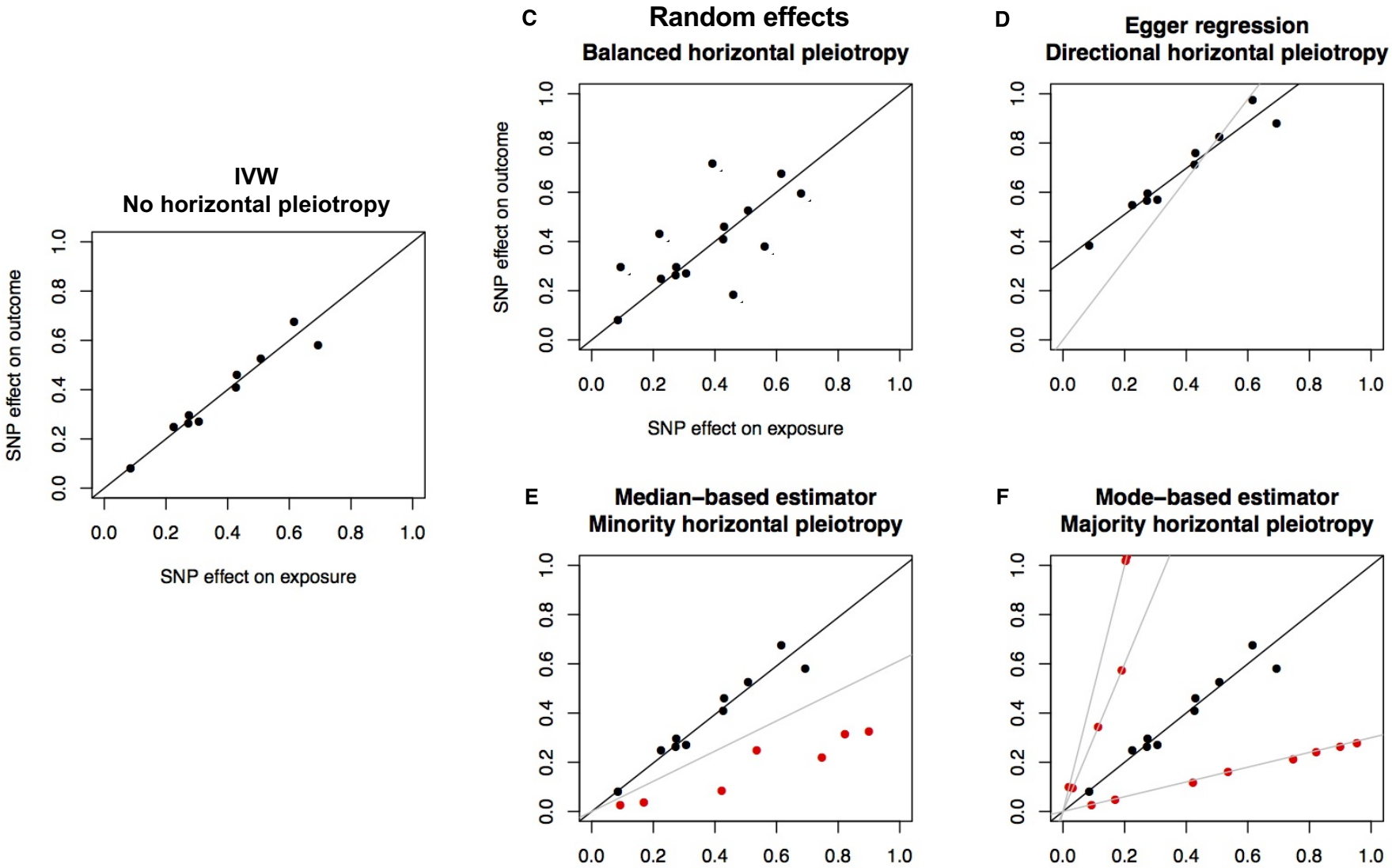


Option 2: Fit a model that is robust to some model of horizontal pleiotropy

- IVW fixed effects estimate assumes all SNPs are valid instruments, and averages across them all

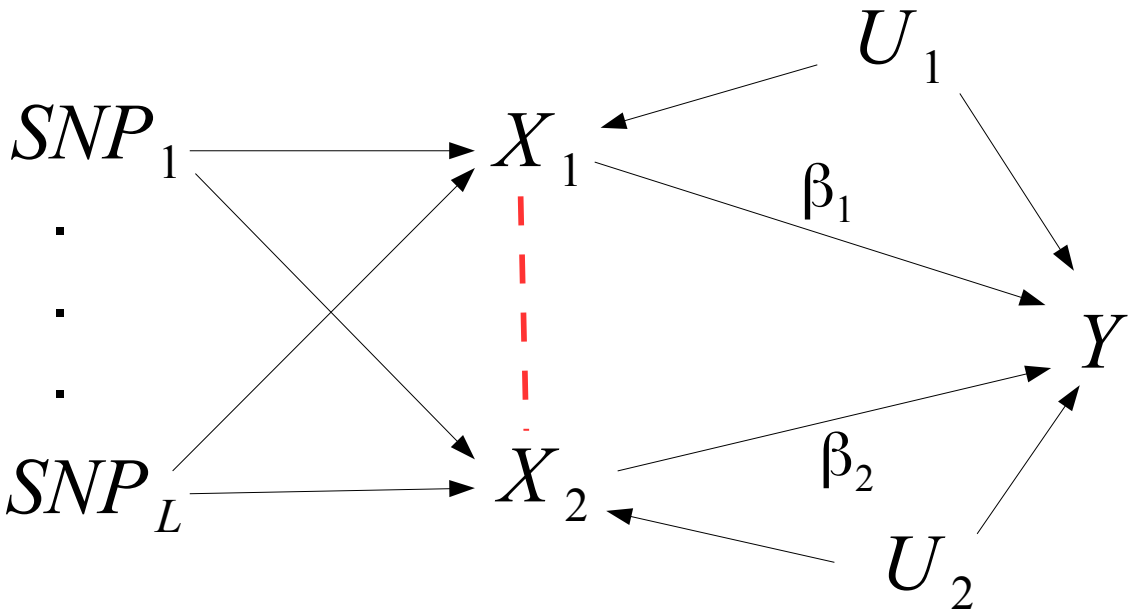
- IVW random effects model allows all SNPs to be drawn from a different distribution – the estimate is the same but the standard error is larger if there is any heterogeneity

- Several others…


IVW
No horizontal pleiotropy
SNP effect on outcome
SNP effect on exposure
C
Random effects
Balanced horizontal pleiotropy
SNP effect on outcome
SNP effect on exposure
D
Egger regression
Directional horizontal pleiotropy
E
Median−based estimator
Minority horizontal pleiotropy
F
Mode−based estimator
Majority horizontal pleiotropy
0.0
0.2
0.4
0.6
0.8
1.0

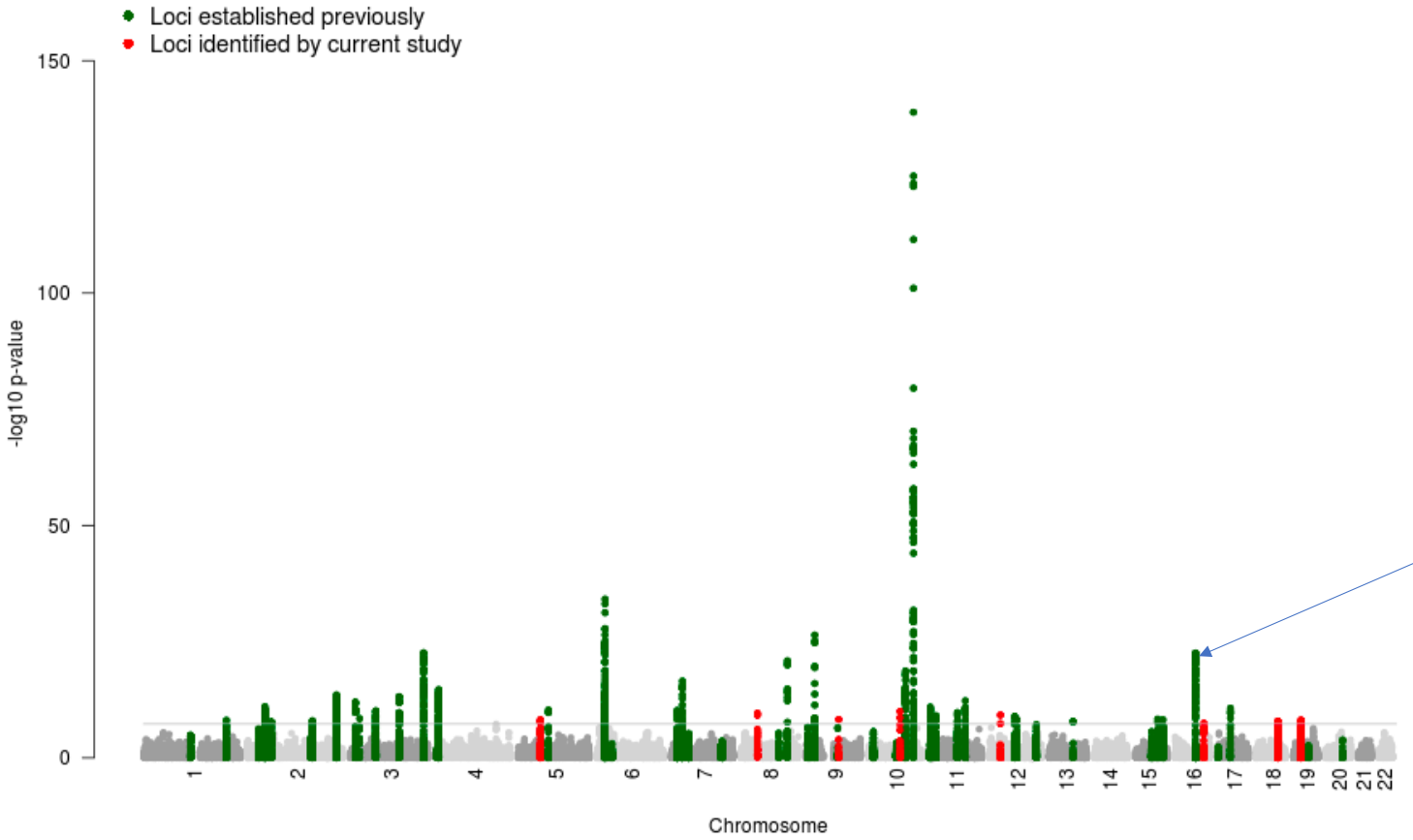
Option 3: Multivariable MR

- We are testing for whether X1 has an influence on Y
- We know that some instruments for X1 also have influences on X2
- This opens up the possibility of horizontal pleiotropy biasing our estimate
- What is the X1-Y association adjusting for X2?

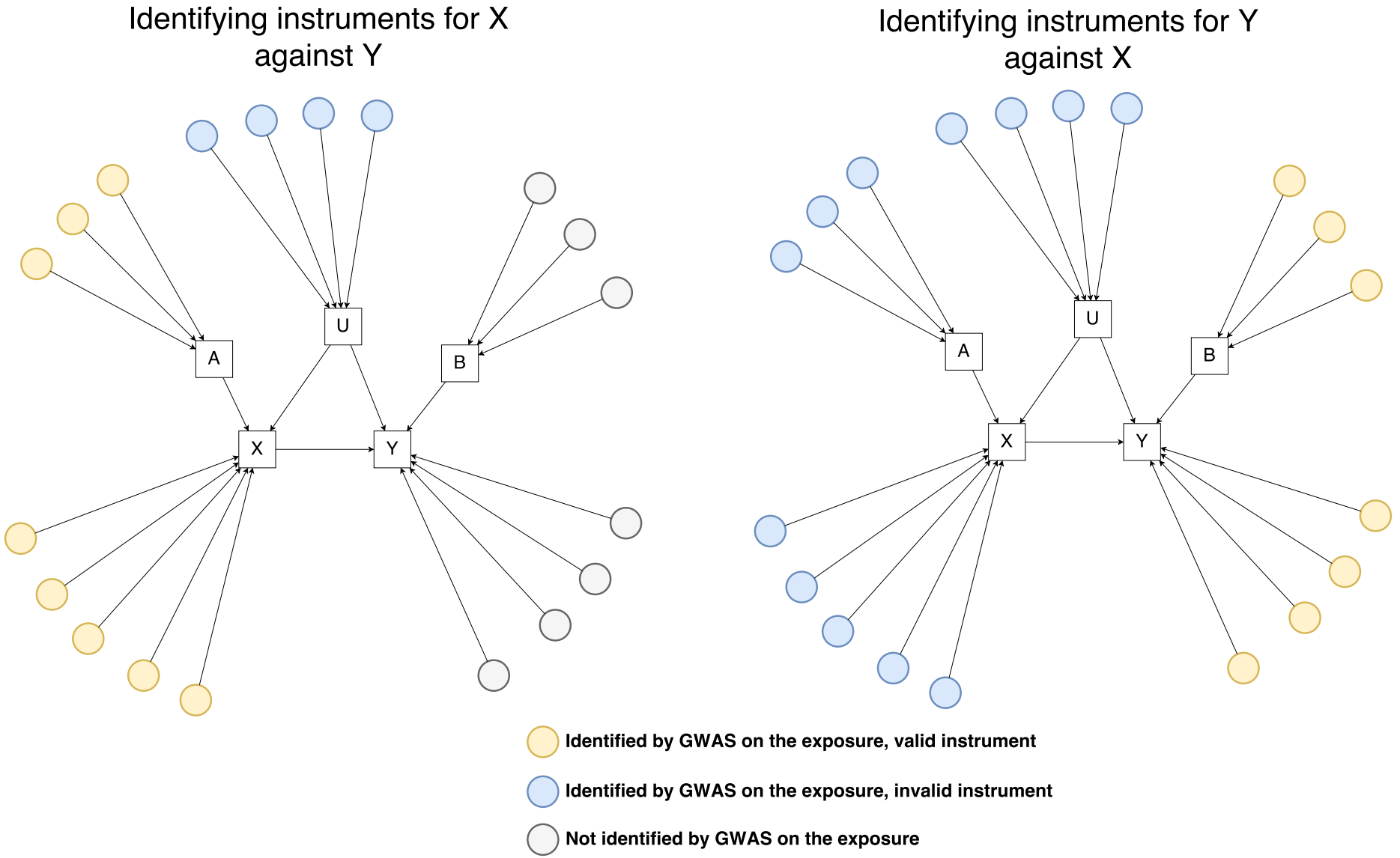


Reverse causal instruments

Problem: MR of type 2 diabetes on BMI

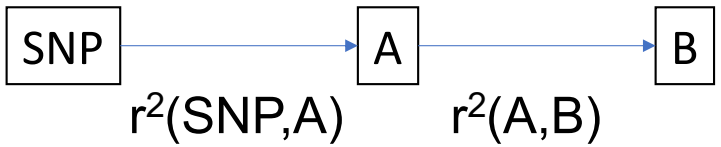


GWAS of T2D reveals *FTO* variant - Famously associated with BMI


Identifying instruments for X against Y
A
U
B
X
Y
Identifying instruments for Y against X
A
U
B
X
Y
Identified by GWAS on the exposure, valid instrument
Identified by GWAS on the exposure, invalid instrument
Not identified by GWAS on the exposure

Can we avoid including reverse-causal SNPs as instruments?

- If A causes B and B causes C
- The effect of A on B should be larger than the effect of A on C



Expect that

$r^2(SNP,B) = r^2(SNP,A) \times r^2(SNP,B)$

Steiger test used to evaluate if $r^2(SNP,A) > r^2(SNP,B)$

If this is not satisfied, infer that this instrument is not influencing the exposure primarily.

Single instrument analyses

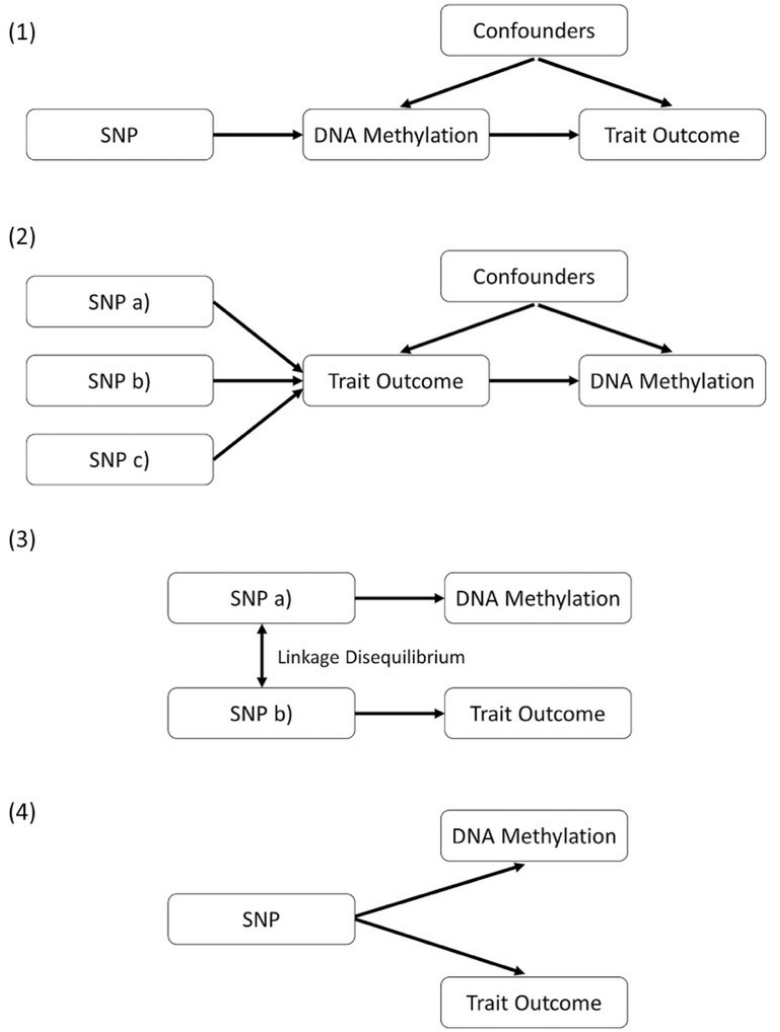
The single SNP scenario:

- Only 1 SNP known to influence the **exposure**
- SNP also associates with the **outcome**

How can this be interpreted?

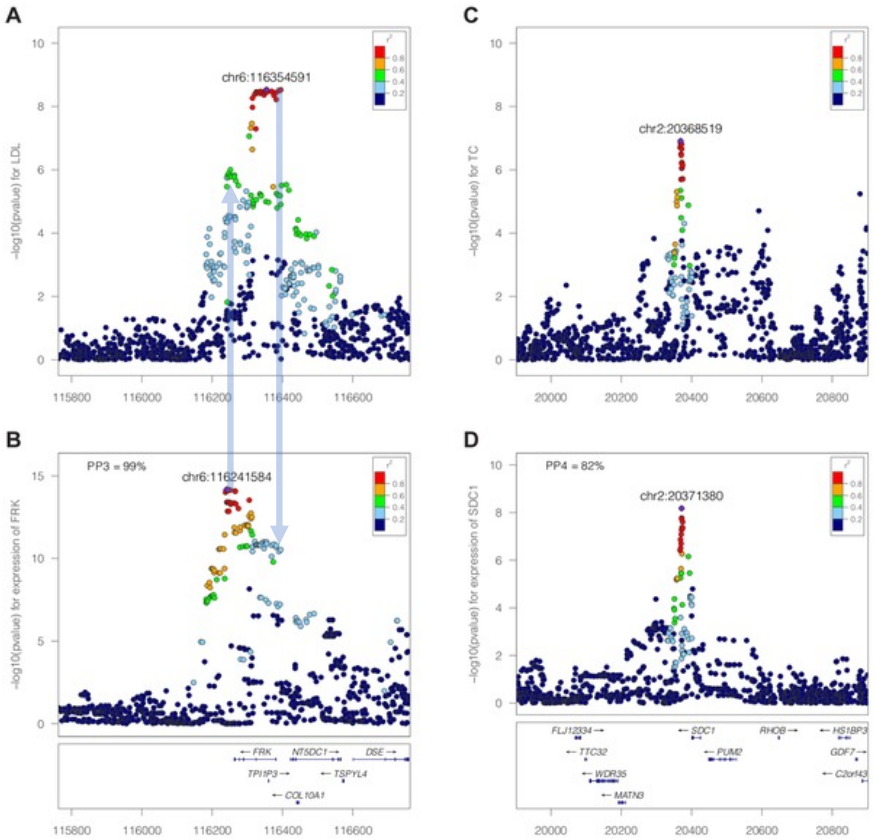
1. Exposure causes outcome
2. Outcome causes the exposure
3. The exposure SNP is in LD with a SNP that influences the outcome (spurious pleiotropy)
4. The SNP confounds the exposure-outcome relationship (horizontal pleiotropy)

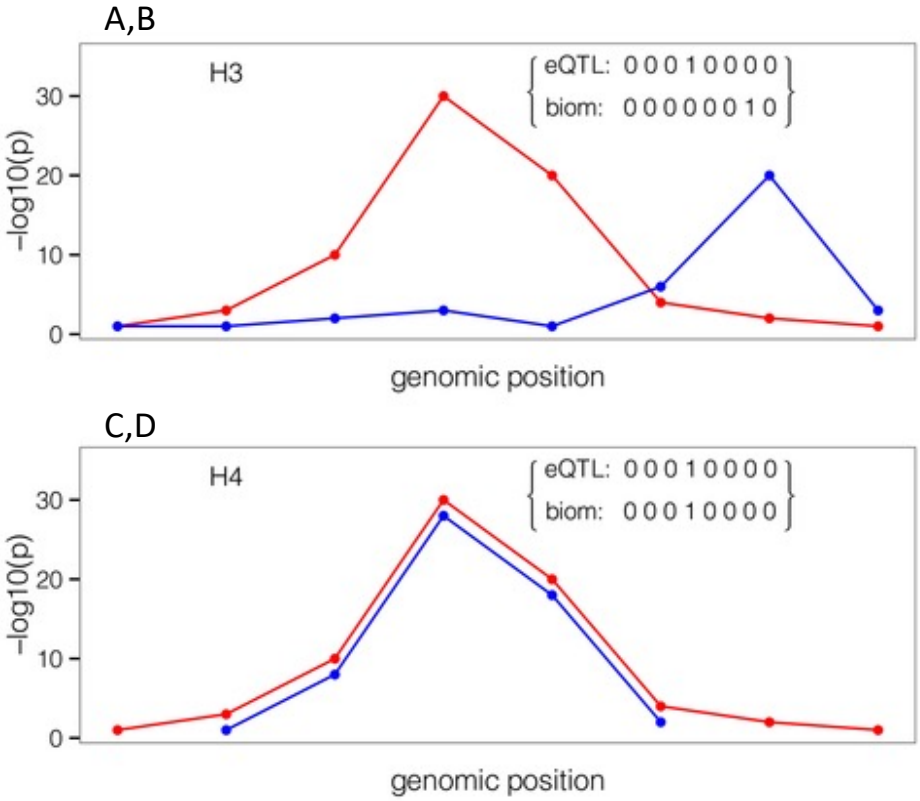
What can be done?



Richardson et al 2017 AJHG

Eliminate possibility of spurious pleiotropy through genetic colocalisation





Summary

- MR uses natural randomization to mimic an RCT
- It is useful, data is abundant, but it is not a panacea for causal inference
- Often valuable for proving that an hypothesised association is not causal
- Crucial to perform sensitivity analyses and obtain metrics regarding the likely reliability of the MR estimates

Background reading

- **Epidemiology—is it time to call it a day?** *International Journal of Epidemiology*, Volume 30, Issue 1, 1 February 2001, Pages 1–11
- **'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?** *International Journal of Epidemiology*, Volume 32, Issue 1, 1 February 2003, Pages 1–22
- **Mendelian randomization: genetic anchors for causal inference in epidemiological studies** *Human Molecular Genetics*, Volume 23, Issue R1, 15 September 2014, Pages R89–R98
- **Evaluating the potential role of pleiotropy in Mendelian randomization studies** *Human Molecular Genetics*, Volume 27, Issue R2, 1 August 2018, Pages R195–R208
- **Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians** *BMJ* 2018; 362
- **Orienting the causal relationship between imprecisely measured traits using GWAS summary data** PLoS Genet 2017 13(11): e1007081
- **Mendelian randomization in cardiometabolic disease: challenges in evaluating causality.** Nat Rev Cardiol. 2017 Oct;14(10):577-590
- **Using genetic data to strengthen causal inference in observational research** *Nature Reviews Genetics* volume 19, pages566–580 (2018)
- **Detecting individual and global horizontal pleiotropy in Mendelian randomization: a job for the humble heterogeneity statistic?** *American Journal of Epidemiology 2018*, kwy185
- **The MR-Base platform supports systematic causal inference across the human phenome** eLife 2018;7:e34408
- **Mendelian randomization accounting for horizontal and correlated pleiotropic effects using genome-wide summary statistics.** https://www.biorxiv.org/content/10.1101/682237v1
- **Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics.** https://www.medrxiv.org/content/10.1101/2020.01.27.20018929v1