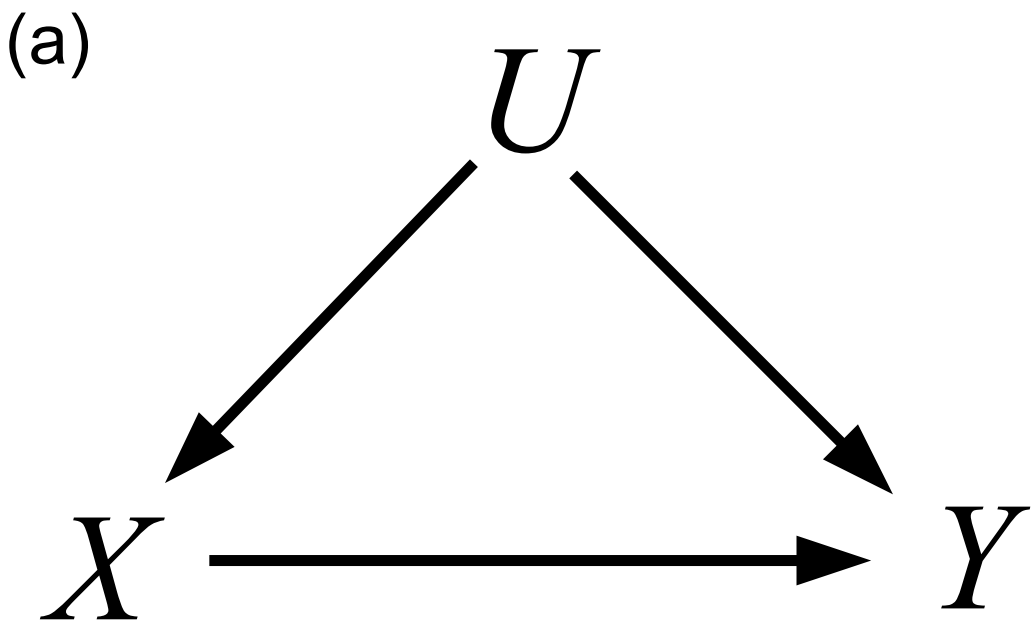
Recent advances in two-sample summary data Mendelian randomization

Jack Bowden

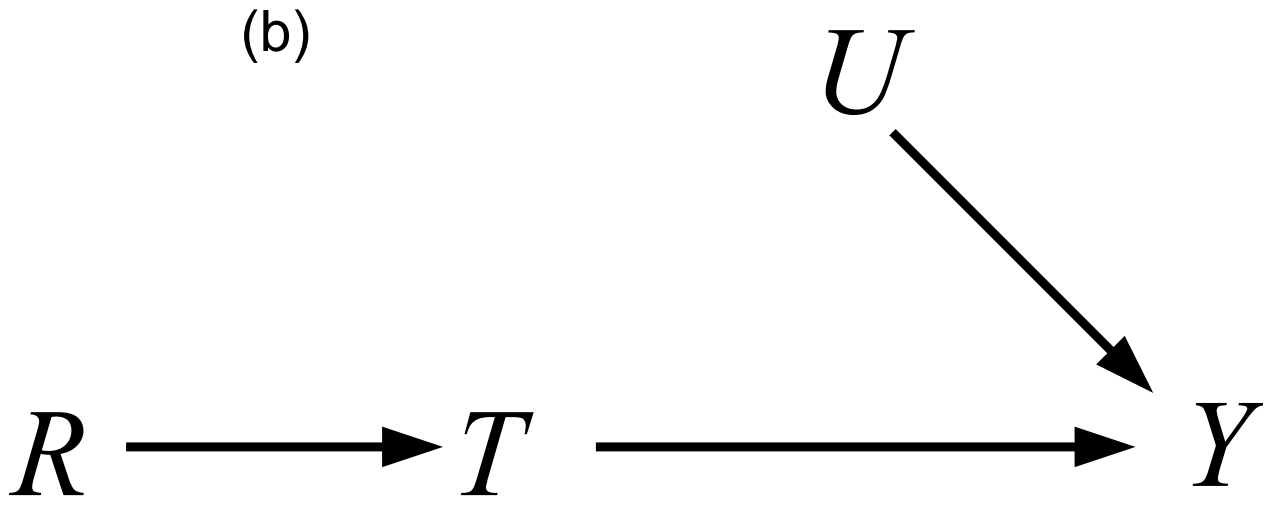
8th October 2018

The problem with observational epidemiology?



- Does health exposure X cause disease condition Y?

Experimental science



- Randomized controlled trials enable scientists to estimate the causal effect of treatment T on outcome Y in a simple and transparent manner
- i.e. Compare outcomes across randomized groups

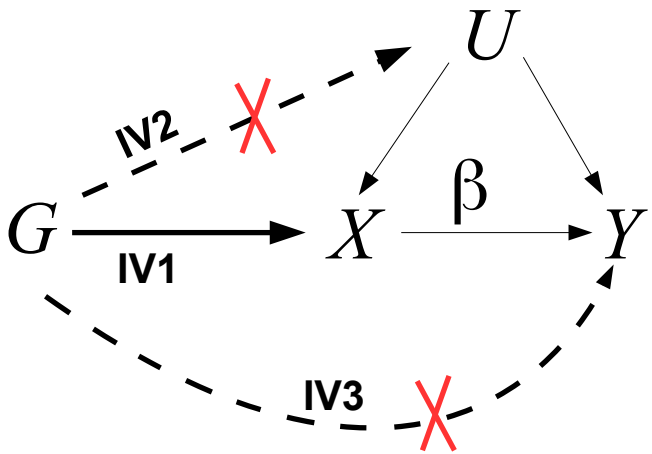
What is Mendelian randomization?

- Do you want to know what 'it' is?
- All I'm offering is the truth, nothing more

You take the blue pill and the story ends. You wake up in your bed and believe whatever you want to believe. You take the red pill and you stay in wonderland, and I show you how deep the rabbit hole goes.

- Like it or not, each and every one of us has been recruited into an experiment, from the moment we were born.
- Random genetic assignment determines (to a small degree) how much we eat, sleep, drink, weigh, smoke, study, worry, play.
- We can learn about causality using our genetic code.

MR requires genets satisfy the IV assumptions



- MR relies on assumptions IV1-IV3 to test for causality
- And then additional modelling assumptions to estimate causal effect

Background

- Traditional MR studies made use of individual level data
  - Genetic, exposure and outcome measured for each individual
- Utilised small number of variants with a known functional effect
  - *CRP* genes used to probe causal effect of CRP on CHD risk
  - *ALDH2* gene used to probe causal effect of alcohol on CHD
- TSLS estimates often combined across studies to improve power
- High level of cooperation and administrative burden required
- Relatively inefficient for the large-scale pursuit of MR analyses.

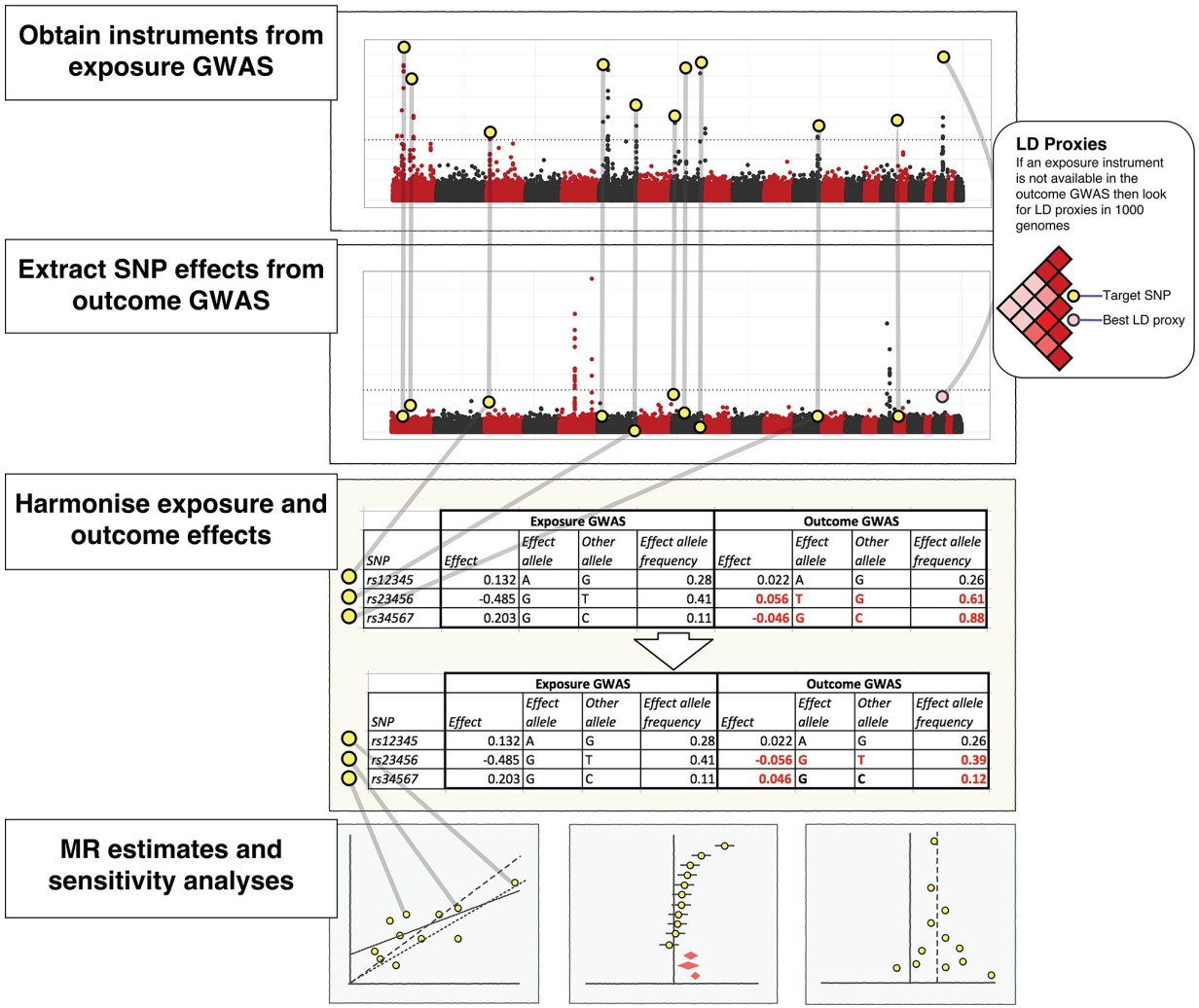
Summarized data and Mendelian randomization

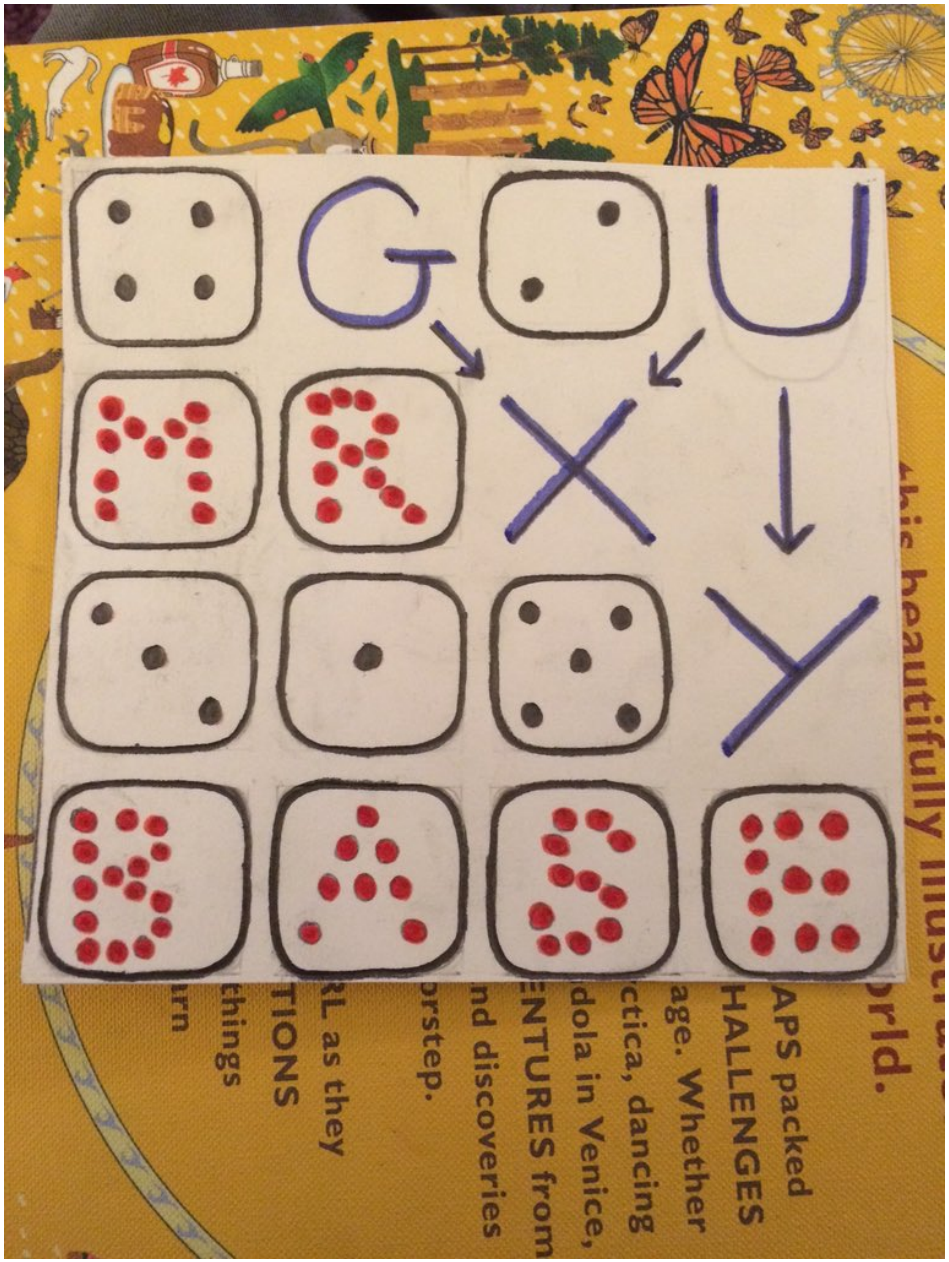
- In recent years, it has become possible in theory for *anyone* to conduct an MR analysis
- Achieved by combining summary estimates of SNP-trait associations from two genome wide association studies (GWAS) with publically available data
- Referred to as two-sample summary data MR

Example disease consortia with publically available data

Trait	Consortium
Alzheimer's	International Genomics of Alzheimer's Project (IGAP)
Anthropometric traits	Genetic Investigation of ANthropometric Traits (GIANT)
Autism	Psychiatric Genomics Consortium (PGC)
Bipolar disorder	
Major depressive disorder	
Blood pressure	International Consortium for Blood Pressure (ICBP)
Coronary heart disease	Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIOGRAM)
Glycaemic traits	Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC)
Lipids	Global Lipids Genetics Consortium (GLGC)
Osteoporosis	GEnetic Factors for OSteoporosis Consortium (GEFOS)
Smoking	Tobacco and Genetics Consortium (TAG)
Type II diabetes	DIAbetes Genetics Replication And Meta-analysis (DIAGRAM)

- A one stop shop containing the summary data and analysis tools





- Didn't develop MR-Base, but it uses some of my methods!

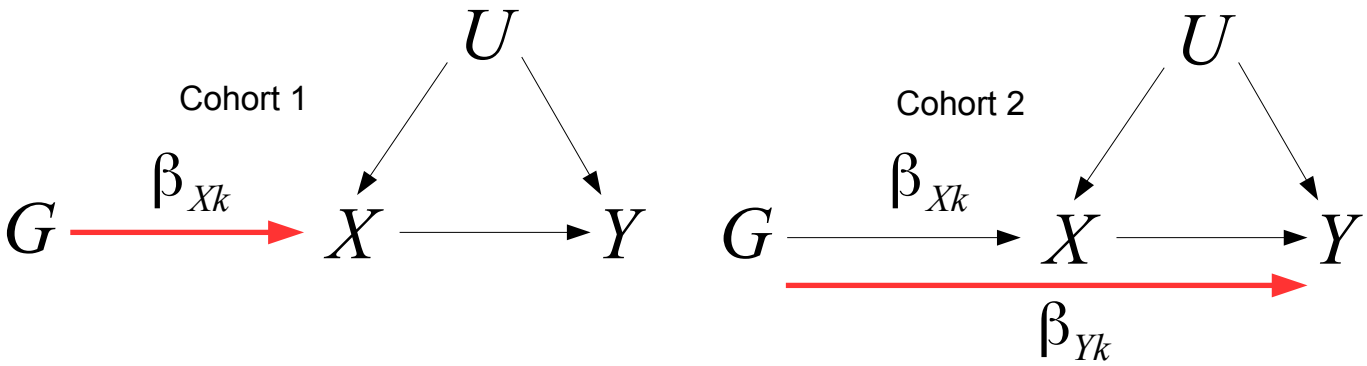
Assumed model for two-sample summary data MR

- Suppose, for each SNP:

  - The underlying SNP-outcome association (from cohort 2)
  - The underlying SNP-exposure association (from cohort 1)

  are linked by the causal effect parameter β in the equation:

$$\beta_{Yk} = \beta\beta_{Xk}$$



- β is the increase in Y when we intervene and change X by 1 unit
- Estimate β via $\hat{\beta}_k = \hat{\beta}_{Yk}/\hat{\beta}_{Xk}$

Combining summary data estimates

- The overall inverse-variance weighted (IVW) estimate of the causal effect combines the ratio estimates of multiple variants using the approximate variance just derived, to give:

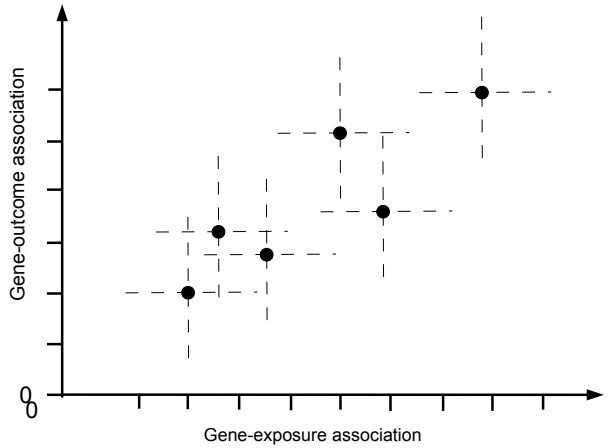
$$\hat{\beta}_{IVW} = \frac{\sum_{k=1}^{K} \hat{\beta}_k w_k}{\sum_{k=1}^{K} w_k}$$

  where w_k is the reciprocal of the variance of $\hat{\beta}_k$

- IVW asymptotically equivalent to TSLS with uncorrelated SNPs

Visual representation

- Plot of $\hat{\beta}_{Yk}$ Vs $\hat{\beta}_{Xk}$



- Standard MR = Regression of $\hat{\beta}_{Yk}$ on $\hat{\beta}_{Xk}$ with no intercept
- Two problems
  1. Genes most likely a mixture of valid and invalid IVs (IV2/3)
  2. Most genes are only weakly associated with X (IV1)
- 2 mitigated by choosing SNPs strongly associated with X
- Still often yields 20-150 SNPs, and allows focus on 1.

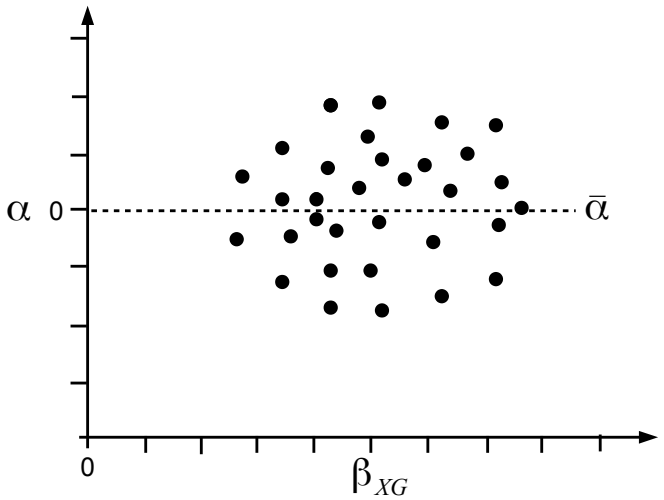
Progress on problem 1: the new 'standard' MR model

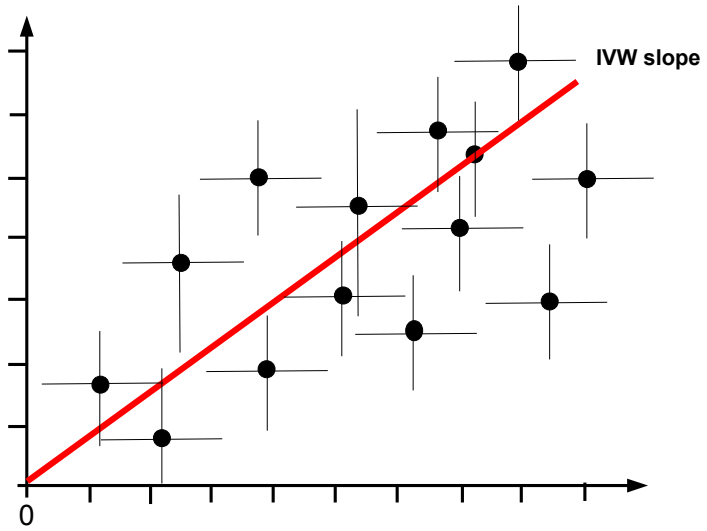
- An extended linear model accounting for SNP invalidity:

$$\beta_{Yk} = \alpha_k + \beta\beta_{Xk}$$

- $\alpha_k = 0$: $\rightarrow$ SNP k valid
- $\alpha_k \neq 0$: $\rightarrow$ SNP k invalid
- **S**ome **I**nvalid **S**ome **V**alid **I**instrumental **V**ariable **E**stimation (SISVIVE) framework (Kang et al, JASA 2016)
- **The Game**: What must we assume about α_k in order to be able to identify and estimate the causal effect β?

Pleiotropy equals a zero mean random effect

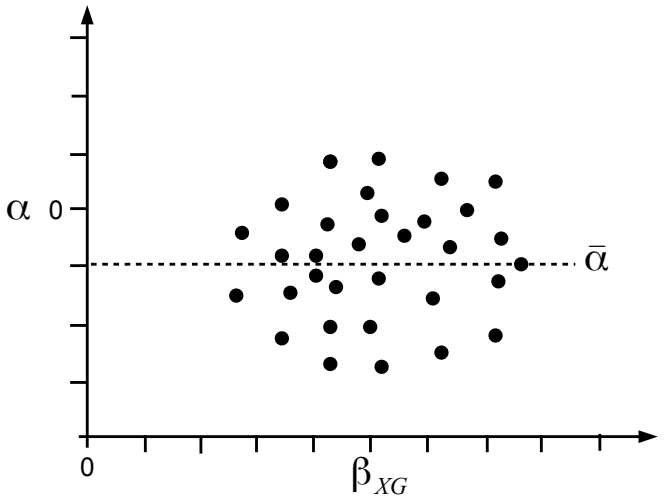


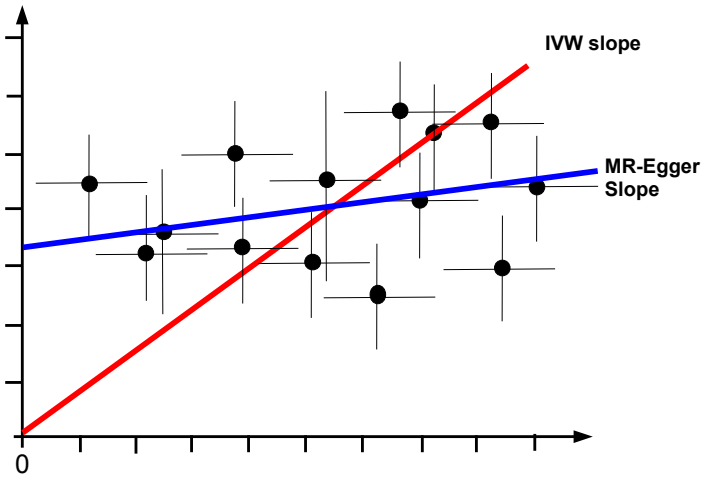


- If: $\alpha_k \perp\!\!\!\perp \beta_{Xk}$ and $\mathrm{E}[\alpha_k] = 0$ then the pleiotropy is said to be 'balanced'
- Justifies simple random effects meta-analysis: $\hat{\beta} = \sum w_k \hat{\beta}_k / \sum w_k$

  w_k inflated to take account of extra variation due to pleiotropy

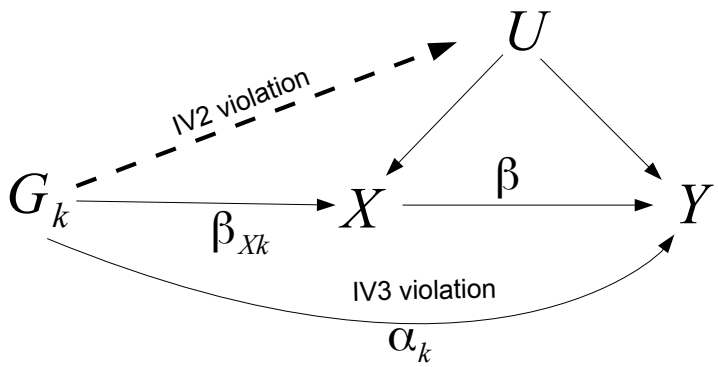
Adjusting for non-zero mean pleiotropy (Bowden et al, 2015)

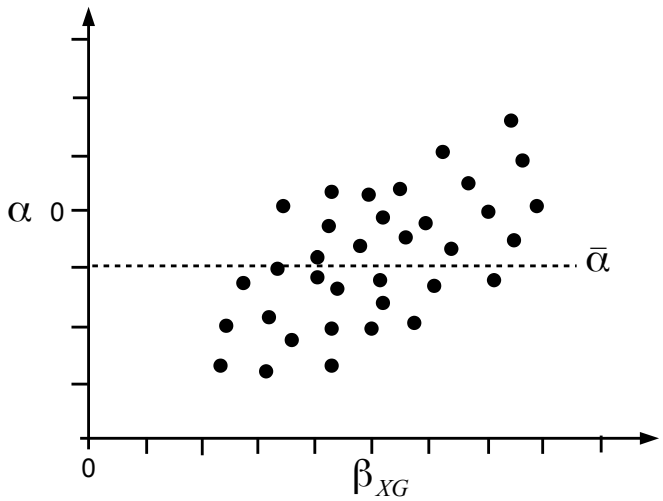




- Perform meta-regression: $\hat{\beta}_{YGk} = \beta_{0E} + \beta_{1E}\hat{\beta}_{XGk}$
- I called it 'MR-Egger regression'
- Valid if $\alpha_k \perp\!\!\!\perp \beta_{Xk}$

Plausability of InSIDE assumption





- MR-Egger & IVW allow 100% of SNPs to be invalid instruments, but requires InSIDE
- InSIDE most likely to be satisfied when pleiotropy occurs via an independent pathway, not via a confounder
- MR-Egger more sensitive to InSIDE violation than IVW

Median-based estimation (Bowden et al (2016)

- Suppose instead that the following were true:

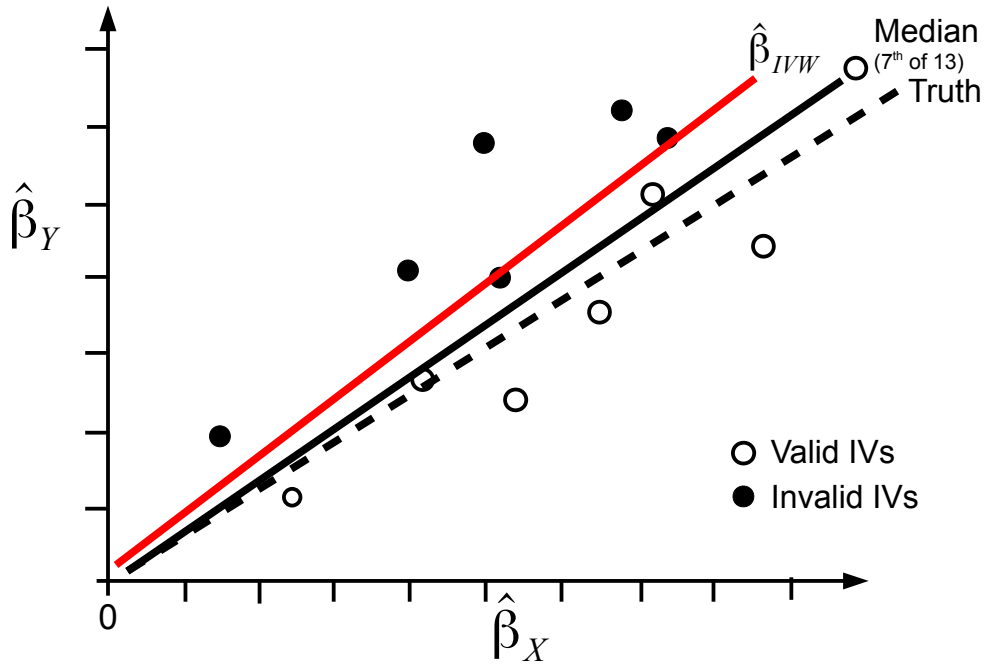
$$\alpha_k = 0 \qquad \text{for} > 50\% \text{ of the variants}$$

- That is, the majority of SNPs are 'valid' IVs
- **No** restrictions need to be placed on the invalid IVs

  - InSIDE not required, violations via IV2 and IV3 are allowed

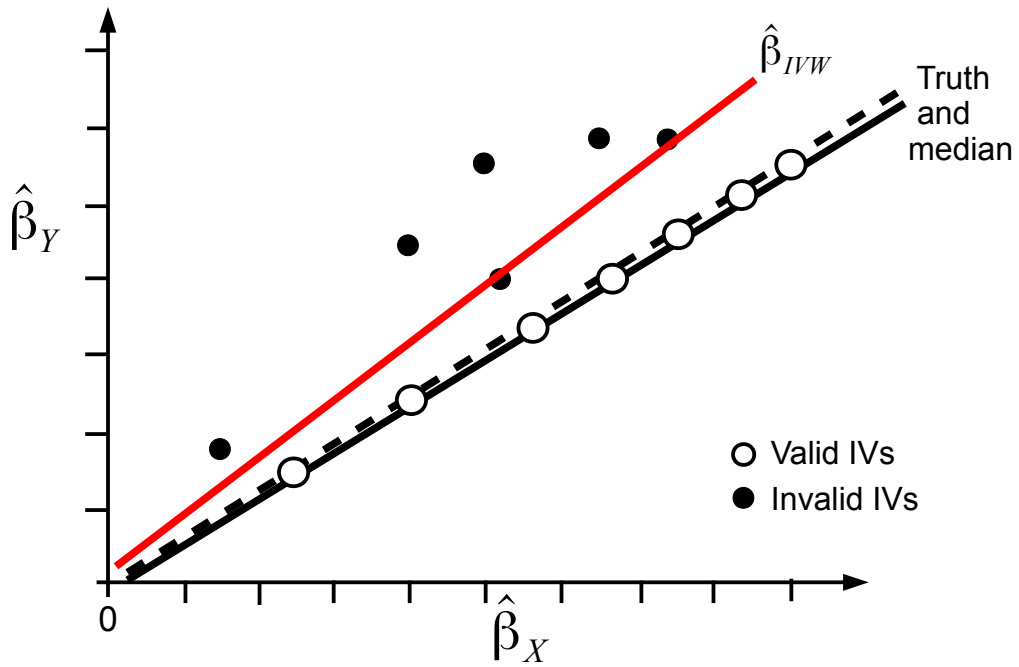
- If true, the median ratio estimate is a more reliable estimate for β

Hypothetical example scatter plot (finite sample data)



- IVW estimate biased
- Median estimate biased, but closer to truth

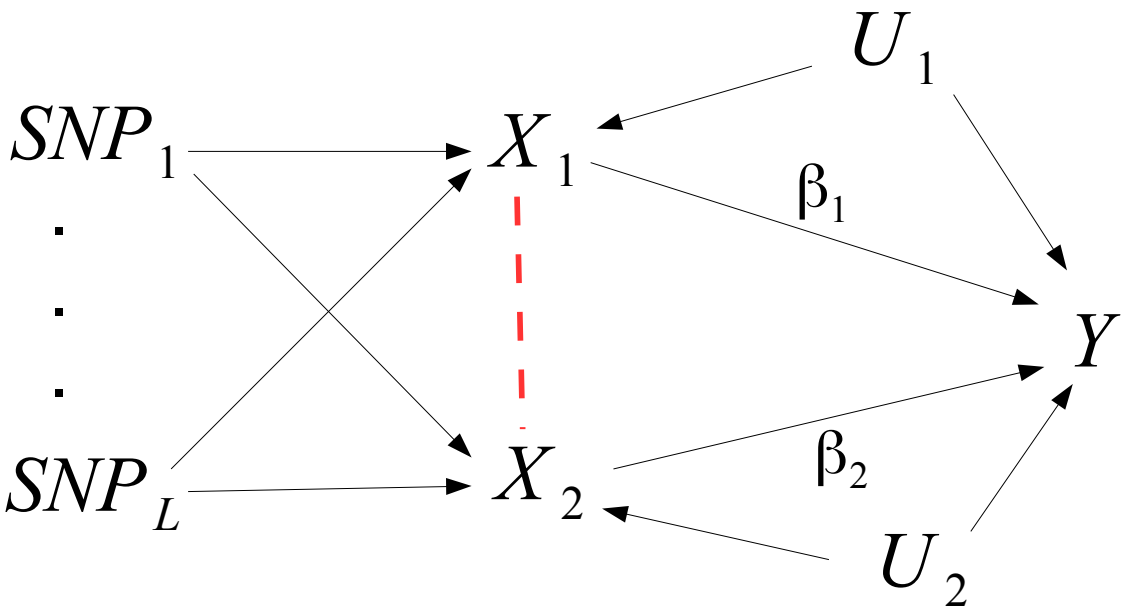
Scatter plot for infinite sample data



- IVW estimate asymptotically biased, but median consistent
- In practice, we calculate a 'weighted median' from a weighted empirical distribution function of ratio estimates
- Similar ideas can be used to define a mode-based estimate

Genetic confounder adjustment: Multi-variable MR

- IV2 could be violated if SNPs affect exposure of interest through correlated phenotype
- e.g. X_1 = LDL cholestorol, X_2 = HDL cholesterol, Y = CHD
- Estimate the causal effect of X_1 on Y adjusting for genetically X_2
- Avoids possible collider bias from adjusting for observed X_2



- Fit model: $\hat{\beta}_{YGk} = \beta_1 \hat{\beta}_{XG1k} + \beta_2 \hat{\beta}_{XG2k}$

Detecting and removing outliers

- If all SNPs are valid IVs & linearity/modelling assumptions hold:

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{L} w_j \hat{\beta}_j}{\sum_{j=1}^{L} w_j} \quad \text{where} \quad w_j = var(\hat{\beta}_j)^{-1}$$

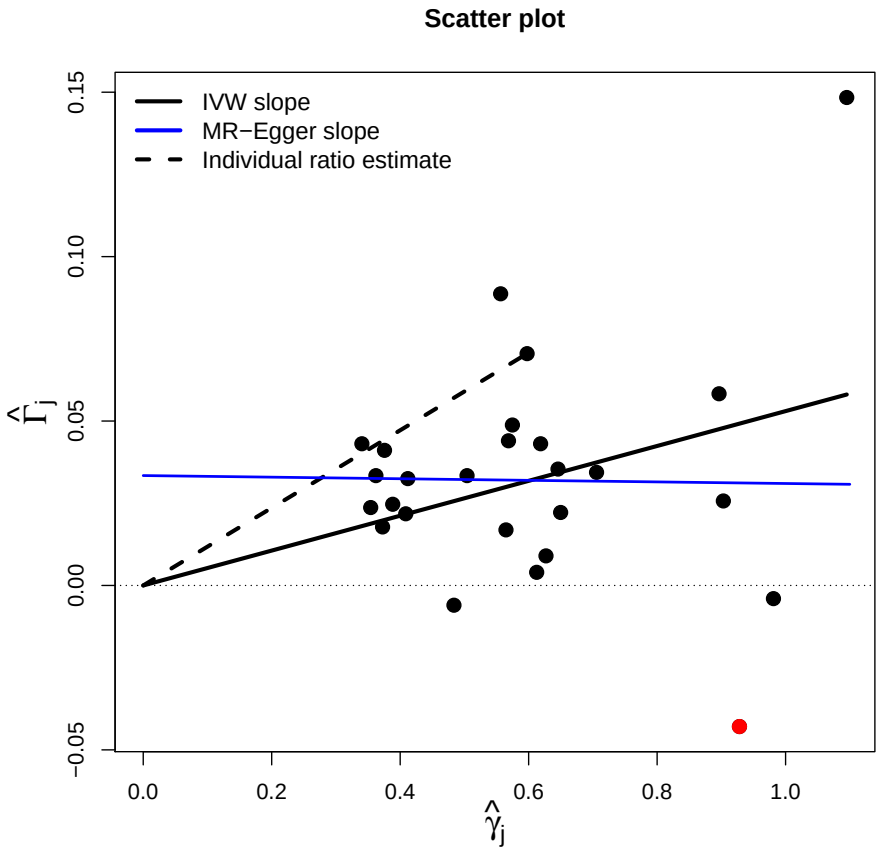
  should be both a consistent and precise measure of causal effect

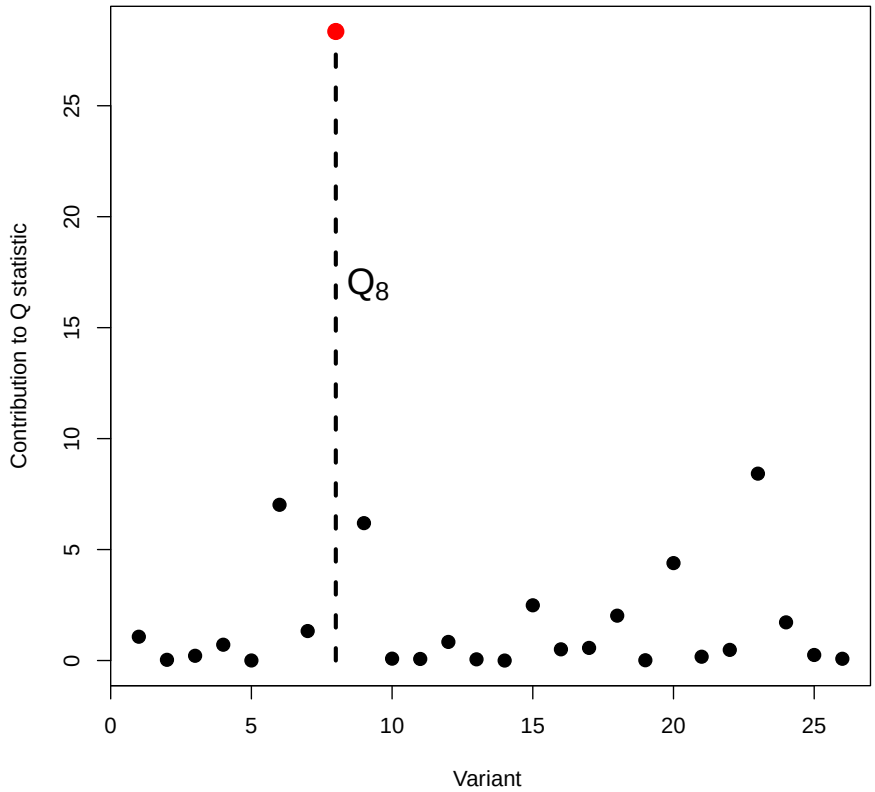
- Heterogeneity can be assessed using Cochran's Q:

$$Q = \sum_{j=1}^{L} Q_j = \sum_{j=1}^{L} w_j (\hat{\beta}_j - \hat{\beta}_{IVW})^2$$

- Substantial heterogeneity a sign of pleiotropy
- Individual Q_j indicate which SNPs are most pleiotropic
- Large outliers could lead to unbalanced pleiotropy

Application: Does SBP causally influence CHD risk?





- 26 SNP-SBP & SNP-CHD estimates from ICBP/CARDIoGRAM
- SNP rs17249754 is a major outlier

Results

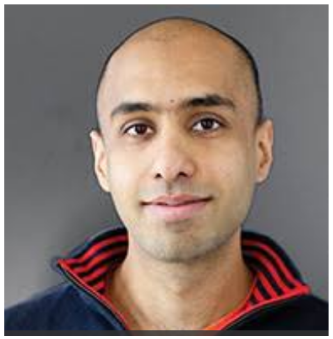
Method (weights)	Estimate (C.I)	S.E.	P-value	Het. Stat (p)	$\hat{\phi}$
All 26 SNPs					
Causal estimate					
IVW	$\hat{\beta}_{IVW}$: 0.054 (0.027,0.082)	0.014	4.60×10^{-4}	$Q = 62.4$ (4.84×10^{-5})	2.61
MR-Egger	$\hat{\beta}_{1E}$: -0.002 (-0.063 0.059)	0.031	0.94	$Q' = 58.6$ (1.00×10^{-4})	2.50
MR-Egger intercept					
MR-Egger	$\hat{\beta}_{0E}$: 0.033	0.018	0.075	-	
Weighted median (1st order weights)					
Weighted Median	$\hat{\beta}_{WM}$: 0.063 (0.042,0.084)	0.011	4.90×10^{-6}	-	-
SNP rs17249754 removed					
Causal estimate					
IVW	$\hat{\beta}_{IVW}$: 0.067 (0.049,0.085)	0.009	8.37×10^{-8}	$Q = 32.8$ (0.108)	1.39
MR-Egger (1st)	$\hat{\beta}_{1E}$: 0.0490 (-0.006,0.104)	0.028	0.09	$Q' = 34.3$ (0.061)	1.35
MR-Egger intercept					
MR-Egger (1st)	$\hat{\beta}_{0E}$: 0.010	0.015	0.51	-	
Weighted median (1st order weights)					
Weighted Median	$\hat{\beta}_{WM}$: 0.065 (0.044,0.087)	0.011	2.33×10^{-6}	-	

- MR-Egger and IVW in good agreement after removal of outlier
- Nice illustration of weighted median's robustness to outliers

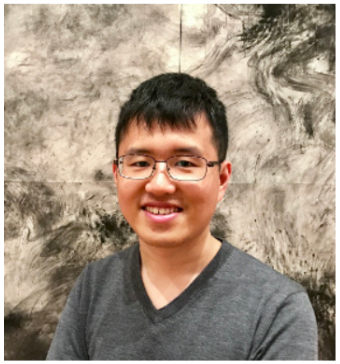
Summary

- MR possible with summary data even in the presence of pleiotropy
- Pleiotropy can be benign if 'balanced' then standard IVW analysis fine
- If pleiotropy has a directional element, then standard IVW analysis may be biased
- MR-Egger regression, Weighted median, Multi-variable MR are useful tools for sensitivity analysis to explore the impact of pleiotropy
- All of these approaches can be implemented in MR-Base

Work not possible without my many collaborators






Come to the MR conference in 2019!

Some references (published papers and ArXiv pre-prints)

Bowden, J., Del Greco, FM., Minelli, C., Davey Smith, G., Sheehan, N. and Thompson, J. (2017). A framework for the assessment of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*,**36**:1783–1802.

Bowden, J., Davey Smith, G., Haycock, P., Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* **40,**: 304–314.

Hartwig FP, Davey Smith G, Bowden J. (2017). Robust inference in two-sample Mendelian randomisation via the zero modal pleiotropy assumption. *International Journal of Epidemiology*. DOI: 10.1093/ije/dyx102

Bowden J, Del Greco F, Minelli C, Lawlor D, Zhao Q, Sheehan N, Thompson J, Davey Smith G. Improving the accuracy of two-sample summary data Mendelian randomization: Moving beyond the NOME assumption. (2018). Under review at *International Journal of Epidemiology* https://www.biorxiv.org/content/early/2018/02/27/159442

Zhao Q, Wang J, Hemani G, Bowden J, Small D. (2018). Statistical inference in two-sample summary data Mendelian randomization using robust adjusted profile score. Submitted to *Annals of Statistics* `https://arxiv.org/abs/1801.09652`

Zhao Q, Wang J, Bowden J, Small D. (2017). Two-sample Instrumental Variable Analyses using heterogeneous samples. Submitted to *Statistical Science* `https://arxiv.org/pdf/1709.00081.pdf`

Extra Slides

Example: 10 SNPs: weighted and unweighted median

- Order ratio estimates from smallest to largest
- $\hat{\beta}_{(1)}, \hat{\beta}_{(2)}, \ldots, \hat{\beta}_{(10)}$

	$\hat{\beta}_{(1)}$	$\hat{\beta}_{(2)}$	$\hat{\beta}_{(3)}$	$\hat{\beta}_{(4)}$	$\hat{\beta}_{(5)}$	$\hat{\beta}_{(6)}$	$\hat{\beta}_{(7)}$	$\hat{\beta}_{(8)}$	$\hat{\beta}_{(9)}$	$\hat{\beta}_{(10)}$
Simple median										
Weight (w_k)	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Percentile (p_k)	5	15	25	35	45	55	65	75	85	95
Weighting 1										
Weight (w_k)	$\frac{1}{30}$	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{4}{30}$	$\frac{5}{30}$	$\frac{5}{30}$	$\frac{4}{30}$	$\frac{3}{30}$	$\frac{2}{30}$	$\frac{1}{30}$
Percentile	1.67	6.67	15.00	26.67	41.67	58.33	73.33	85.00	93.33	98.33
Weighting 2										
Weight (w_k)	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Percentile (p_k)	2.78	9.72	27.78	52.78	70.83	81.94	88.89	93.06	95.83	98.61

- Simple median $= \frac{\hat{\beta}_5 + \hat{\beta}_6}{2}$
- Weighted median: $\hat{\beta}_{WM} = \hat{\beta}_3 + (\hat{\beta}_4 - \hat{\beta}_3) \times \frac{50 - 27.78}{52.78 - 27.78}$

Accounting for weak/pleiotropic instruments

- View Q as an estimating equation for β
- Extend to allow w_j to depend on β

$$Q_m(\beta) = \sum_{j=1}^{L} w_j(\beta)(\hat{\beta}_j - \beta)^2 \quad \text{where} \quad w_j(\beta) = \left(\frac{\sigma^2_{Yj} + \beta^2 \sigma^2_{Xj}}{\hat{\gamma}^2_j} \right)^{-1}$$

- Either:
  1. 'Iteratively' re-calculate estimates for $\hat{\beta}_{IVW}$, and $Q_m(\hat{\beta}_{IVW})$

     - Summary data analogue of 'two-step GMM'
  2. Find β that 'exactly' minimises Q statistic

     - Summary data analogye of LIML

Mendelian randomization: Target validation

- MR used extensively in epidemiology to understand causal determinants of disease

  - Feeds into public health policy (e.g. on importance of healthy BMI)

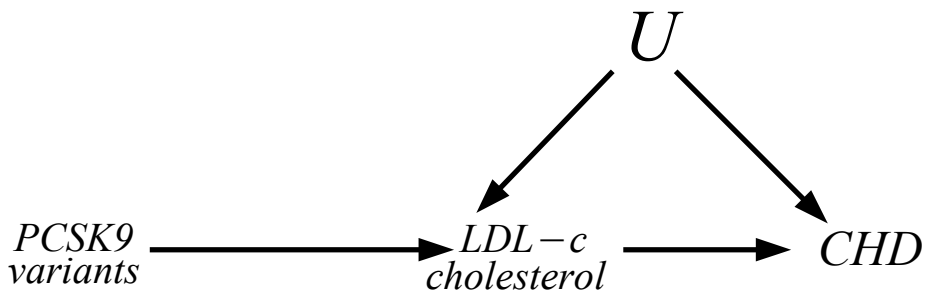
- But also useful more directly in drug development

- For example, Statins, which inhibit *3-hydroxy-3-methylglutaryl-CoA reductase* (HMGCR), are currently the most commonly used drug to reduce LDL cholesterol

- *PCSK9* protein binds to LDL receptors on the surface of the liver, decreasing the capacity of the liver to remove LDL cholesterol from circulation

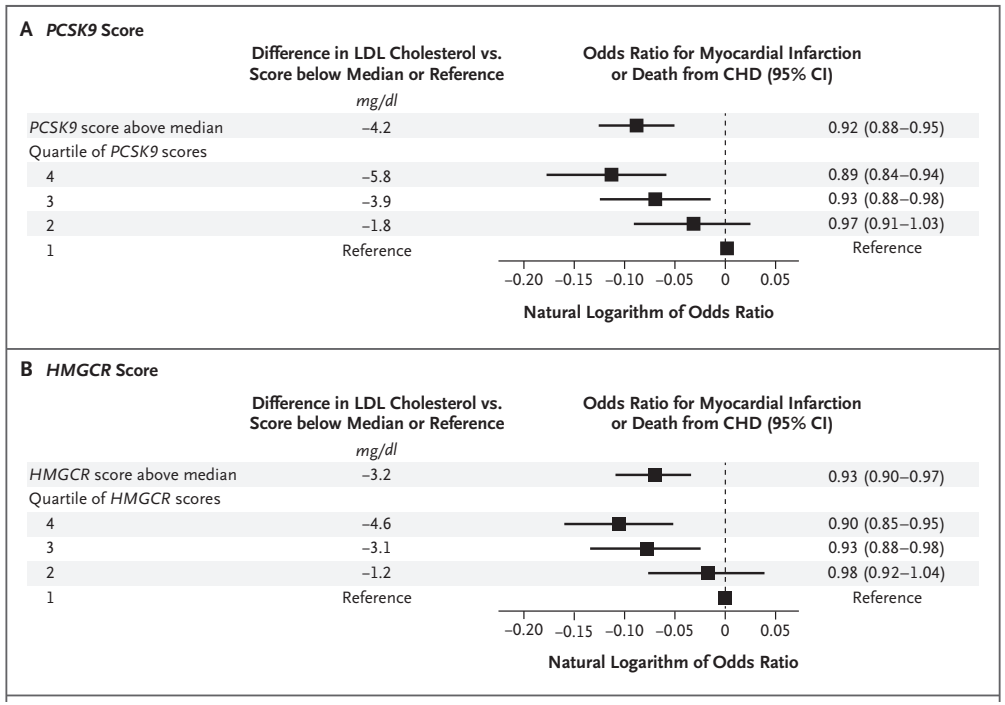
  - *PCSK9* inhibitors have been proposed as a way to reduce LDL cholesterol

- Use genes known to influence *PCSK9* expression & MR to predict effect on CHD/MI risk in trials



Mendelian randomization: Target validation

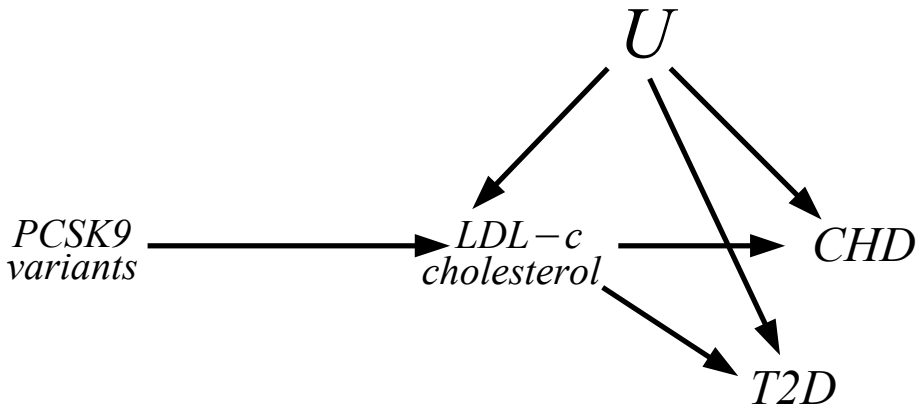
- 'In this study, variants in PCSK9 had approximately the same effect as variants in HMGCR on the risk of cardiovascular events and diabetes per unit decrease in the LDL cholesterol level. The effects of these variants were independent and additive'



- Ference, B. A.et al.Variation in PCSK9 and HMGCR and Risk of Cardiovascular Disease and Diabetes. *NEJM* 375,2144–2153 (2016).

Mendelian randomization: Side effect prediction

- MR can predict the same side effects of LDL lowering treatment is an increased risk of type II Diabetes



- The 'genetic support' provided by MR analyses is becoming increasingly invaluable in drug development
- It is under-pinned by IV theory
- Lotta et al. Association Between Low-Density Lipoprotein Cholesterol-Lowering Genetic Variants and Risk of Type 2 Diabetes A Meta-analysis. JAMA 2016: 1383–1391