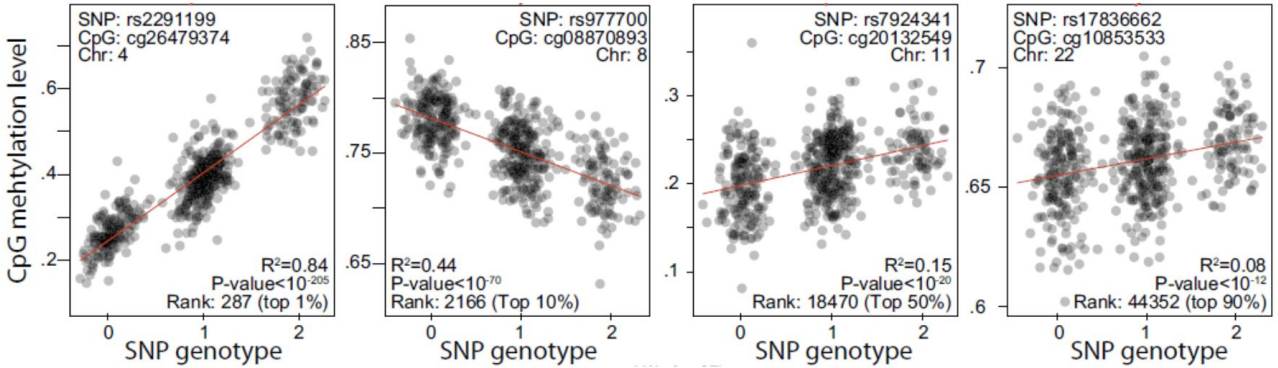
Recitation 8

GWAS + Heritability + stats

eQTL

- Expression quantitative loci (eQTLs) are genomic loci that explain variation in expression levels of mRNAs or proteins
  - Quantitative traits are phenotypes that can be measured (e.g. height, weight)
- There are two kinds of QTLs:
  - cis-eQTLs are local to the gene of interest
  - trans-eQTLs are at a distance



How Do We Detect eQTLs?

- By assaying gene expression and genetic variation simultaneously on a genome-wide basis for a large number of individuals
- After, use statistical genetic methods to map factors underpinning differences in quantitative levels of expression for individuals
- Defining DNA variants that cause this variation are usually difficult though, and sometimes require multiple experimentation rounds
    - especially for trans-eQTLs since that do not benefit from strong prior probability that relevant variants are in the immediate vicinity of the parent gene

Linear Mixed Models (LMM)

- Primarily used for analyzing non-independent, hierarchical, longitudinal, or correlated data
  - Particularly used when there is non-independence in the data

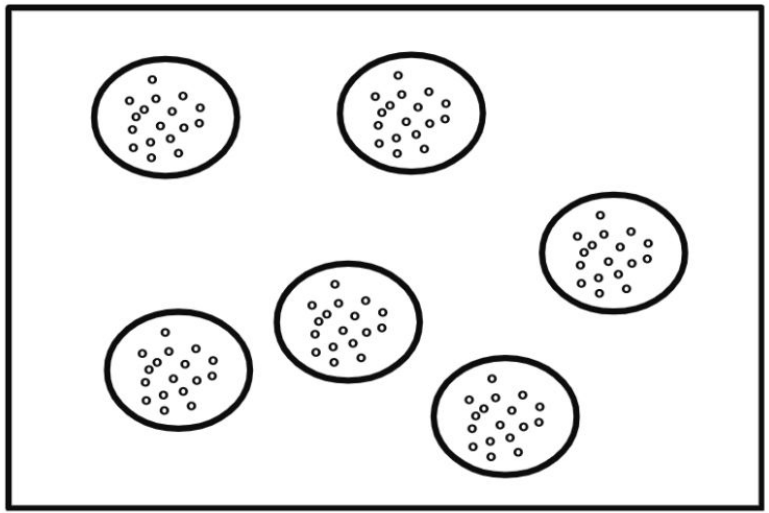
Likelihood model:

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{u} + \epsilon.$$

(Empirical) prior knowledge:

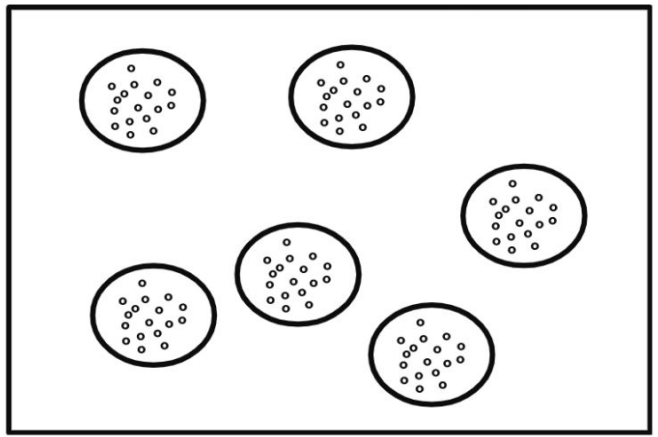
$$\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \mathrm{K})$$

- For example, students could be sampled from within classrooms, or patients from within doctors.
- When there are multiple levels, such as patients seen by the same doctor, the variability in the outcome can be thought of as being either within group or between group.
- Patient level observations are not independent, as within a given doctor patients are more similar.
- Units sampled at the highest level (in our example, doctors) are independent.
- The figure below shows a hierarchical data set (such as the example where the dots are patients within doctors, the larger circles)



Possible Solutions

- Aggregate
  - Does not really take advantage of all the data
- Analyzing data from one unit at a time
  - This does work, but there are many models, and each one does not take advantage of the information in data from other doctors
  - This can also make the results “noisy”



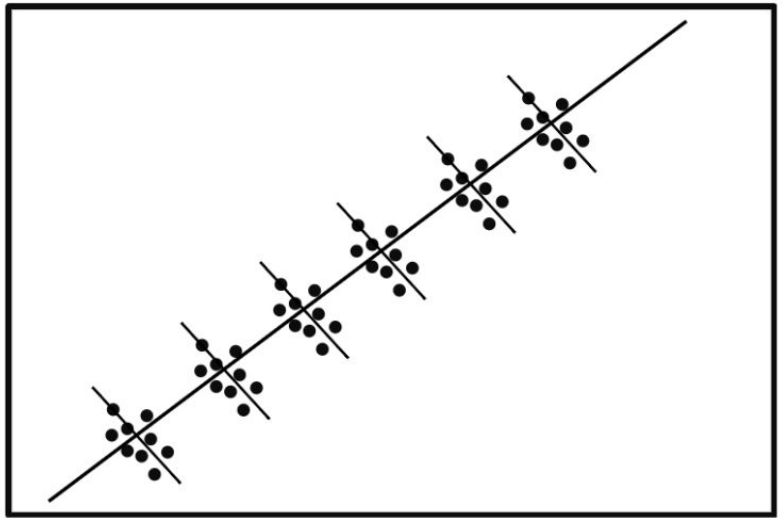
Using Linear Mixed Models

- Tradeoff between the two alternatives
- There are important reasons to explore the difference between effects within and between groups

Within each doctor, the relation between predictor and outcome is negative

Between doctors, the relation is positive

LMMs allow us to explore and understand these important effects



Fixed Effect

- Parameter that does not vary
- For example, we'll assume there is some true regression line in the population θ and we get some estimate of it

$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{u} + \epsilon$$

Random Effect

- Parameters that are themselves random variables
- For example, we could say that **u** is distributed as a random normal variate with mean 0 and covariance K,

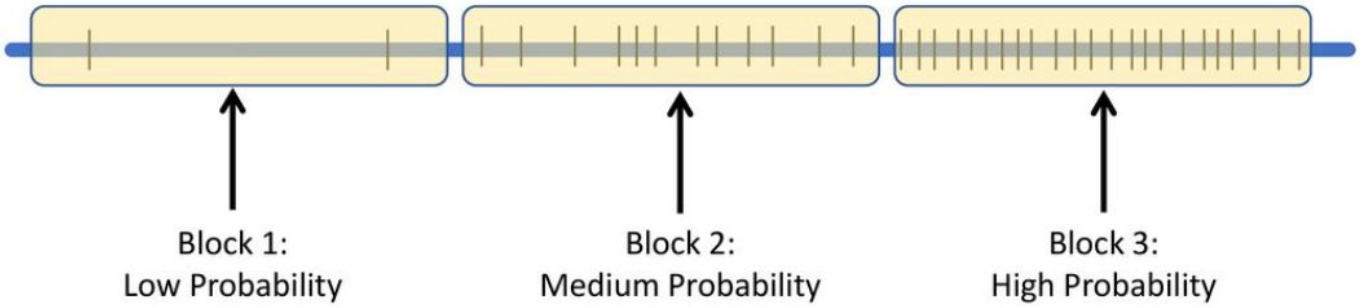
$$\mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{u} + \epsilon$$

$$\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \mathrm{K})$$

Linkage Disequilibrium Score Regression (LDSR)

- One of multiple methods used to estimate magnitude and direction of shared genetic effects between phenotypes
  - E.g. shared genetic influences on schizophrenia and bipolar disorder
- Other methods include twin/family studies, PRS, and GCTA

Logic = a causal variant in a haplotype block in strong disequilibrium is more more likely to have a high association with each loci than one in a block with weak disequilibrium.

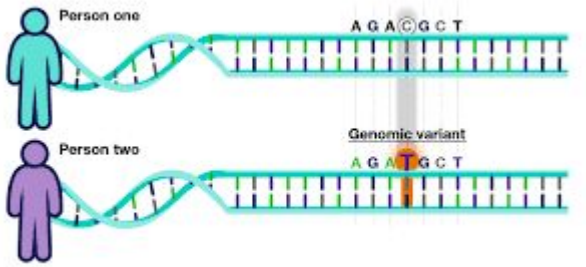


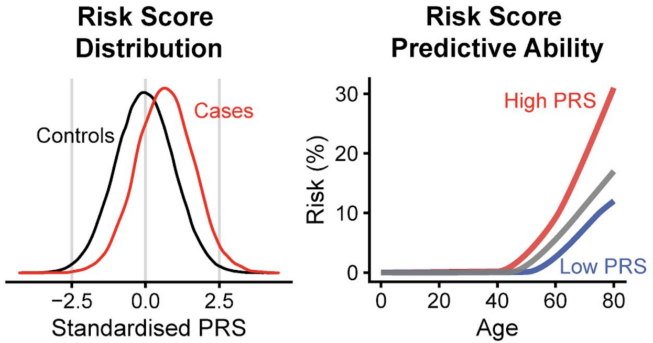
Twin and Family Studies

- Use information about phenotypes and family relations
- Infer relative importance of genetic factors in population (heritability)
- Quantify genetic effects between 2 phenotypes
- Capture all inherited genetic effects
- Have limitations

Polygenic Risk Scoring (PRS)

- Uses GWAS to construct individual-level metrics of genetic risk
- Outcome = the amount of phenotypic variance explained in a phenotype of interest
- Useful for *existence* of shared genetic effects
- Not useful for *magnitude*
- Time consuming and less widely applicable than LDSR





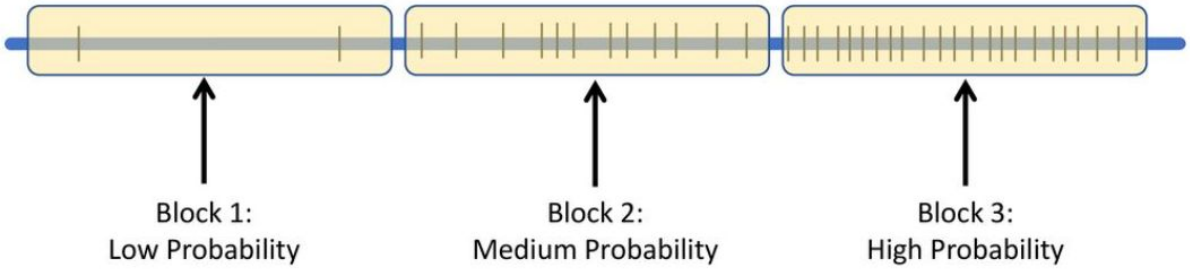
Genome-Wide Complex Trait Analysis (GCTA)

- Uses molecular genetic data to estimate heritability and shared genetic effects
- Assessed using maximum likelihood
- GCTA estimates of shared genetic effects have smaller standard errors than LDSR
- Better for N <= 3k

LDSR

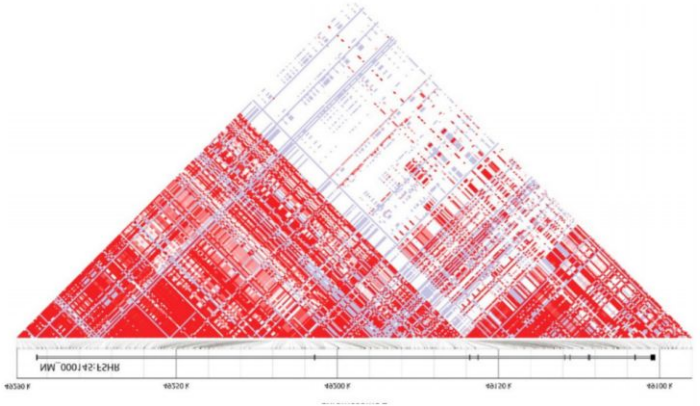
- Regression based model → estimates sample overlap & population stratification, heritability, shared genetic effects
- Applied to summary statistics from GWAS
- Overlap among participants is permissible
- Unbiased estimates of genetic correlation can be obtained even in presence of overlap

Logic = a causal variant in a haplotype block in strong disequilibrium is more more likely to have a high association with each loci than one in a block with weak disequilibrium.



Chi-Squared Statistic Inflation from Tagging SNPs in LD Regions

- The genome is comprised of regions of strong intermarker linkage disequilibrium (LD)
  - These are called LD regions or blocks
- Causal variants drawn uniformly at random from the genome are more likely to come from larger LD blocks
  - The longer a block, the higher the LD score regression
  - Chi-Squared statistic is ***inflated***



Inflation lets us infer heritability

$$\mathbb{E}(\chi^2 \text{statistic}) = 1 + (h_g^2 N/M)\text{LDscore}$$

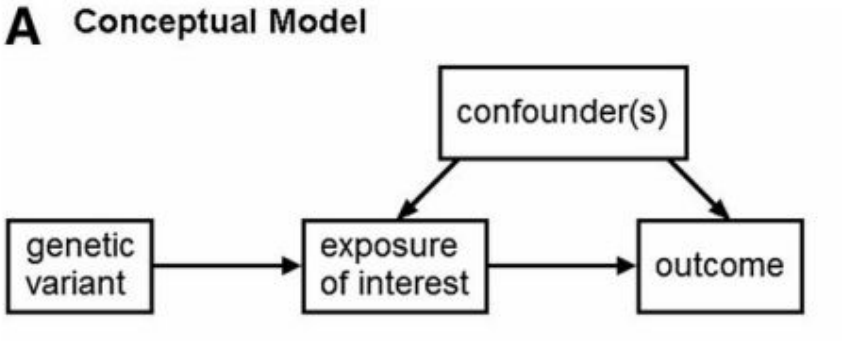
can calculate

can calculate

$(M = \text{markers},\ N = \text{samples})$

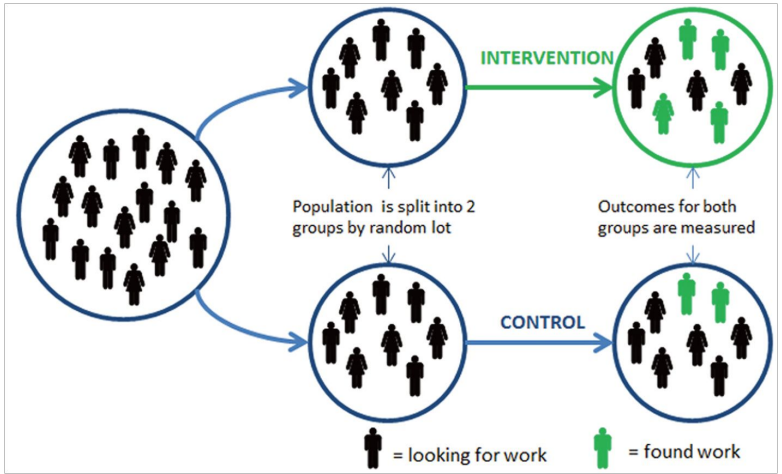
Mendelian Randomization (MR)

- Valuable tool → especially when randomized controlled trials to examine causality are not feasible and observational studies provide biased associations because of confounding or reverse causality
- MR addresses issues by using genetic variants as instrumental variables for tested exposure
- Time and cost efficient approach



Other Methods - Randomized, Controlled Trials (RCTs)

- Gold standard to establish causal relationships
- Proper randomization ensures study groups are comparable in all characteristic minus the exposure of interest
- Differences in outcome can be directly assign to effect of exposure
- RCTs can be costly, impractical, or sometimes unethical



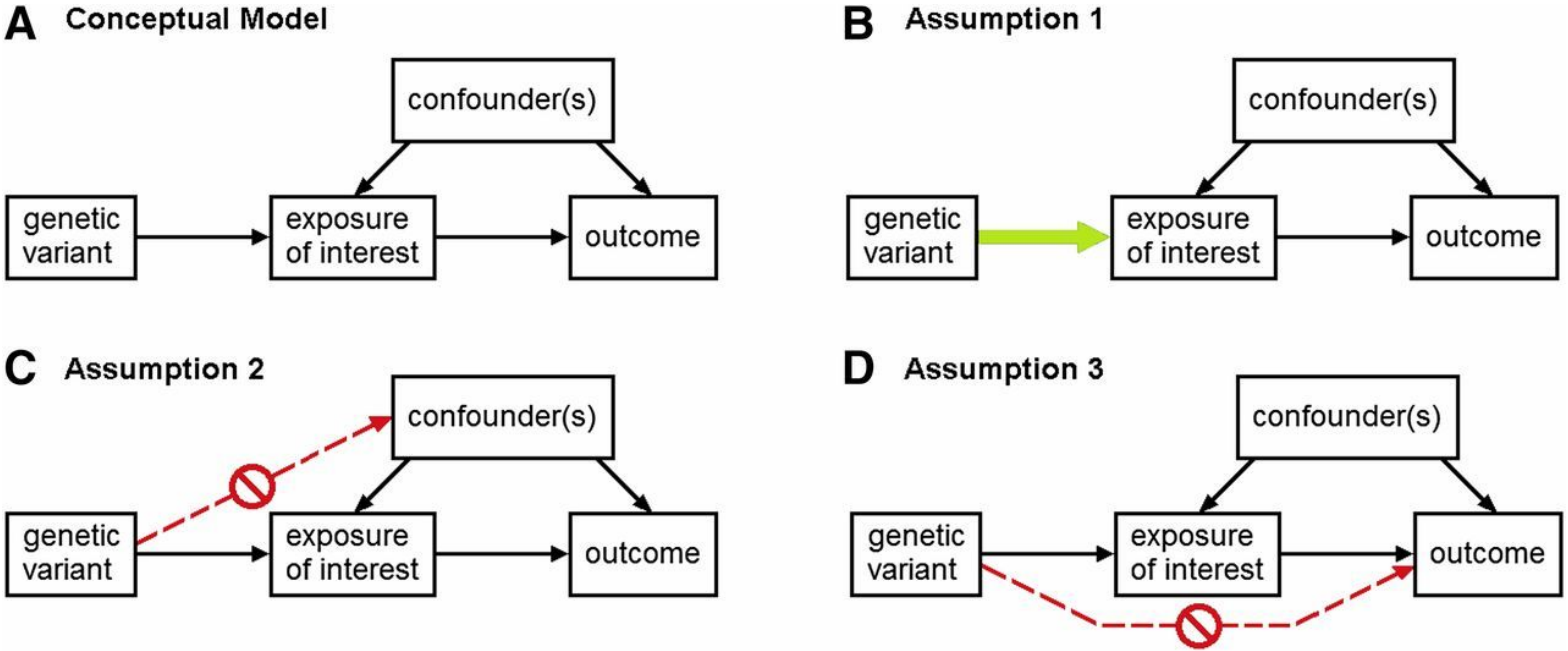
Other Methods - Observational Studies

- Study groups differ in several observed and unobserved characteristics
- Differences in outcome can then be attributed to any of these characteristics (or a combination!)
- Cannot directly establish causality
- Exposure-outcome association may not be a causal relationship but confounding or reverse causation

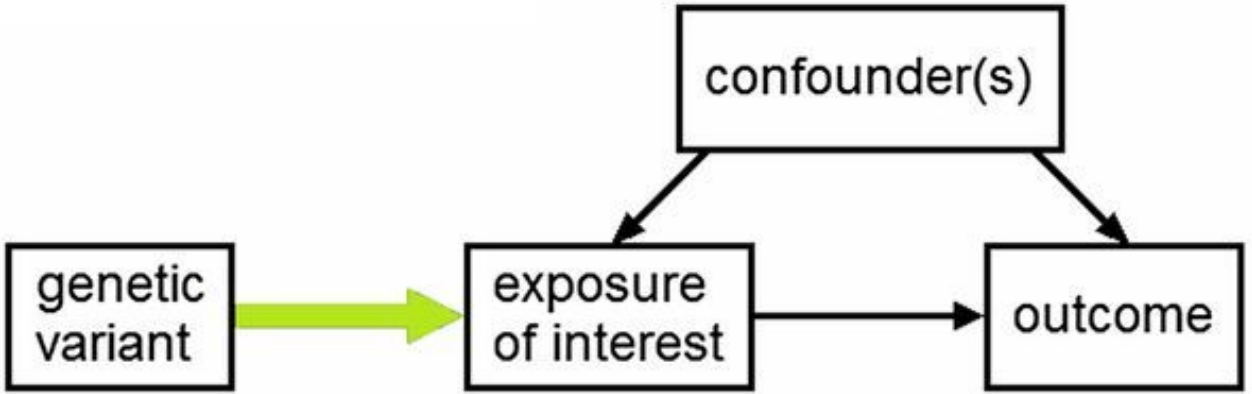
Instrumental Variables

- First introduced by economists and then adopted for medical statistics
  - Proposed as alternative statistical method to examine causality outcome associations while controlling for any confounder
- Mimic the randomized allocation of individuals to the express and ensure comparability of groups with respect to any confounder
- If available, effect of the exposure on the outcome can be unbiasedly estimated and causality of an observed association can be assessed

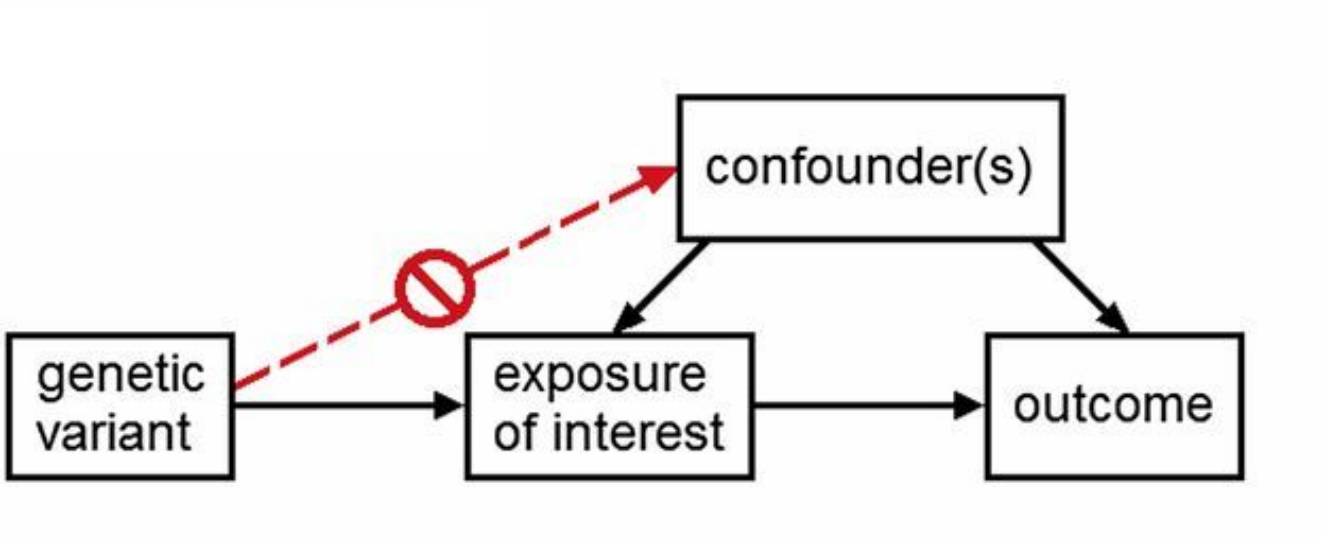
Core Assumptions for Choosing Genetic Instrumental Variable (GIV)



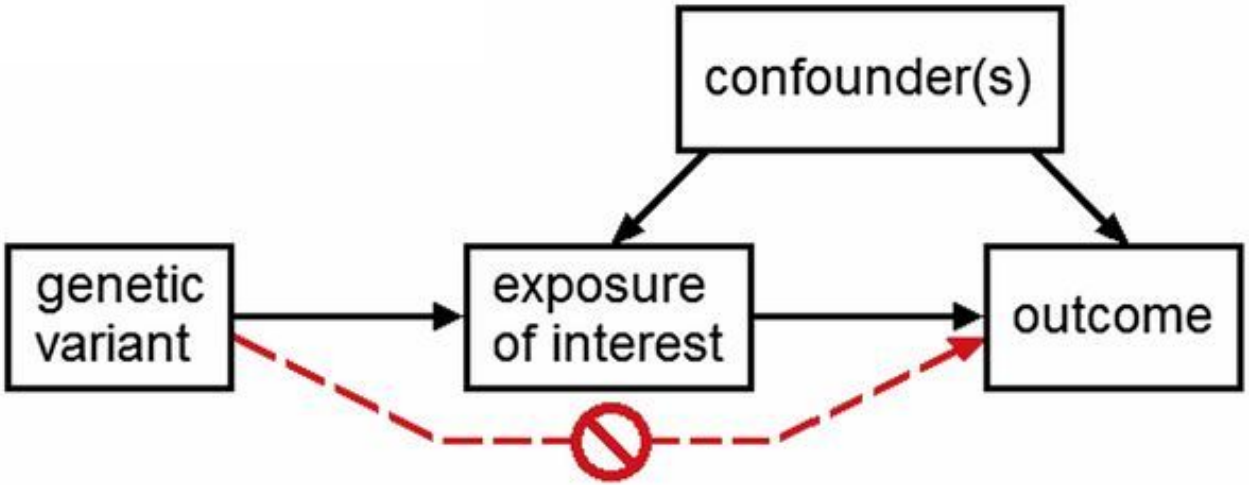
1: GIV must be reproducibly and strongly associated with the exposure



2: GIV must not be associated with confounders



3: GIV is only associated with the outcome through exposure



Conducting an MR Study

1) Define the presumed causal association to be investigated
2) Choose (≥ 1) genetic variant to be used as the instrumental variable
3) Evaluate core assumptions and discuss applicability
4) Carry out statistical MR analysis
5) Interpret and discuss results

Simple Linear Example of a Statistical MR Analysis

- We will look at effects on the exposure of C-reactive protein (CRP) via the GIV to estimate the outcome of body mass index
- causal effect of exposure X = CRP
- outcome Y = BMI
- G = GIV
- Linear model for effect estimation can be given by: $\hat{\beta}_{MR} = \hat{\beta}_{Y \sim G} / \hat{\beta}_{X \sim G}$
- $\hat{\beta}_{MR}$ is Wald ratio estimate which represents causal effect estimate obtained from $\hat{\beta}_{Y \sim G}$ and $\hat{\beta}_{X \sim G}$ which are regression coefficients from the regression of the outcome on the GIV and the regression of the exposure on the GIV (respectively)