UMR 1283 8199
Fonctional (Epi)genomics and Mechanisms of Type 2 Diabetes and Related Diseases

Mendelian Randomization

Mathilde Boissel
mathilde.boissel@univ-lille.fr / mathilde.boissel@cnrs.fr

16/03/2021

Context

Training

The **Mendelian Randomization** (**MR**) course hold in November 2020, provided by the MRC Biostatistics Unit from Cambridge University (UK). (More detail here http://mendelianrandomization.com/index.php)

In this feedback, we will see

- ✓ Introduction & motivations of MR
- ✓ Paradigm & Assumptions
- ✓ Methods on Individuals level data
- ✓ Methods on Summary level data and Two Samples MR
- ✓ Robust Methods, Issues and Sensitivity analyses

What is Mendelian Randomization

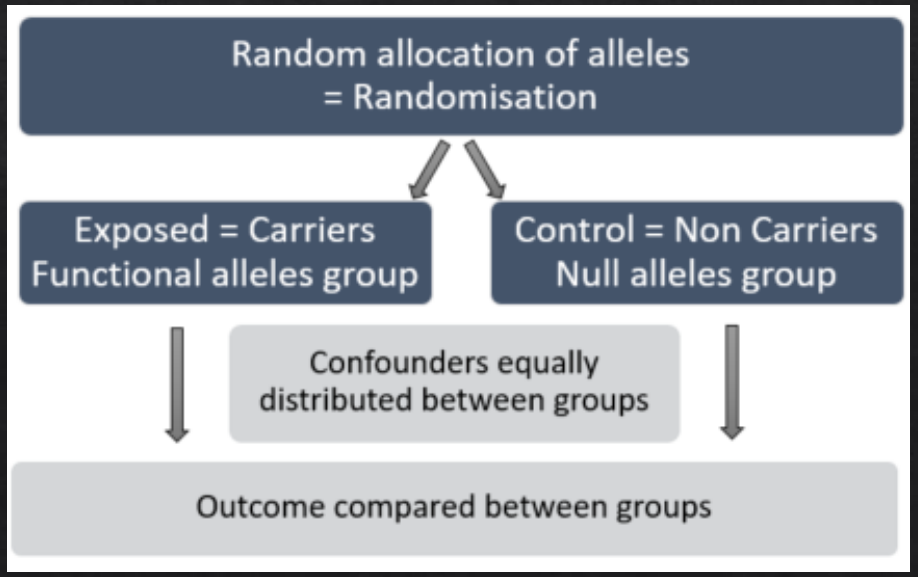
- Instrumental variable (**IV**) analysis using genetic variants to assess causal relationships
- Candidate gene study using biological knowledge about function of genetic variants
- Target validation for pharmaceutical priorization of modifiable risk factor / pathways
- Natural experiment to exploit naturally-occurring variation analogous to a randomized trial
- Use of genetic variants as proxies for treatments / interventions to exploit human as the model organism
- Investigation of shared heritability of exposure and outcome variable

- Short TED Med Live : https://youtu.be/rjMwcTttKoQ
- Lecture by George Davey Smith, NTNU trial : https://youtu.be/Whut4Yo-x-A

Introduction

- Correlation is not causation.
- MR is not a simple association test : with its instrumental variable it answers to causality. Use something stable (like genetic) as an instrumental variable to represent a risk factor of interest. Because it's determined at conception, the genetic sequence can't be influence by the disease outcome.
- Question of MR study : variation in risk factor impact the outcome ?

MR is analogous to a randomized trial : treatment vs placebo, with all confounders equally distributed.

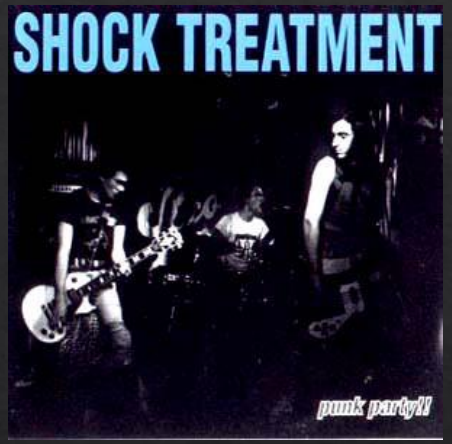


Motivation

- Financing concern.
- Failure in clinical trials.

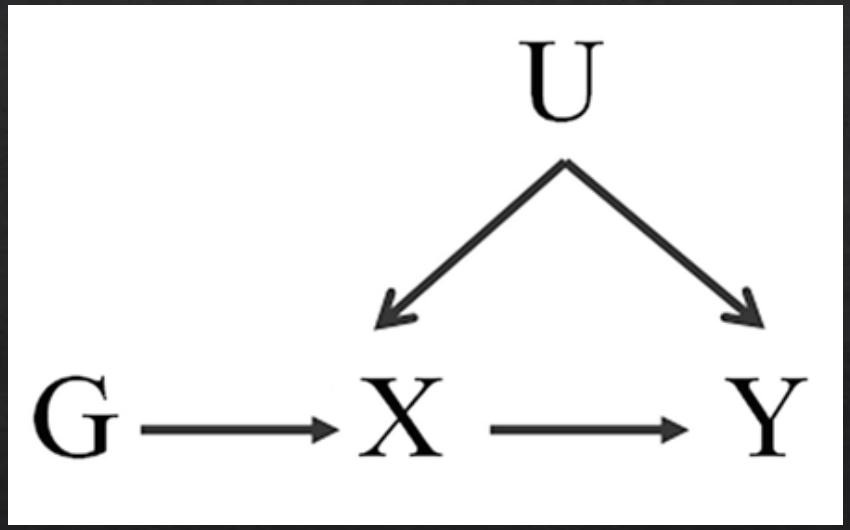
  Genetic help prevent ? X wasn't a good target, dev drug to modify this pathway won't change the outcome.

- Exploits a natural experiment assuming that genetic variants are quasi-randomized. Genetic variants are proxies for disease treatment (formally instrumental variables)
- **In MR context**, an association between a genetic variant and the outcome implies a causal effect of the risk factor on the outcome.



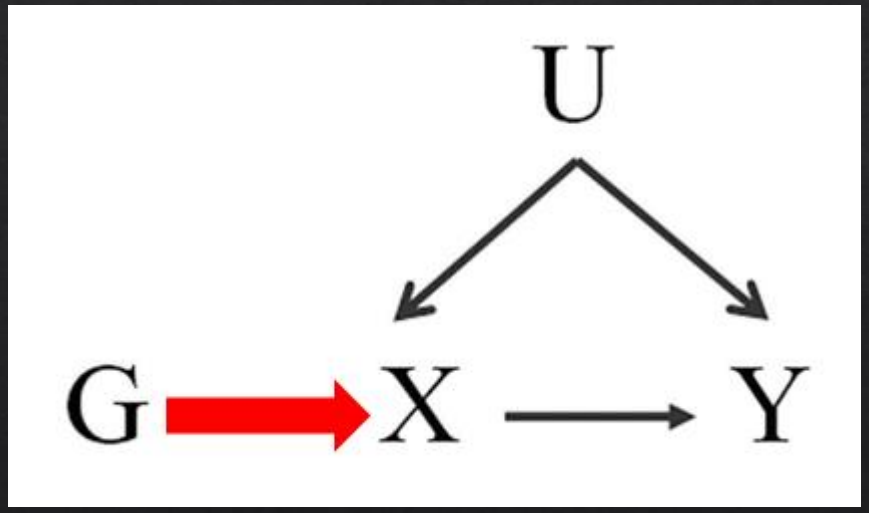
Paradigm

We want to study the causal effect of exposure X on outcome Y using a genetic variant G, in the presence of confounding U.



Assumptions

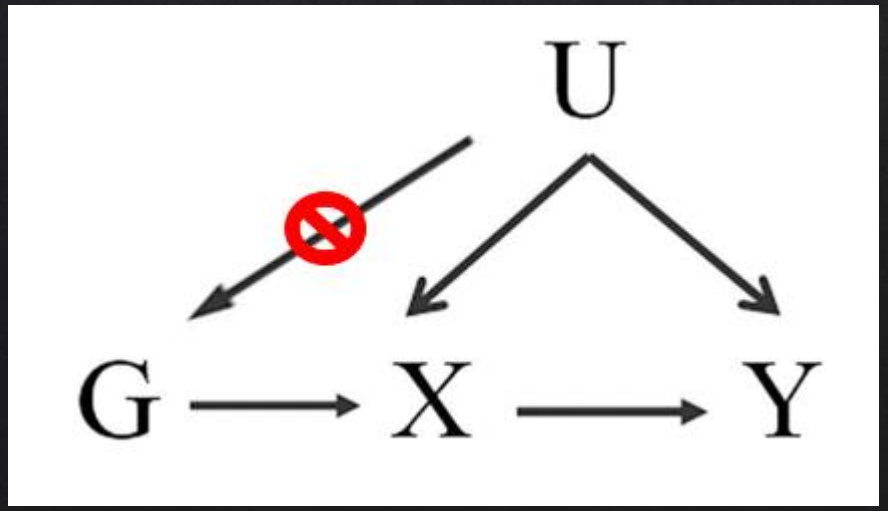
i. **Relevance**
   G is associated with the risk factor (of interest) X
   Not necessary for G to be a "causal variant" (consequence of LD)

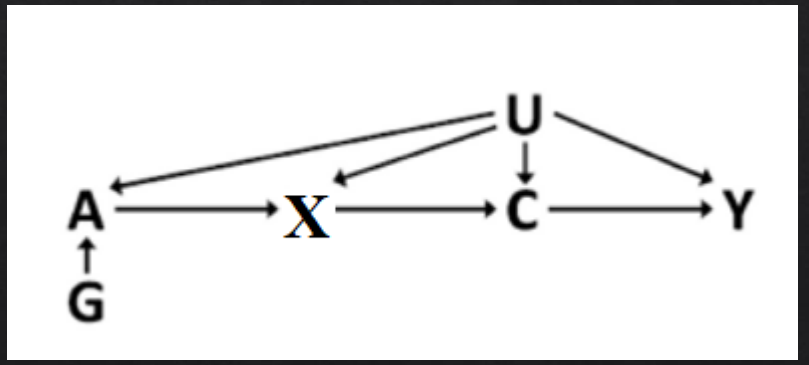


Assumptions

i. **Relevance**
   G is associated with the risk factor (of interest) X
   Not necessary for G to be a "causal variant" (consequence of LD)

ii. **Exchangeability i.e. "No pleiotropy" assumption**
   G is not associated with any potentially confounding variable (U)



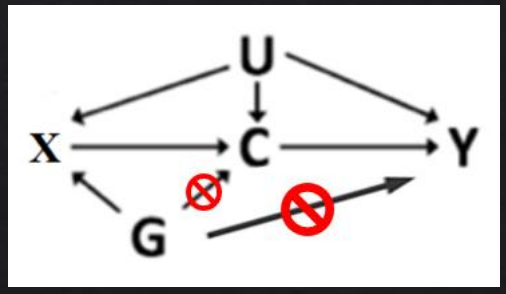


Assumptions

i. **Relevance**
   G is associated with the risk factor (of interest) X
   Not necessary for G to be a "causal variant" (consequence of LD)

ii. **Exchangeability i.e. "No pleiotropy" assumption**
   G is not associated with any potentially confounding variable (U)

iii. **Exclusion restriction**
   The only way G can influence (Y) the outcome is via risk factor (X)
   If you intervene on the genetic variant but keep the risk factor constant, then the outcome will not change, *i.e.*
   G and Y are conditionally independent given the risk factor X and U confounders.



Assumptions

Reasons for violations of assumptions

- Pleiotropy : some genetic variants act in multiple causal pathways
- Genetics variant considered in the MR study must be independent :
  be carefull to prunned variants, based on linkage desequilibrium
- Population stratification (like different ethnic group) :
  so important to adjust with PCs or restrict to population not mixed
- Canalization : in some cases, a disruptive variant may mean that a compensatory mechanism is switched on to compensate for the lost biological function

MR test H_0 : causal effect $= 0$ vs H_1 : causal effect $\neq 0$

Estimation vs Testing

To estimate the magnitude of causal effect of the risk factor on the outcome previous assumption are needed to test H0, but not sufficient to estimate the magnitude.

Additional assumptions needed to estimate the magnitude of causal effect.

Assumption for Causal Estimation

- **Average causal effect** : $E[Y(X = x + 1) - Y(X = x)]$
  - Homogeneity : the effect of X on Y is assumed constant in all individuals
  - Linearity : the effect of X on Y is assumed linear
  - No effect modification : the effect of X on Y is assumed not to depend on U

- **Local average causal effect** / compiler average causal effect
  - Monotonicity : the effect of G on X is the same direction for all individuals
  - Genetic "compliers" are individuals in pop for whom genetic variant affects the risk factor

Bradford-Hill's Criteria

- Plausibility (function of genetic variant (G) acts related to the risk fact / outcome ?)
- Consistency (is same direction of causal effect predicted by multiple G ?)
- Biological gradient (dose-response relationship to G association with risk fact and outcome ?)
- Strength (G explain much variance in the risk factor ?)
- Specificity (G association with many risk fact / outcomes ?)
- Coherence (causal relation concordant with observational data ?)

Methods on Individual Level Data

Ratio Method

Genetic association estimates are related via a simple linear regression model

$$(1)\ \ \widehat{\boldsymbol{\beta}}_{\boldsymbol{Y_j}} = \boldsymbol{\theta} \times \widehat{\boldsymbol{\beta}}_{\boldsymbol{X_j}} + \boldsymbol{\varepsilon_j}$$

- where $var(\varepsilon_j) = se^2\left(\hat{\beta}_{Y_j}\right)$
- Genetic association with the outcome equals to a causal effect θ times genetic association with the exposure.
- Interpretation : **θ is increase in Y when we intervene and change X by 1 unite**.
- linearity assumption : X linearly affects Y
- $\beta_y = 0 <=> \theta = 0$

Ratio estimate for variant j $\quad (2)\ \ \theta_{Y_j} = \dfrac{\widehat{\beta}_{Y_j}}{\widehat{\beta}_{X_j}}$

The causal effect of the risk factor X on the continuous outcome Y can be estimated as the per allele genetic association with the outcome $\widehat{\beta}_{Y_j}$ divided by the per allele genetic association with the risk factor $\widehat{\beta}_{X_j}$.

Ratio Method

The standard error of the causal estimate can be calculated simply as the standard error of the genetic association with the outcome $se(\hat{\beta}_{Y_j})$ divided by the genetic association with the risk factor $\hat{\beta}_{X_j}$ .This is the simplest form of the standard error, and is the first-order term from a delta method expansion for the standard error of a ratio.

Standard error of ratio estimate (first order)

$$(3)\quad se_{first}(\theta_{Y_j}) = \frac{se(\hat{\beta}_{Y_j})}{\left|\hat{\beta}_{X_j}\right|}$$

The above approximation does not account for the uncertainty in the denominator of the ratio estimate. This can be taken into account using the second term of the delta method expansion:

Standard error of ratio estimate (second order)

$$(4)\quad se_{second}(\theta_{Y_j}) = \sqrt{\frac{se(\hat{\beta}_{Y_j})^2}{\hat{\beta}_{X_j}^2} + \frac{\hat{\beta}_{Y_j}^2 se(\hat{\beta}_{X_j})^2}{\hat{\beta}_{X_j}^4}}$$

Ratio Method

In practice, in R, you can just run

```
by1 = lm(y~g1)$coef[2] #Genetic asso with the outcome
bx1 = lm(x~g1)$coef[2] #Genetic asso with the exposure
beta.ratio1 = by1/bx1
beta.ratio1 #Ratio estimate for g1
```

With Binary Outcome

Gene-outcome association : Use logistic regression

```
glm(y.bin~g1, family=binomial)
```

Gene-risk association : Use linear regression **in controls only**

```
lm(x[y.bin==0]~g1[y.bin==0])
```

In a case control setting, it is usual to regress the risk factor on the genetic variant in controls only because the case-control sample is generally unrepresentative of the population and if measurements of the risk factor are made post-disease = **Avoid possible reverse causation**.

Two-Stage Least Squares Method

When multiple variants G as Instrumental Variables

Two-Stage Least Squares (**2SLS or TSLS**) is performed by:

$$(5)\quad X \sim G_1 + G_{\ldots} + G_j$$

i.e. regressing the risk factor on all the genetic variants in the same model, and storing $\hat{X}$ the fitted values of the risk factor.

$$(6)\quad Y \sim \hat{X}$$

i.e. a regression is then performed with the outcome on the fitted values of the risk factor.

- With a single instrumental genetic variation, 2SLS and Ratio method are identical.
- With multi genetic variants as IV, the 2SLS estimate is a weighted average of the ratio estimates.
- 2SLS need to adjust the standard error
  (by hand must adj calculation, or adj ever implemented in MR R-packages)

Two-Stage Least Squares Method

In practice, in R, you may want to run
step 1 `fitted.values=lm(x~g1+g2+g3+g4)$fitted`
step 2 `lm(y~fitted.values)`

In real practice,

Performing two-stage least squares by hand is generally discouraged, as the standard error in the second-stage of the regression does not take into account uncertainty in the first-stage regression.

The R package performs the two-stage least squares method using the `ivreg` function: `ivreg(y~x|g1+g2+g3+g4, x=TRUE)`

(from the R package `AER`, loaded with `ivpack`)

With Binary Outcome
First stage with a binary outcome is only conducted on controls
Second step is done with a logistic regression

F-Statistic

The F-statistic from the regression of the risk factor on the genetic variant(s) is used as a measure of 'weak instrument bias', with smaller values suggesting that the estimate may suffer from weak instrument bias.

Care should be taken when interpreting the F-statistic from observed data. Some studies recommend excluding genetic variants if they have a F-statistic less than 10. Using such a stringent cut-off may introduce more bias than it prevents, as the estimated F-statistic can show considerable variation and may not provide a good indication of the true strength of the instrument.

Methods on Summary Level Data

Summary Level Data

Many consortia release GWAS summary
- GWAS Catalog
The NHGRI-EBI Catalog of human genome-wide association studies
- Uk Bio Bank
361,194 subjects, 4,203 phenotypes, covariates adjustement as 2 pcs, sex and age (http://www.nealelab.is/uk-biobank), and pan-ancestry genetic analysis of the UK BioBank (https://pan.ukbb.broadinstitute.org)
- FinnGen consortium
96,499 samples, 16,152,119 variants and 1,122 endpoints phenotypes available
- PhenoScanner
A database of human genotype-phenotype associations, Available in `library(MendelianRandomization)` with function `pheno_input()`

Two Samples (2-Samples) MR

When exposure (risk factor) estimates is in a cohort 1 (G => β_x => X) and **outcome estimates is in a cohort 2 (G => β_Y => Y).**

Pitfalls in match2-samples MR

- Harmonization :
  match effect and non effect allele + genome build
  very important to harmonize effect allele REF/ALT
  (may consider $-\beta$ if REF/ALT must be flipped)
- Population structure (similar allele freq in both cohorts)
- No sample overlap between cohorts (induce bias)
- Both studies have same adjustment (collider bias)
- Instrument selection available in both studies
- Effect heterogeneity (robust method with summarized data)

Inverse Variance Weighted Method

Recall ratio equation (**1**) $\widehat{\boldsymbol{\beta}}_{\boldsymbol{Y}_j} = \boldsymbol{\theta} \times \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}_j} + \boldsymbol{\varepsilon}_{\boldsymbol{j}}$ with following assumptions :
X linearly affects Y and additive effect of genetic variant j, all J variants are valid IVs and homogenous effects in the two cohorts.

If assumptions hold :

- Clear linear trend in G asso, starting at the origin with a gradient $\hat{\theta}$
- No heterogeneity, e.g. Cochran's $Q \approx J - 1$

How to combine n different ratio estimates $\widehat{\theta_j}$?With the Inverse Variance Weighted (IVW) estimate

$$(7)\ \widehat{\boldsymbol{\theta}}_{\boldsymbol{IVW}} = \frac{\sum_{j=1}^{n} \widehat{\theta}_j / var(\widehat{\theta}_j)}{\sum_{j=1}^{n} 1/var(\widehat{\theta}_j)} \qquad (8)\ var(\hat{\theta}_j) = \frac{var(\beta_{Y_j})}{\beta_{X_j}^2} = \frac{\sigma_{Y_j}^2}{\beta_{X_j}^2}$$

Important assumption : genetic variant are independent thus θ_j are independent.

Assessing homogeneity : using the Cochran's heterogeneity stat Q.

Inverse Variance Weighted Method

In practice, we can use summarized data (the genetic associations with the risk factor and with the outcome, with their standard errors) to estimate the causal effect of the risk factor on the outcome via the IVW method:

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{IVW}}\ estimate = \frac{\sum_j \widehat{\beta_{Y_j}}\,\widehat{\beta_{X_j}} se(\widehat{\beta_{Y_j}})^{-2}}{\sum_j \widehat{\beta_{X_j}}^{\,2}\, se(\widehat{\beta_{Y_j}})^{-2}}$$

The standard error of the IVW estimate is:

$$standare\ error\ of\ IVW\ estimate = \sqrt{\frac{1}{\sum_j \widehat{\beta_{X_j}}^{\,2}\, se(\widehat{\beta_{Y_j}})^{-2}}}$$

In real practice, in R, we use the functions `mr_input` and `mr_ivw` to run the model **IVW**.

Equivalently we can run a **weighted regression**

```
lm(betaY ~ betaX - 1, weights = 1/betaYse^2)
```

N.B. : `- 1` is important for no intercept

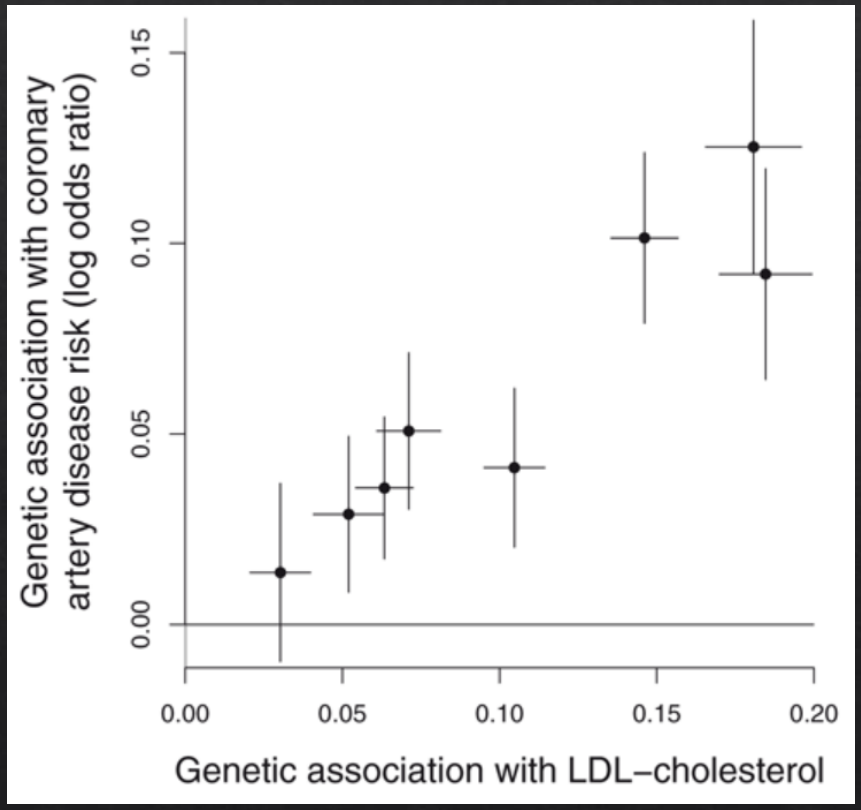
Graphics

Scatter plot of genetic associations with the outcome (vertical axis) $\widehat{B_Y}$ against genetic associations with the exposure (horizontal axis) $\widehat{B_X}$.

The lines on each points represent the 95% confidence intervals for the genetic associations with the exposure and with the outcome.

The ratio estimate for each variant is the gradient of the line from the origin to the data point.

The slope of the regression combining n θ.

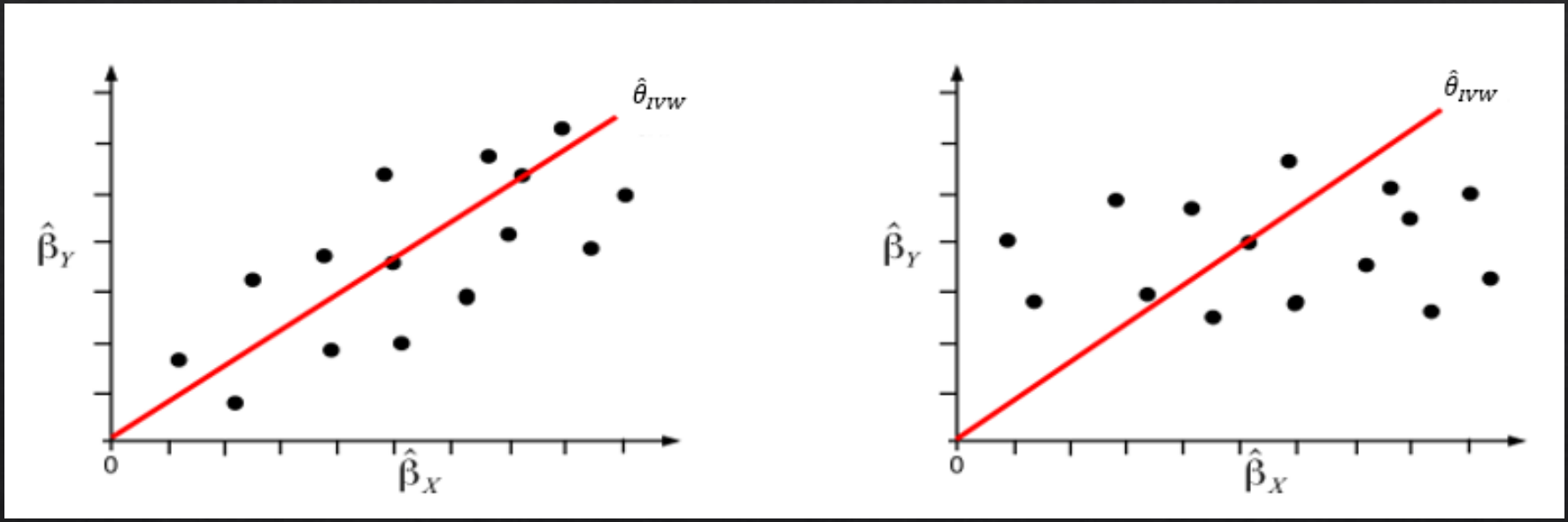


Guidelines for performing Mendelian randomization investigations, Stephen Burgess et al.

Graphics

Pleiotropic variants may appear as outliers on the scatter plot or

fit with intercept ≠ 0 => need Robust Methods.



Balanced vs directional pleiotropy, by Apostolos Gkatzionis

Robust Methods

IVW and MRE Model

Recall **IVW method,** under standard assumptions, use model **without any intercept term**

$$\hat{\beta}_{Y_j} = \theta \times \hat{\beta}_{X_j} + \varepsilon_j \text{ where } var(\varepsilon_j) = se^2(\hat{\beta}_{Y_j})$$

Bias due to pleiotropy

$\boldsymbol{\beta_{Y_j} = \alpha_j + \theta \times \beta_{X_j}}$ where the **intercept** $\boldsymbol{\alpha_j}$ represents the pleiotropic effect of the variant on the outcome (i.e. not via X)

Condition for no bias in IVW method

- Strength independent of direct effect (the **"InSIDE"** assumption) :
  The pleiotropic effects α_j are independent of the instrument strength β_{X_j}
- The mean pleiotropic effect is zero, i.e. $\overline{\alpha} = 0$

These scenario represent **"balanced" pleiotropy**, there is increased heterogeneity but non bias.

Multiplicative Random Effects (MRE) model can account for additional heterogeneity with ϕ its over-dispersion parameter. So equation becomes

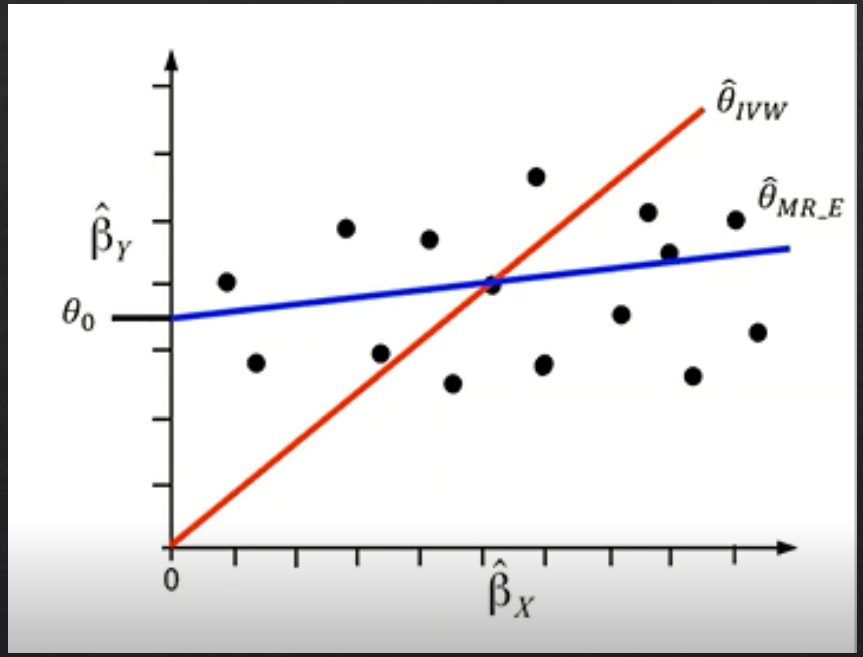
$$\hat{\beta}_{Y_j} = \theta \times \hat{\beta}_{X_j} + \varepsilon_j \text{ where } var(\varepsilon_j) = \phi se(\hat{\beta}_{Y_j})^2$$

MR-Egger

MR-Egger regression relaxes the assumption that the average pleiotropic effect is zero (allow for directional pleiotropy) with an intercept regression model.

Still need the the **"InSIDE"** assumption : pleiotropic effects α_j are independent of the instrument strength β_{X_j}

- MR-Egger allow 100% of variants to be invalid, but require the InSIDE to be valid
- MR-Egger is more sensitive to InSIDE violations than IVW
- MR-Egger low power
- MR-Egger sensitive to outliers



Robust methods with summarized data session by Apostolos Gkatzionis

Median Based Estimation

- G with outlying ratio estimates
- The median of ratio estimates is more robust to outliers and can be used as estimate for θ
- The median is accurate estimate if a majority (>50%) of variants are valid instruments, especially with large samples sizes
- No InSIDE assumption required
- “Weighted Median” is estimated to incease precision

Mode Based Estimation

- Like median more robust than IVW if outliers
- So take ratio estimates, fit a smooth curve (kernel) to model their distribution and use the mode of that distribution as the overall causal effect estimate
- Required a plurality of genetic variants to be valid (not >50%)
- Simple and weighted version, like median
- Can be sensitive to weak instruments.

Recent Robust Methods

MR-Presso (Verbanck et al. 2018)
Outlier detection and deletion method, performs an overall heterogeneity test and test of validity for each variant. Works well when there are few strongly pleiotropic variants, not so well with many moderately pleiotropic variants
MR-Robust (Rees et al. 2019)
Fits weighted linear regression like IVW but adds a loss function to reduce the influence of outlying variants
Contamination Mixture (Burgess et al 2019)
Fits a mixture model to ratio estimates. Relied on a plurality assumption, similar to the mode
Can yield disjoint confidence intervals, has low MSE but may have inflated Type I errors
MR-Mix (Qi & Chatterjee 2019)
Also a mixture-model approach
Requires summary statistics for variants not associated with X
Good uncertainty quantification (nominal Type I error rates) but occasionally lacks power
MR-Ras (Zhao et al 2018)
Models pleiotropic term as normally distributed random effects, estimates using profile regression
Guarantees of asymptotic performance, adj for weak instrument bias
MR-clust (Foley et al 2020)
Clusters genetic variants into groups with similar ratio estimates
Tries to identify variants with similar biological function

More detail about comparison : Review paper by Slob & Burgess (2020)

Issues in MR

- **Heterogeneity of estimates :** Assessing homogeneity using the **Cochran's heterogeneity stat Q** and decision to how to account for it.
- **Overidentification** : Not all genetic variants are valid IV.
  if all IVs identify the same causal effect, then the estimates should all be similar.
  Equivalently, the residual from $(Y - \hat{\beta}X)$ should be uncorrelated with the IVs.
  Overidentification test is call **Sargan's test**.
- **Harmonise** effect alleles : **Palindromic SNP** to exclude.
- **Winner's curse** : the variants that "won" the competition to have the lowest p-value tends to have an overestimated association.
- **Power calculation** : Sample size calculation with a continuous or binary outcome can be affected by pleiotropy.
- **Weak instruments :** The **F-statistic** from the regression of the risk factor on the genetic variant(s) is used as a measure of 'weak instrument bias'. G that explain a small amount of variation in the exposure (small R2). asymptotically unbiais estimates but bias with small samples. IVs with F-stat less than 10 are considered to be "weak" (do not use for splited studies)

Issues in MR

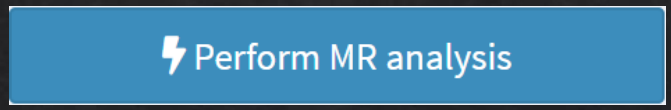
- **Allele scores :** A way to minimized weak instrument bias is to use an allele score (GRS) with unweighted score : $\sum_k g_{ik}$ (or weighted score : $\sum_k w_k g_{ik}$ with w_k reflecting the effect of genetic variant k on the exposure)
  - Advantages : The GRS is used as a single IV, Ratio method for a single IV has a median bias close to 0, Parsimony (for testing asso with cov)
  - Desadvantages : Each contributing variant must be a valid IV, Cannot investigate heterogeneity of estimated causal effects, One or few varaints may dominate the associations with the exposure
- **Selection / collider bias :** X and Y totally independent, or only select values > 1.5 Examples : intelligence and attractiveness, inappropriate adj (blood pressure and BMI), survivor bias in elderly (eg parkinson's disease), non representative samples (T2D and BMI), secondary prevention (is risk factor is a cause of first disease), Time to event data for disease progression
- **Binary Exposure** : Uncommon, interpretation difficult, can represent effect per % absolute change in exposure or per unit change in log odds of exposure.

Sensitivity Analyses

- **Strategy 1 : Exclude a key variant**
  A variant that explain 3 times more variance than other variants may dominate the MR estimates. Assessing whether there is still evidence for a causal effect when this variant is excluded from the analysis.
- **Strategy 2 : Check for other associations**
  - If a genetic variant is also asso with other unrelated trait, then it is potentially pleiotropic.
  - If a genetic var is asso with any other trait at $p < 5 \times 10^{-}8$ (set a genome wide threshold), then exclude that variant from the analysis.
  - If a genetic var is asso with a key potential competing risk factor at $p < 0.01$, then exclude that variant from the analysis.
- **Other strategies for increasing confidence in a finding …**
  - Scatter plot and heterogeneity test ; look consistency across different variants.
  - Robust methods (weighted median, MR-Egger, …)
  - MR-Egger intercept test avoid pleiotropy
  - Multivar MR
  - Positive Negative controls
  - MOST IMPORTANT : **use biological knowledge**

Usefull Links

- MR-Base web-app that allows the fitting of models without using R.
  May help to have fast results, examples of Cohorts, Traits, and just few QC…
  - Choose risk factors, outcomes, many datasets…
  - LD clumping : to not double count the same information, filter correclated SNPs
  - LD proxies : if in the 2nd dataset I can't found my SNP, use proxies (1=> super correlated)
  - palindromic SNPs : difficult to harmonize, may filter it
  - Allele harmonization : positive strand ?
  - Select methods, and go for it !



- Sample size calculation with a continuous or binary outcome:
  https://shiny.cnsgenomics.com/mRnd/
- Sample size calculation with a continuous or binary outcome:
  http://sb452.shinyapps.io/power/
- Fieller's theorem for confidence intervals with a single instrumental variable:
  http://sb452.shinyapps.io/fieller/
- Likelihood-based method (with heterogeneity test) with summarized data:
  http://sb452.shinyapps.io/summarized/

I thank speakers

Stephen Burgess
Apostolos Gkatzionis
Amy Mason
Verena Zuber
Dipender Gill
Andrew Grant

And Thank You For Your Attention !

Guidelines

Checklist for Reviewing Mendelian Randomization Investigations

- What is the primary hypothesis of interest? What is the motivation for using Mendelian randomization? What is the scope of the investigation? What and how many primary analyses?
- Data sources
  What type of Mendelian randomization investigation is this? One-sample or two-sample? Sample overlap? Summarized data or individual-level data? Drawn from same population? Relevance to applied research?
- Selection of genetic variants – how were the genetic variants chosen? Single or multiple gene regions?
  - Biological rationale?
  - GWAS analysis? If so, what dataset? What was the p-value threshold? Clumping?
  - Were genetic variants excluded from the analysis? Associations with pleiotropic pathways?
  - How else was the validity of genetic variants as instrumental variables assessed?
- Variant harmonization (for two-sample analyses)
  Was it checked that genetic variants were appropriately orientated across the datasets?
- Primary analysis
  What was the primary analysis? What was the statistical method? How implemented? Multiple testing? - Supplementary and sensitivity analyses
  What analyses were performed to support and assess the validity of the primary analysis?
  For example: stricter criteria for variant selection, assess heterogeneity, robust methods, subgroup analysis, positive/negative controls, 'leave-one-out' analyses, colocalization (single gene region)
- Data presentation
  How are the data and results presented to allow readers to assess the analysis and assumptions?
  For example: scatter plot, forest/funnel/radial plot, R2/F statistics, comparison of methods, power
- Interpretation
  How have results been interpreted, particularly any numerical estimates?

Flowchart while Reading a MR Paper …