Using genetic variants for causal inference using Mendelian randomization

2020-03-02 - Marc-André Legault

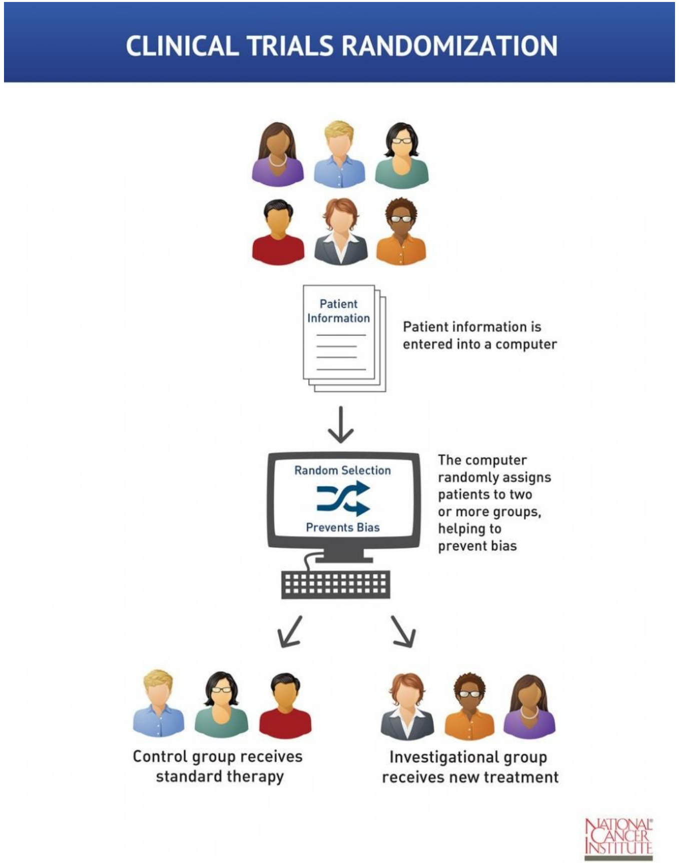
Causality in the medical sciences

A large fraction of medical research focuses on:

- Prediction of disease onset or disease course (prognosis)
  - Having a causal understanding may improve predictive power or generalizability
- Determining safe and efficacious interventions
  - Drugs
  - Surgical interventions
  - Recommendations, diets, etc.

The Randomized Clinical Trial (RCT)

- Start from a population and randomize
  - Any latent **confounder** should be distributed in the same way in both arms
  - No possibility of **reverse causation** of treatment
  - Investigator and patient **blinded** to treatment (accounts for placebo or experimenter bias)
- Statistical challenges
  - Compliance
  - Non-random dropout
  - Design, power, etc.


CLINICAL TRIALS RANDOMIZATION
Patient Information
Patient information is entered into a computer
Random Selection
Prevents Bias
The computer randomly assigns patients to two or more groups, helping to prevent bias
Control group receives standard therapy
Investigational group receives new treatment
NATIONAL CANCER INSTITUTE

https://www.cancer.gov/about-cancer/treatment/clinical-trials/what-are-trials/randomization/clinical-trial-randomization-infographic

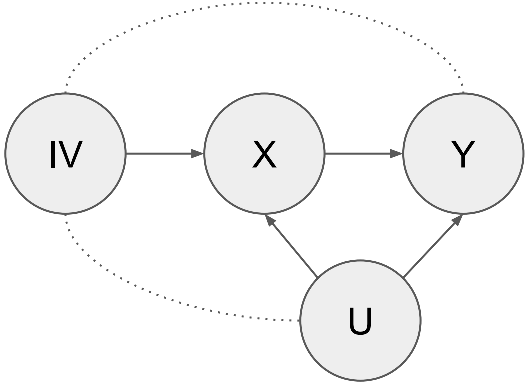
The Randomized Clinical Trial (RCT)

- One of the highest quality level of evidence (if well done)
- Often long and costly
  - Example of major adverse cardiovascular event trials where a pre-defined number of events need to occur
- Answers a very specific question: “Does treatment X work in clinical population Y and protocol Z”
  - Secondary or post-hoc analyses are usually considered hypothesis-generating
- (Rightfully) required for approval of treatment by health authorities

Instrumental variable approaches

Another approach to estimating causal effects is to use “instrumental variables”

- From the world of econometrics
- Goal: Estimate the effect of X on Y
  - In epidemiology X is the “exposure” (*e.g.* cholesterol levels) and Y is the “outcome” (*e.g.* cardiovascular disease)
- Traditional set of assumptions:
  - Relevance: IV correlated with X
  - Exclusion-restriction: IV not associated with Y conditional on U and X



Instrumental variable approaches

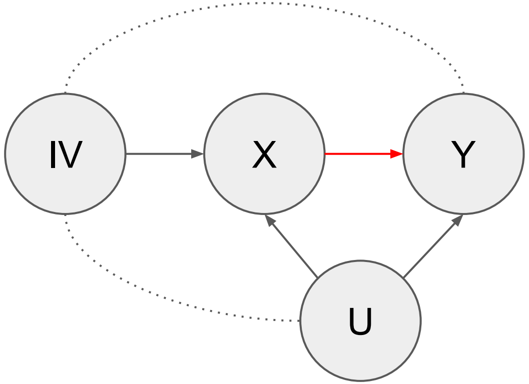
Another approach to estimating causal effects is to use “instrumental variables”

- If the assumptions are met, the causal effect of X on Y can be estimated
- The 2-stage least squares is the traditionally used method:
    - Stage 1. Estimate

      $$X \sim IV + covariates$$

    - Stage 2. Use predicted X from stage 1 to estimate causal effect

      $$Y \sim \hat{X} + covariates$$



What to use as the instrument variable?

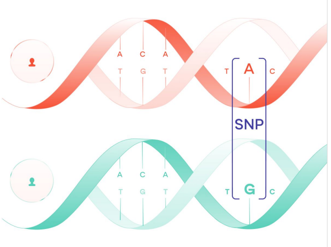
Few variables truly satisfy the IV assumptions

- Frequent source of conflicts in econometrics
- Assumptions are hard to verify statistically

Mendelian Randomization (MR)

Genetic variants are good candidate IVs!

- Common “polymorphisms” in the population



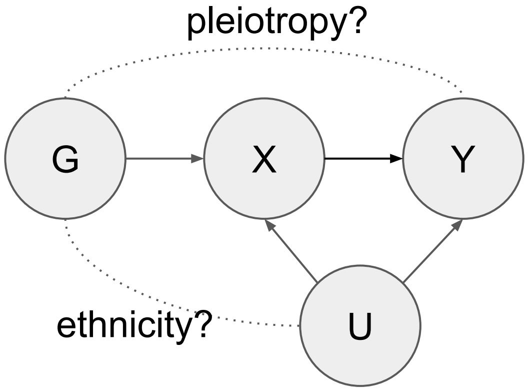
- “Easy” and affordable to screen in large cohorts
- Usually of low to moderate effect size because of natural selection
- Can be seen as a binomial random variable with n=2 (number of copies of the polymorphism)

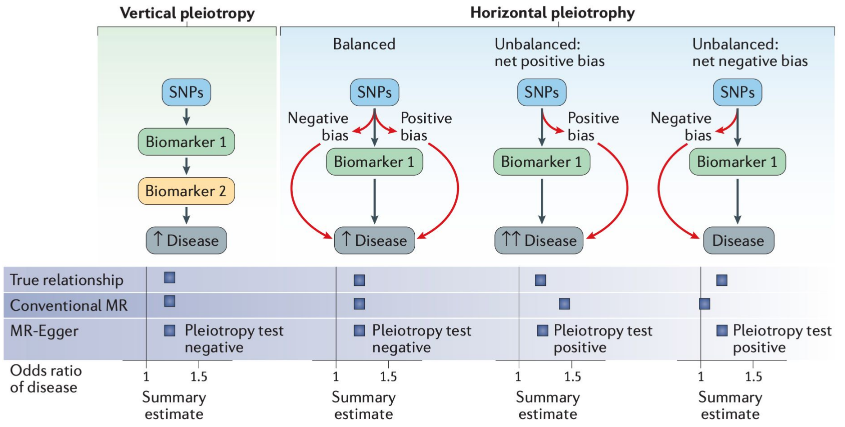
Image from: https://medium.com/@sanogenetics/snp-of-the-week-58e23927c188

Mendelian Randomization (MR)

Genetic variants are good candidate IVs!

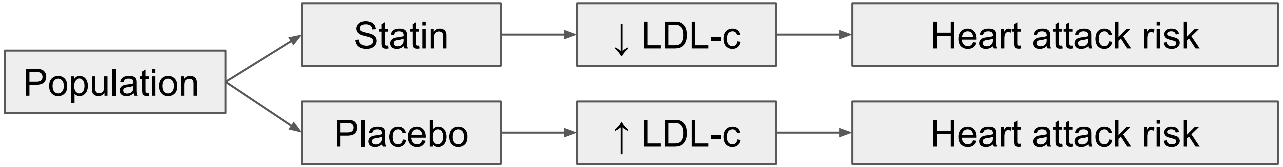
- Are inherited at birth and don’t change
  - Can’t suffer from reverse causation
  - Not very susceptible to confounding (except by ethnicity)
- The main problem is pleiotropy
  - **Vertical pleiotropy** (OK): G acts somewhere in the X pathway, but there are mediators
  - **Horizontal pleiotropy** (not OK): G acts on the X pathway, but also on an independent pathway that affects Y





From Holmes MV. *et al.* (2017) Nature Reviews. Cardiology

Example of LDL-c



Population

Statin

↓ LDL-c

Heart attack risk

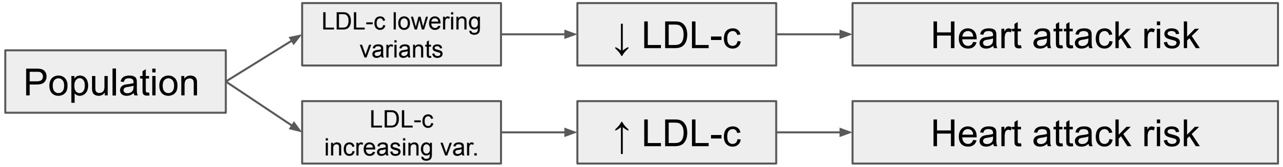
Placebo

↑ LDL-c

Heart attack risk

Artificial randomization

Randomization by the
random allocation of alleles at
birth



Population

LDL-c lowering variants

↓ LDL-c

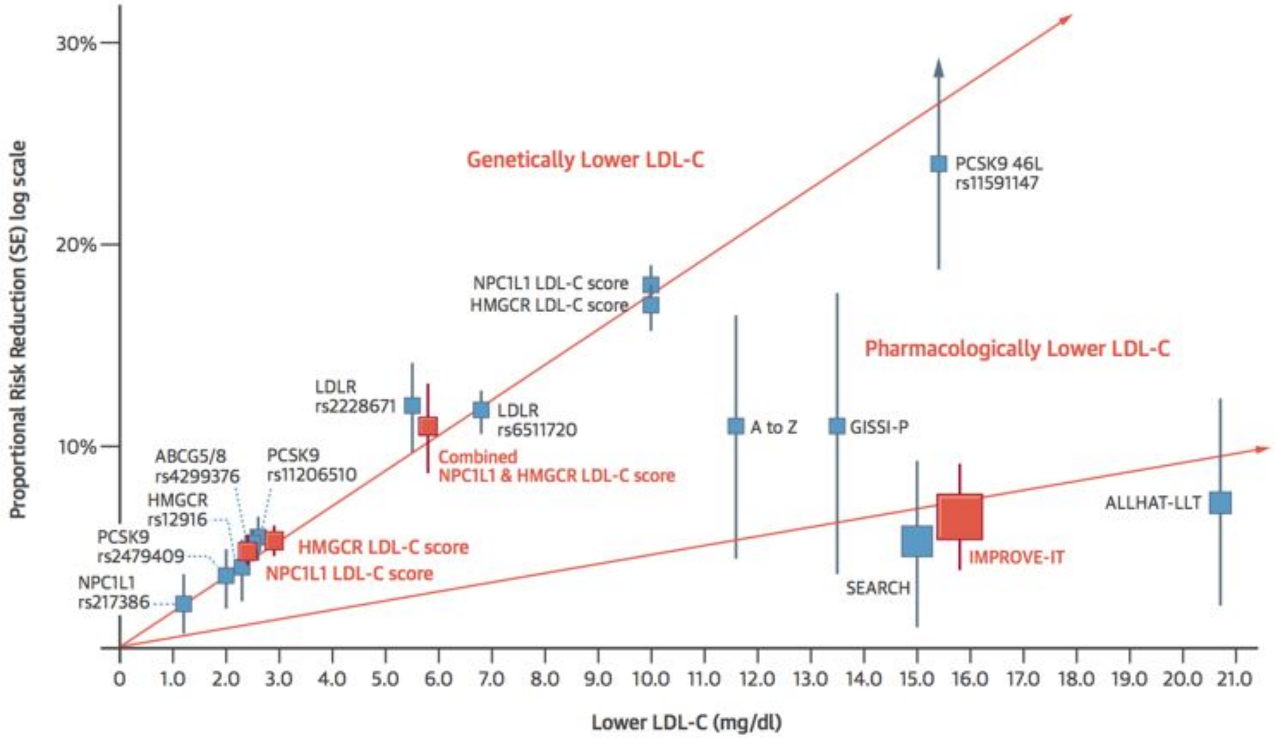
Heart attack risk

LDL-c increasing var.

↑ LDL-c

Heart attack risk

Example of LDL-c



Ference, B.A. et al. J Am Coll Cardiol. 2015; 65(15):1552-61.

- High concordance of effect estimates from RCTs and MR
- Overall, benefit of genetic reduction of LDL-c seems higher, why?
- When known drug target, risk of pleiotropy is small

Can we replace RCTs with MR?

No! Testing an intervention in a RCT remains the only way to truly test its causal effect

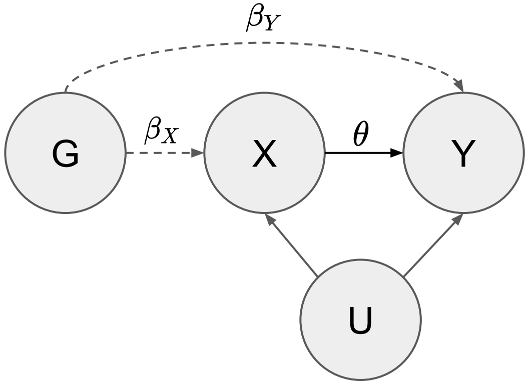
But,

- MR can predict result of RCT (which take time)
- MR can estimate the dose-response curve to inform RCT design
    - What is the expected effect?
    - How many individuals or events are needed to have the statistical power to detect this effect?
- MR can assess differential effects in other clinical populations
    - RCTs are characterized by many inclusions and exclusions
- MR can further understanding of causal mechanism

Ratio estimate

The simplest MR estimate for a single genetic variant is the “ratio estimate”

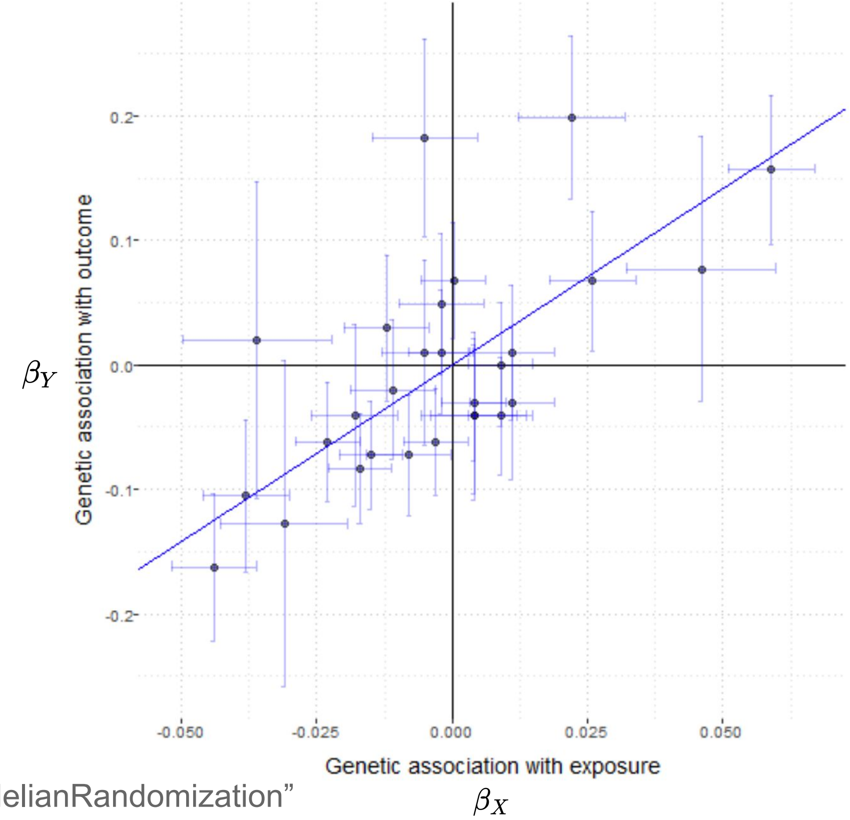
$$\hat{\theta} = \frac{\hat{\beta}_Y}{\hat{\beta}_X}$$



Inverse variance weighted (IVW)

With multiple variants, estimates can be combined using the inverse variance weighted approach

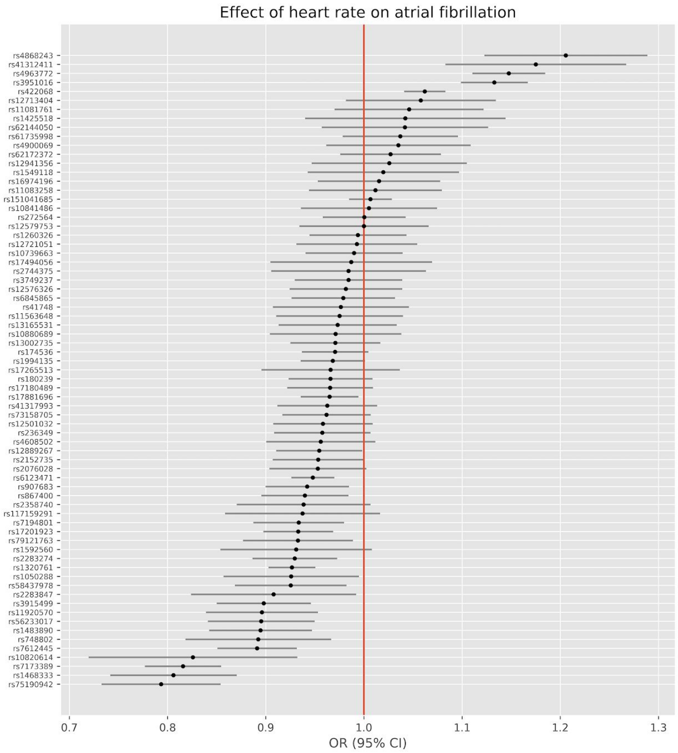
- Ratio estimates weighted by their precision (se^{-2})
- Idea from meta-analysis literature (equivalent of a fixed-effect meta-analysis of ratio estimates)
- Can be seen as a (weighted) linear regression

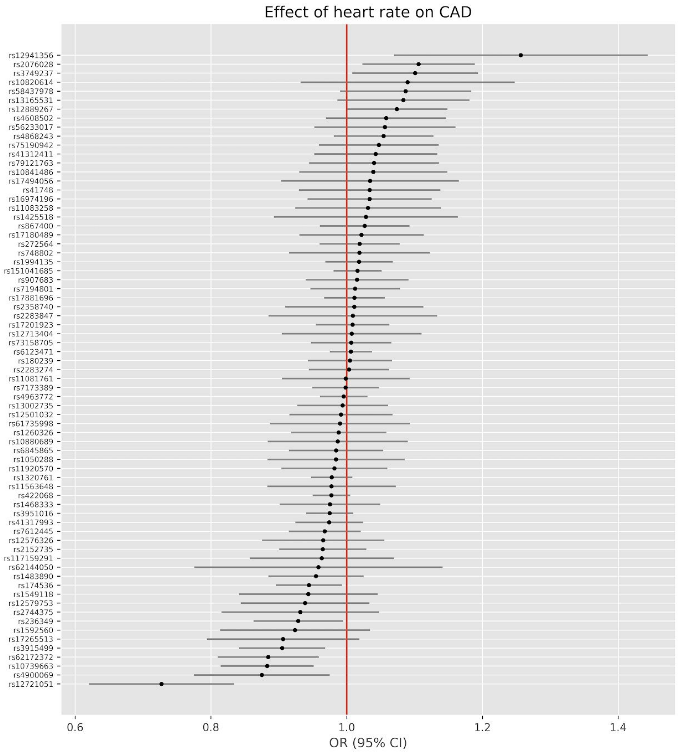


From R package “MendelianRandomization”

Real-life example of ratio estimates and the IVW method

Heterogeneity in direction of effect motivates other approaches


Effect of heart rate on atrial fibrillation


Effect of heart rate on CAD

Effect of HR increase on atrial fibrillation:

- IVW: 0.98 (0.96, 1.00); p=0.021

Effect of HR increase on CAD:

- IVW: 0.99 (0.98, 1.01); p=0.305

MR-Egger

- Builds on the analogy to meta-analysis
- Similar to the IVW method, but allows for an intercept term: directional pleiotropy
- Implication: Only useful to correct for the mean direct effects (G -> Y) across tested variants

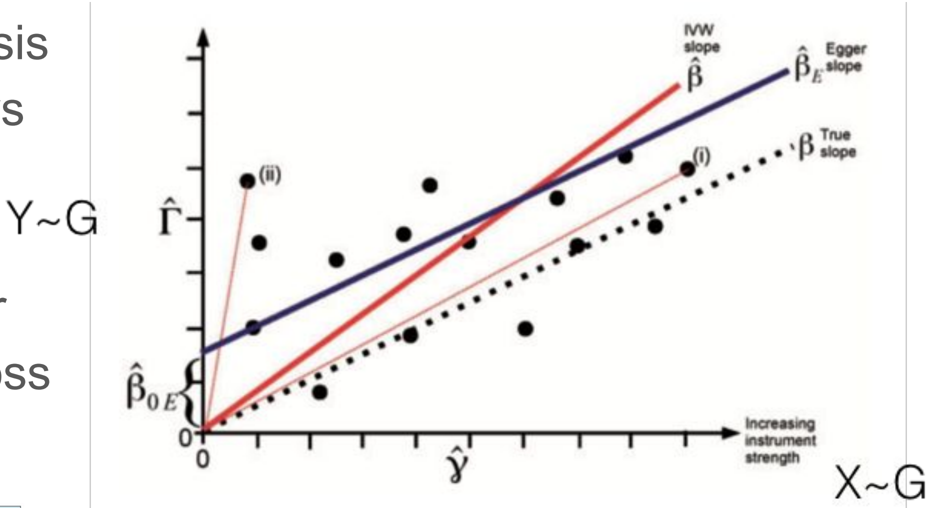
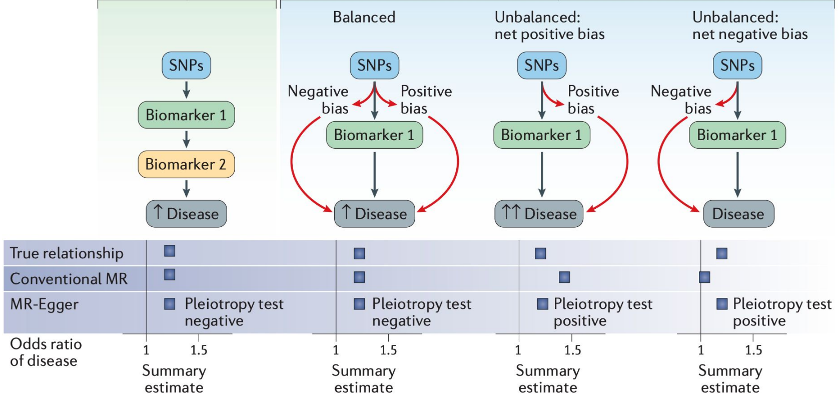


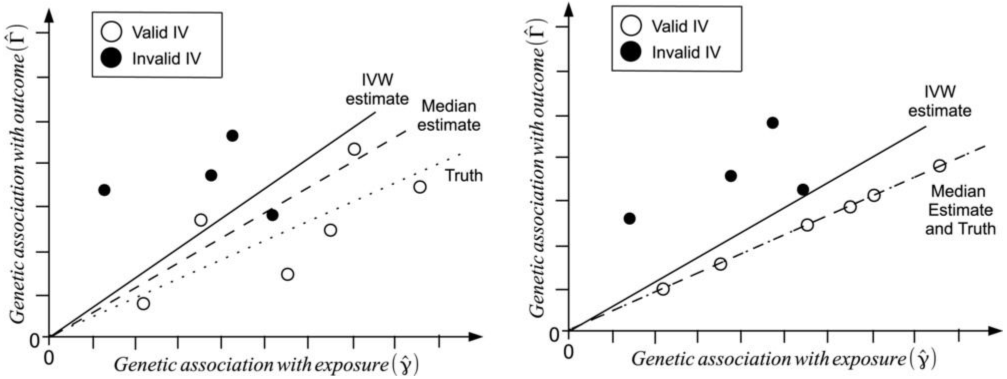
Figure 2. Plot of the gene–outcome ($\hat{\Gamma}$) vs gene–exposure ($\hat{\gamma}$) regression coefficients for a fictional Mendelian randomization analysis with 15 genetic variants. The true slope is shown by a dotted line, the inverse-variance weighted (IVW) estimate by a red line, and the MR-Egger regression estimate by a blue line. Refer to text for explanation of points (i) and (ii).



From Burgess S *et al.* (2017) Eur. J. of Epi.

Weighted median

- As for all standard (linear) regression models, IVW and MR-Egger are sensitive to outliers
  - They assume that **all** included variants are “valid IVs” under similar assumptions
- The simple median estimator is simply the median ratio estimate
  - **Significantly relaxes MR assumptions** (works if majority of IVs are valid)
- The weighted median is the same, but estimates are weighted by their precision



Bowden *et al.* (2016) Genet. Epidemiol.

Mendelian Randomization Pleiotropy RESidual Sum and Outlier

Assumption:

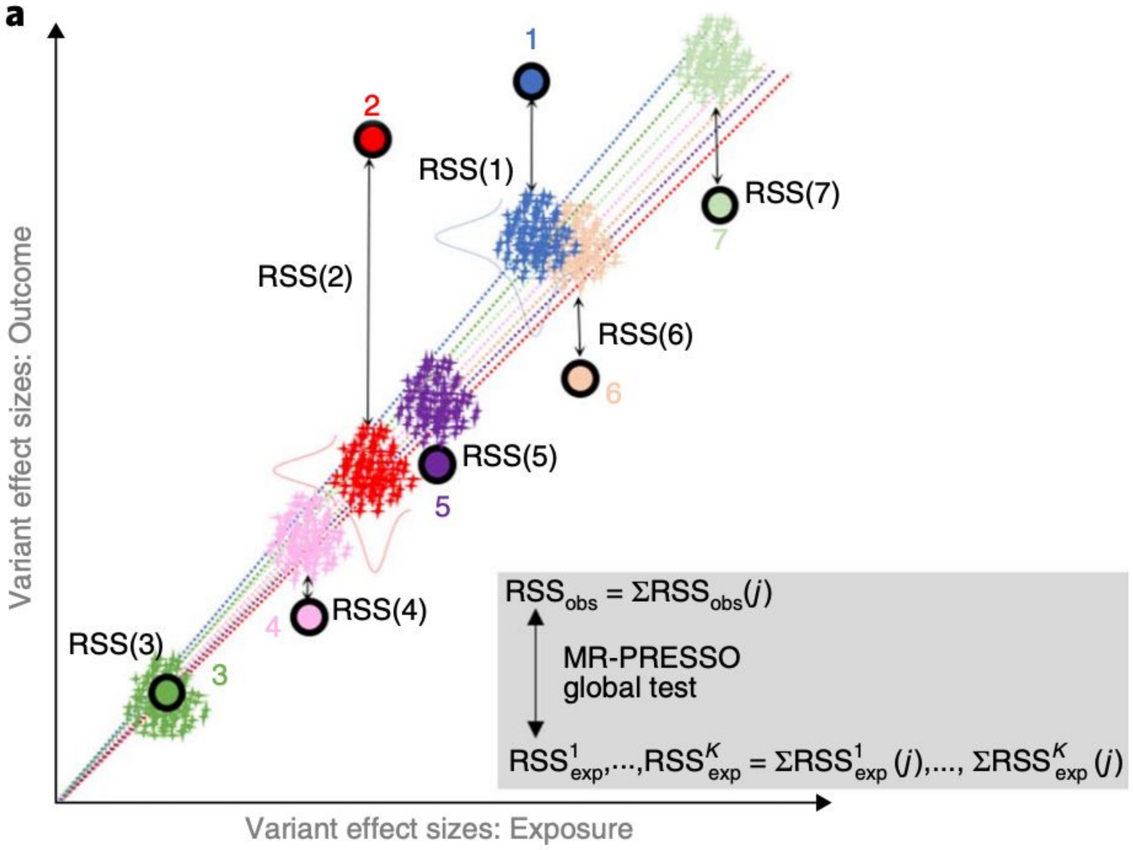
Largest set of variants with homogeneous causal effect estimates represent the true effect

- MR-PRESSO provides:
  - Global test (detection of horizontal pleiotropy)
  - Outlier test for every variant
    - And subsequent correction by removing outliers
  - Distortion test (did removing outlier significantly change the causal estimate?)

MR-PRESSO

High level overview of the method (global test):

- IVW estimate obtained by iteratively excluding variant *j*
- Compare observed RSS(j) (distance to predicted variant-outcome effect) with an expected distribution sampled from the IVW estimate


a
1
2
3
4
5
6
7
RSS(1)
RSS(2)
RSS(3)
RSS(4)
RSS(5)
RSS(6)
RSS(7)
Variant effect sizes: Outcome
Variant effect sizes: Exposure
$RSS_{obs} = \Sigma RSS_{obs}(j)$
MR-PRESSO global test
$RSS^{1}_{exp},...,RSS^{K}_{exp} = \Sigma RSS^{1}_{exp}(j),..., \Sigma RSS^{K}_{exp}(j)$

Verbanck M, *et al.* (2018) Nature Genetics

Other flavors - Multivariable MR

- Leveraging known pleiotropy to estimate causal effects
- For example: correlated exposures like blood lipids
- Same methods as before, but multivariable
  - 2SLS
  - Likelihood based method
  - Regression-based methods
    - Iteratively fits risk factors and keeps residuals

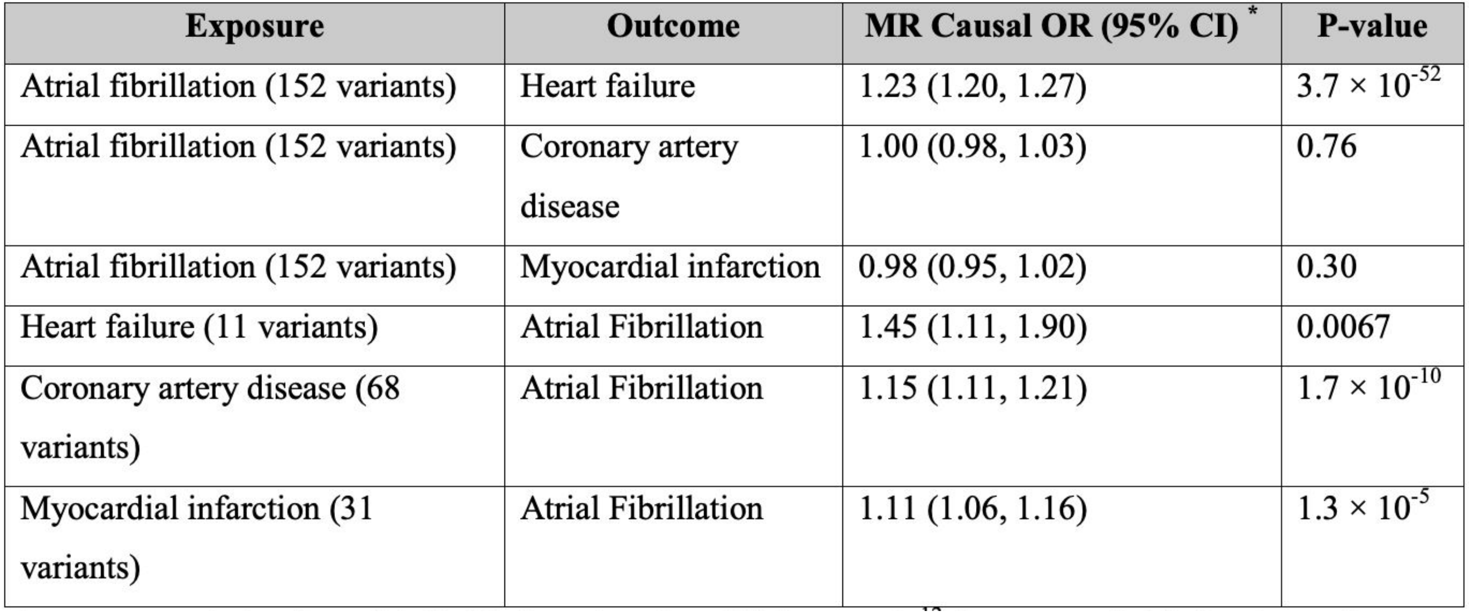
$$\begin{pmatrix} X_{1j} \\ X_{2j} \\ Y_j \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} \xi_{1j} \\ \xi_{2j} \\ \beta_1 \xi_{1j} + \beta_2 \xi_{2j} \end{pmatrix}, \begin{pmatrix} \sigma^2_{X1j} & \rho_{12}\sigma_{X1j}\sigma_{X2j} & \rho_{1Y}\sigma_{X1j}\sigma_{Yj} \\ \rho_{12}\sigma_{X1j}\sigma_{X2j} & \sigma^2_{X2j} & \rho_{2Y}\sigma_{X2j}\sigma_{Yj} \\ \rho_{1Y}\sigma_{X1j}\sigma_{Yj} & \rho_{2Y}\sigma_{X2j}\sigma_{Yj} & \sigma^2_{Yj} \end{pmatrix} \right).$$

Example model to be fit using probabilistic programming software (Bayesian or maximum likelihood).
X_{1j} is the effect of variant *j* on exposure 1
X_{2j} is the effect of variant *j* on exposure 2
etc.

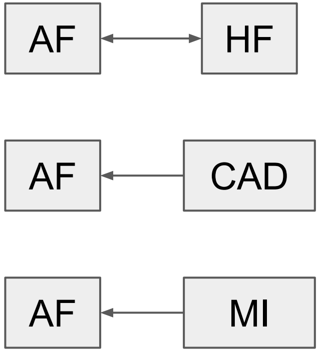
Other flavors - Bi-directional MR

Do the MR causal estimate in both directions

Example from our group (Legault MA *et al.* (2020) preprint on medRxiv)



Exposure	Outcome	MR Causal OR (95% CI) *	P-value
Atrial fibrillation (152 variants)	Heart failure	1.23 (1.20, 1.27)	3.7 × 10^{-52}
Atrial fibrillation (152 variants)	Coronary artery disease	1.00 (0.98, 1.03)	0.76
Atrial fibrillation (152 variants)	Myocardial infarction	0.98 (0.95, 1.02)	0.30
Heart failure (11 variants)	Atrial Fibrillation	1.45 (1.11, 1.90)	0.0067
Coronary artery disease (68 variants)	Atrial Fibrillation	1.15 (1.11, 1.21)	1.7 × 10^{-10}
Myocardial infarction (31 variants)	Atrial Fibrillation	1.11 (1.06, 1.16)	1.3 × 10^{-5}



Try it!

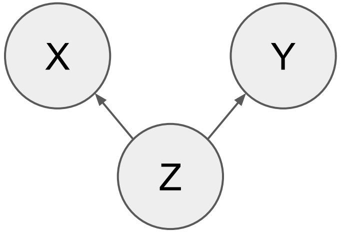
A platform for Mendelian randomisation using summary data from genome-wide association studies

http://www.mrbase.org/

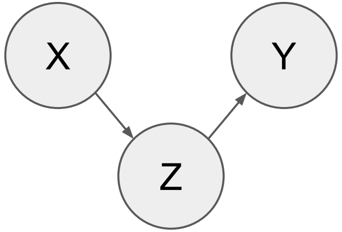
Perspectives for ML

- Opportunities for ML research in Mendelian randomization
  - Going beyond linearity (estimate the causal y = f(X))
  - Different (complex) outcomes:
    - age at onset, disease trajectory
    - ECG or imaging

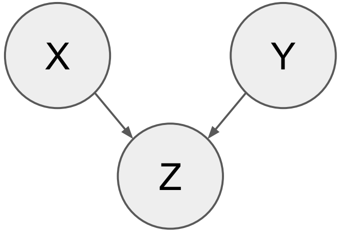
Glossary



- Fork
- Confounder
- Common cause



- Chain
- Mediator



- Collider

Extra slides

Formal regression

When I write Y ~ X I’m referring to the model:

$$Y = \beta X + \epsilon$$

With $\epsilon \sim \mathcal{N}(0, \sigma_e)$

When I write “effect” or “coefficient”, I refer to the beta term from this model or its estimate (by OLS or maximum likelihood, formally $\hat{\beta}$)

MR-Egger assumptions

- IV1: Genetic variant independent of confounders U;

$$G \perp\!\!\!\perp U$$

- IV2: Genetic variant associated with exposure X;

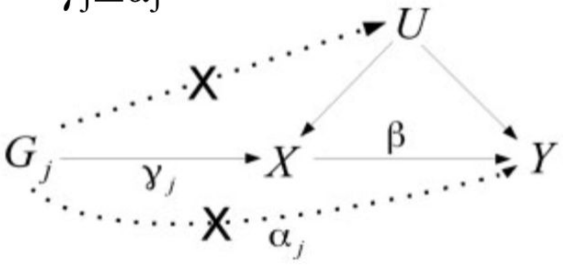
$$\gamma_j \neq 0$$

- ~~IV3~~: Genetic variant independent of the outcome Y given X and U

$$G \perp\!\!\!\perp Y \mid X, U$$

- InSIDE (Instrument Strength Independent of Direct Effects)

$$\gamma_j \perp\!\!\!\perp \alpha_j$$



Ratio estimate derivation (French)

On cherche l’effet causal de X sur Y

Effet génétique des variants sur X

Confounders

Résidus

Exposition (par ex. LDL)

$$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_i^X \qquad (1)$$

Outcome (par ex. CAD)

$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \beta X_i + U_i + \epsilon_i^Y . \qquad (2)$$

Effets génétiques directs (par ex. pléiotropie)

* Effet de l’exposition

La méthode du ratio utilise le ratio des coefficients des régressions Y~G / X~G

$$Y_i = \Gamma_j G_{ij} + \epsilon_{ij}'^Y$$

$$= (\cancel{\alpha}_j + \beta\gamma_j) G_{ij} + \epsilon_{ij}'^Y$$

IV3

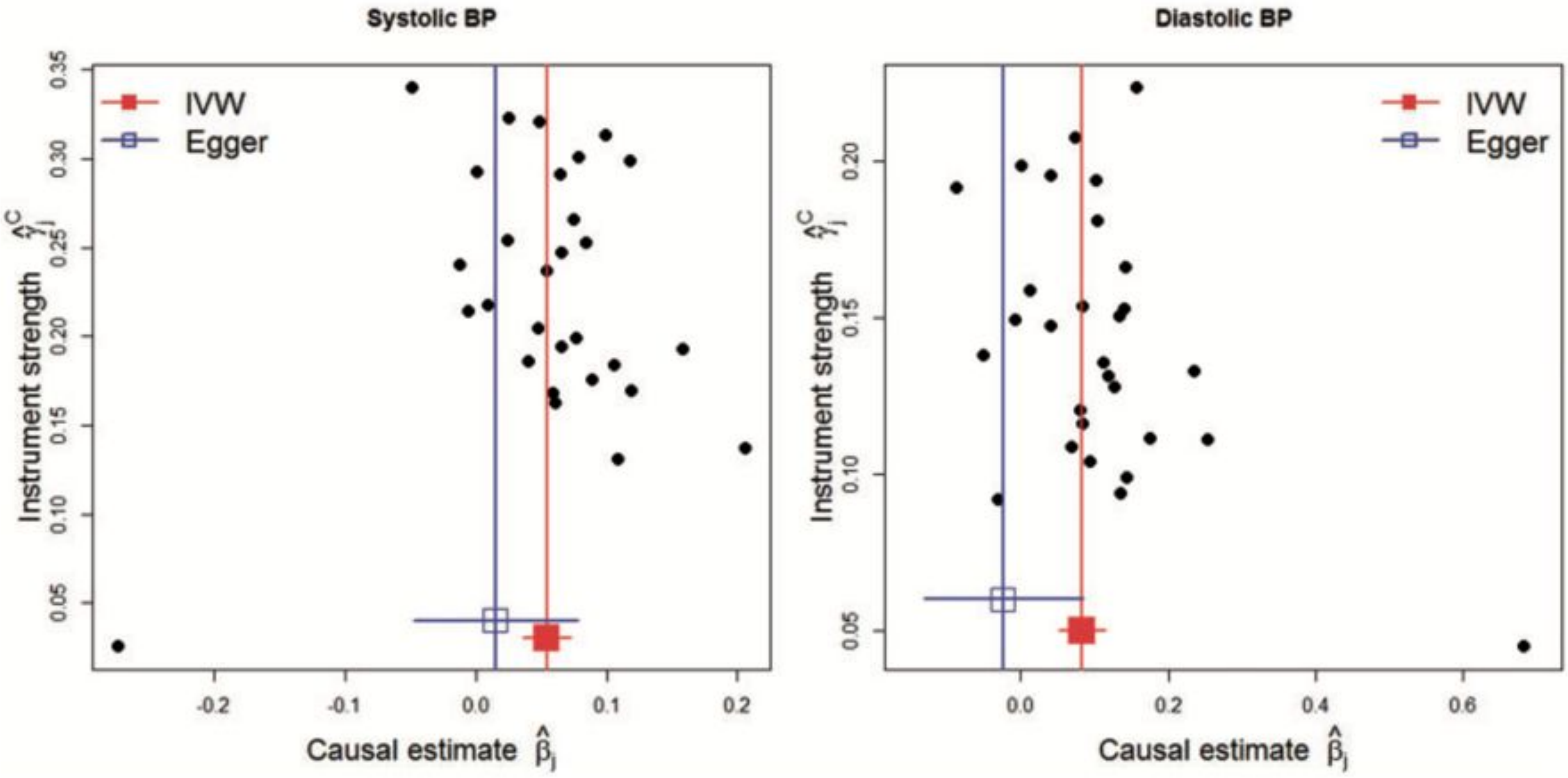
Donc, avec nos suppositions:

$$\Gamma = \beta\gamma$$

$$\frac{\Gamma}{\gamma} = \beta$$

Soit le ratio du coefficient de la régression G sur Y et de la régression G sur X

MR-Egger funnel plot


Systolic BP
IVW
Egger
Instrument strength $\hat{\gamma}_j^C$
Causal estimate $\hat{\beta}_j$
Diastolic BP
IVW
Egger
Instrument strength $\hat{\gamma}_j^C$
Causal estimate $\hat{\beta}_j$