SPRING 2023

Spring Meeting Giovani Ricercatori

SID SIIA SIMI SIPREC SISA

# GUIDA ALLA RANDOMIZZAZIONE MENDELIANA:

## *perché e come applicarla*

*Federica Galimberti*

*IRCCS MultiMedica*
*Sesto San Giovanni (MI)*

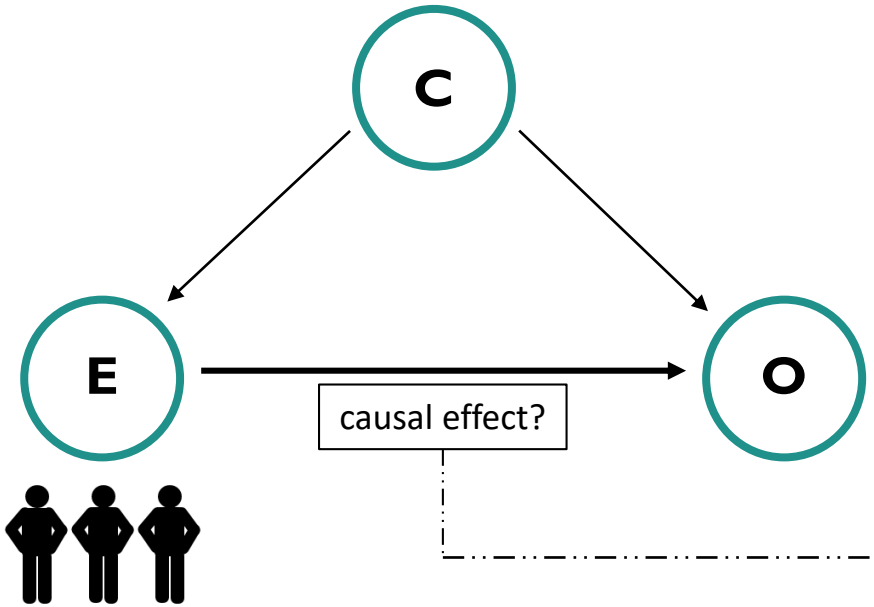
IRCCS Istituto di Ricovero e Cura a Carattere Scientifico MultiMedica

*WORKSHOP: "Come approcciarsi ai test statistici: road to Mendelian Randomization"*

# TABLE OF CONTENTS

# Medical Research

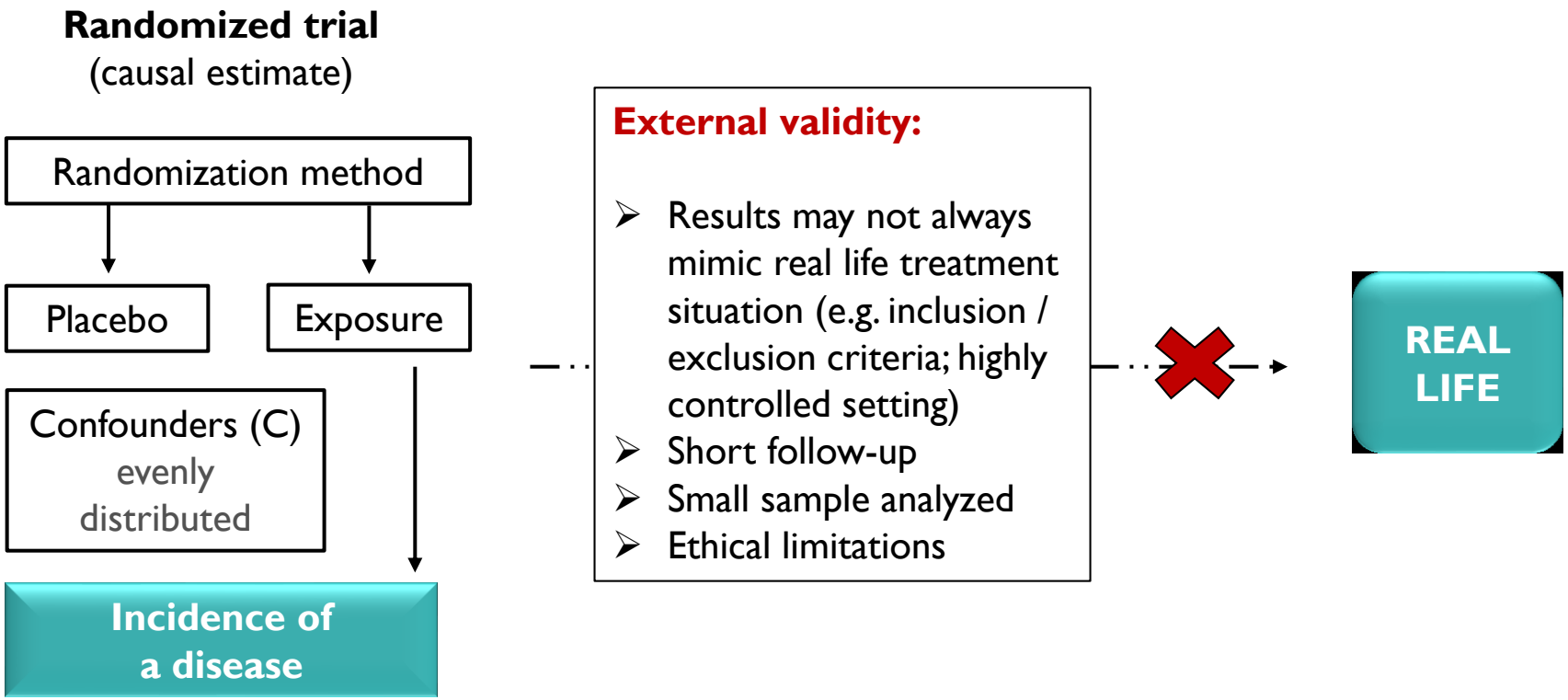
One of the major aims of medical research is to identify exposures (E), also called risk factors or intermediate phenotypes, which are causal to the manifestation of a specific outcome (O), such as disease initiation, disease progression, or response to therapy (efficacy and safety).




Once identified, causal risk factors can enable preventive measures and represent attractive therapeutic targets

# «Gold Standard»

The optimal way to answer questions of cause-effect relationship is to design randomized controlled trials (RCTs), the "gold standard" for the empirical testing of a hypothesis. Here, randomization ensures that study groups are comparable in all characteristics, except for the exposure of interest.

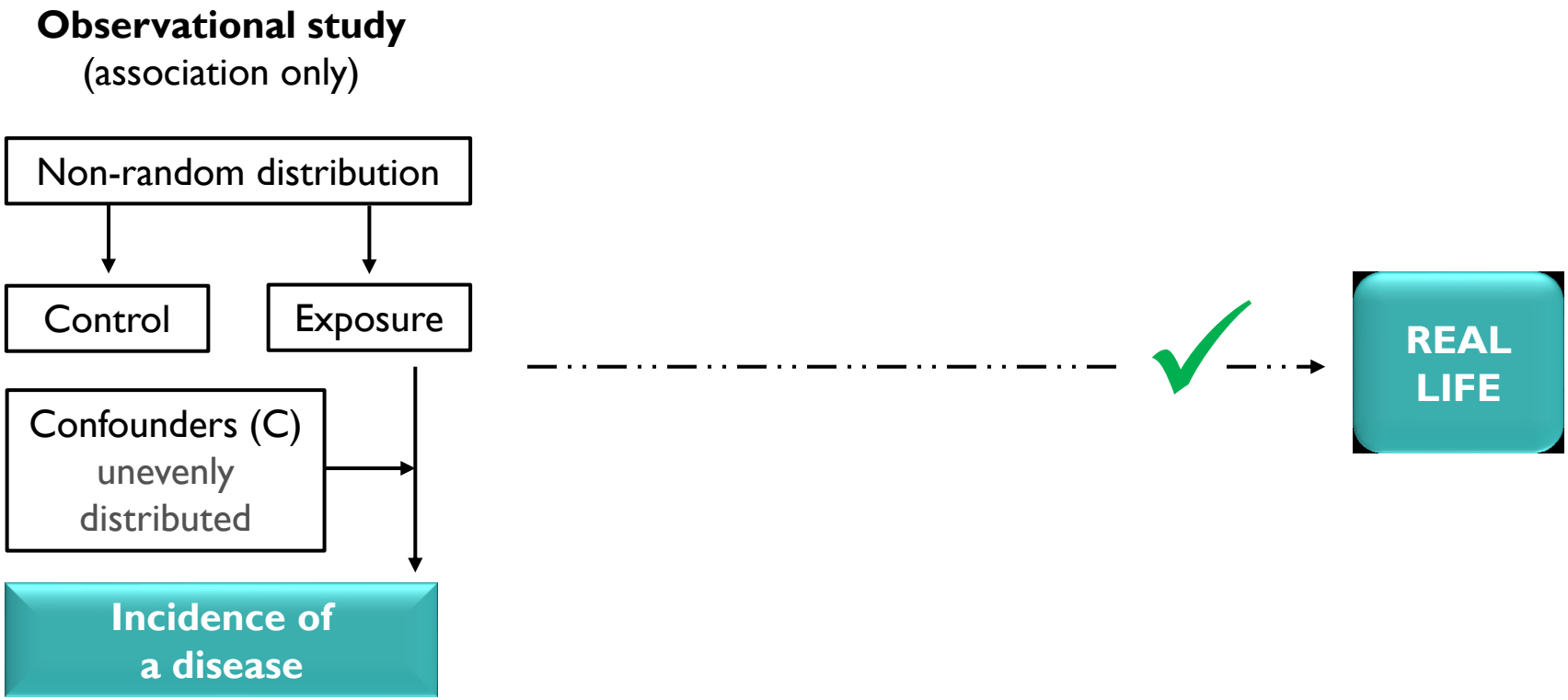


# Expectations vs Reality

*Efficacy* vs *Effectiveness*

# Observational Studies

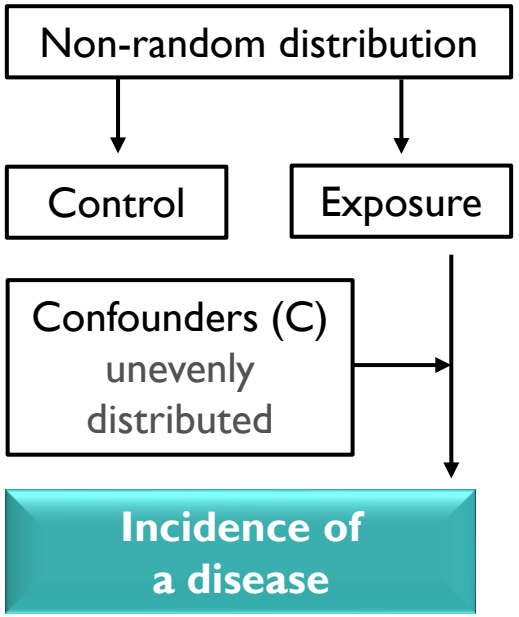
Researchers observe the effect of a risk factor, treatment or other intervention without changing who is or isn't exposed to it. Here, study groups usually differ in not only the exposure of interest but also in several observed and unobserved characteristics.

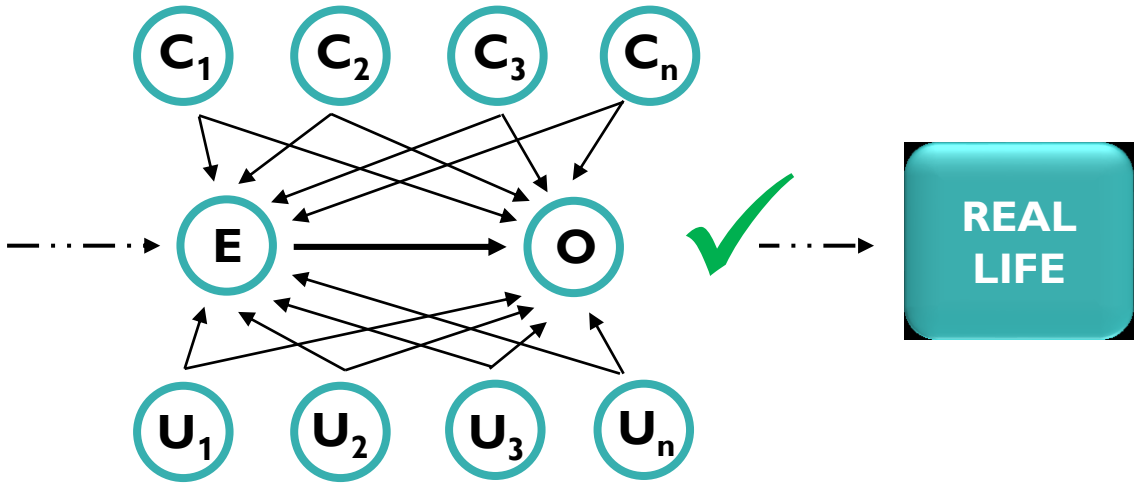


# Observational Studies

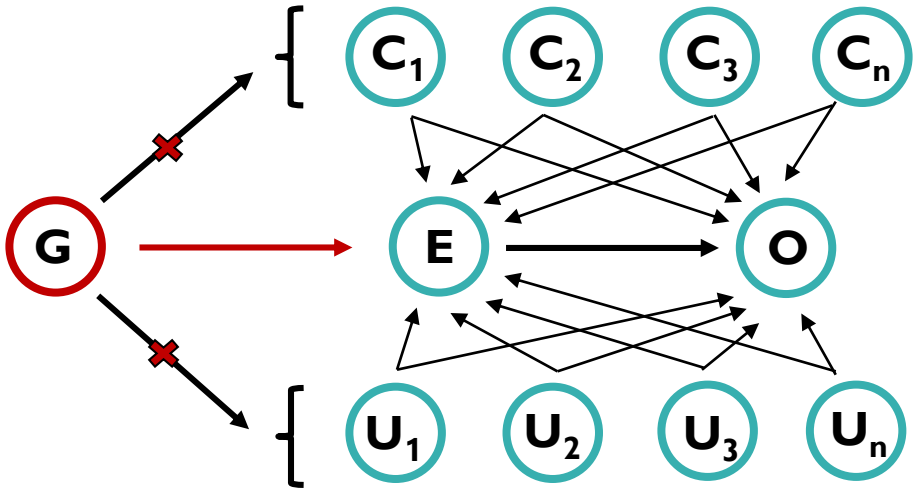
Researchers observe the effect of a risk factor, treatment or other intervention without changing who is or isn't exposed to it. Here, study groups usually differ in not only the exposure of interest but also in several observed and unobserved characteristics.

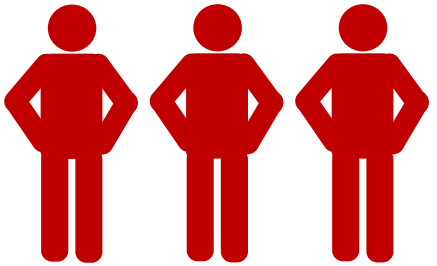
**Observational study**
(association only)

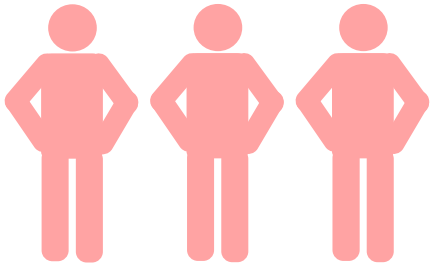


# How to recreate randomization in a real-life setting?





A variable G that either alter the level of, or imitate the biological effects of, a modifiable biomarker that is causal in disease

# Genetics

**Mendel's laws of inheritance**

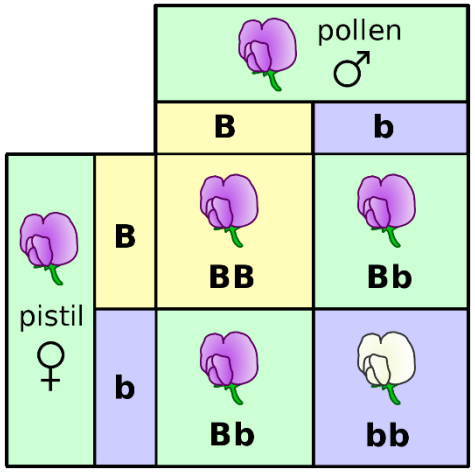
1. **Law of segregation**
   **(Randomization)**

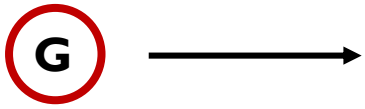
During gamete formation, the alleles for each gene segregate from each other so that each gamete carries only one allele for each gene

2. **Law of independent assortment**
   **(⊥ from confounders)**

Genes of different traits can segregate independently during the formation of gametes

Gregor Mendel
*the Father of Modern Genetics*




**Genetic Variant**

# Genetics

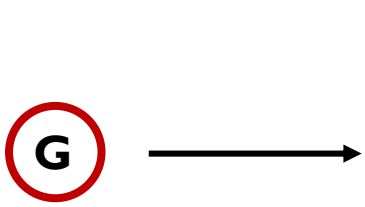
**Mendel's laws of inheritance**

1. **Law of segregation**
   **(Randomization)**

During gamete formation, the alleles for each gene segregate from each other so that each gamete carries only one allele for each gene

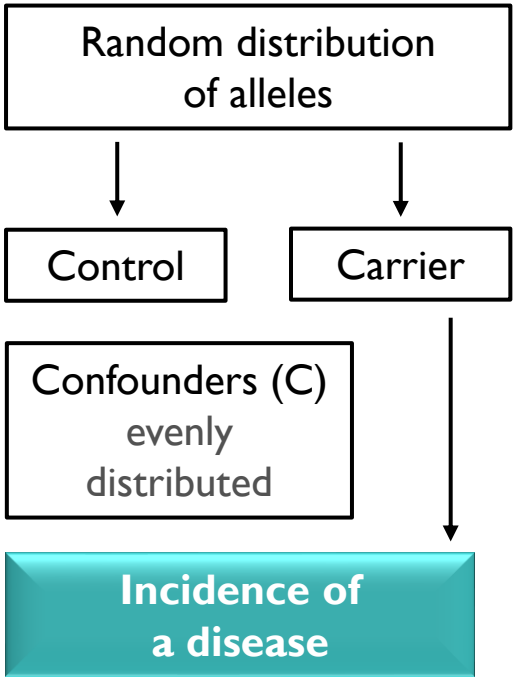
2. **Law of independent assortment**
   **(⊥ from confounders)**

Genes of different traits can segregate independently during the formation of gametes

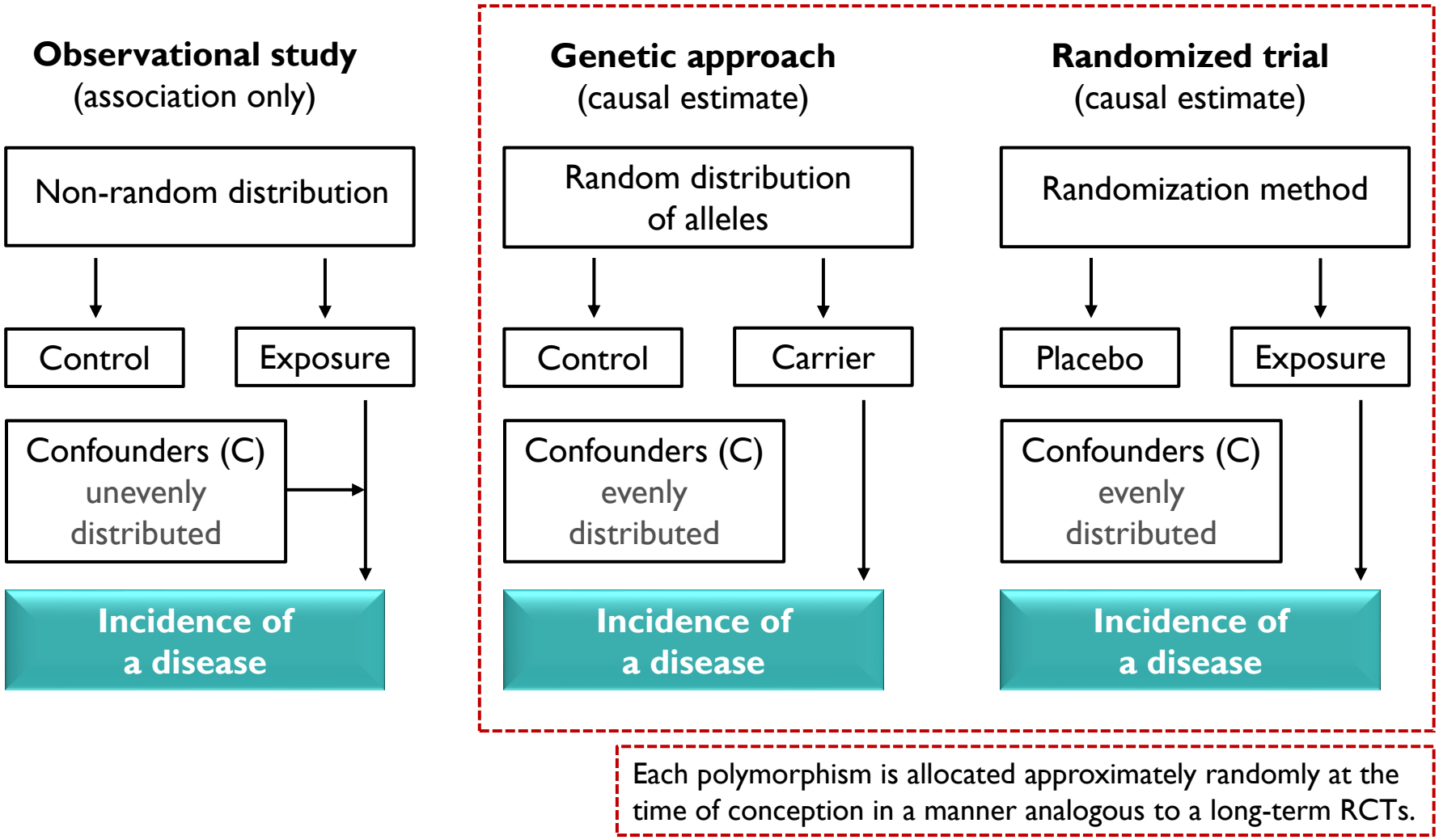



**Genetic Variant**


**Genetic approach**
(causal estimate)



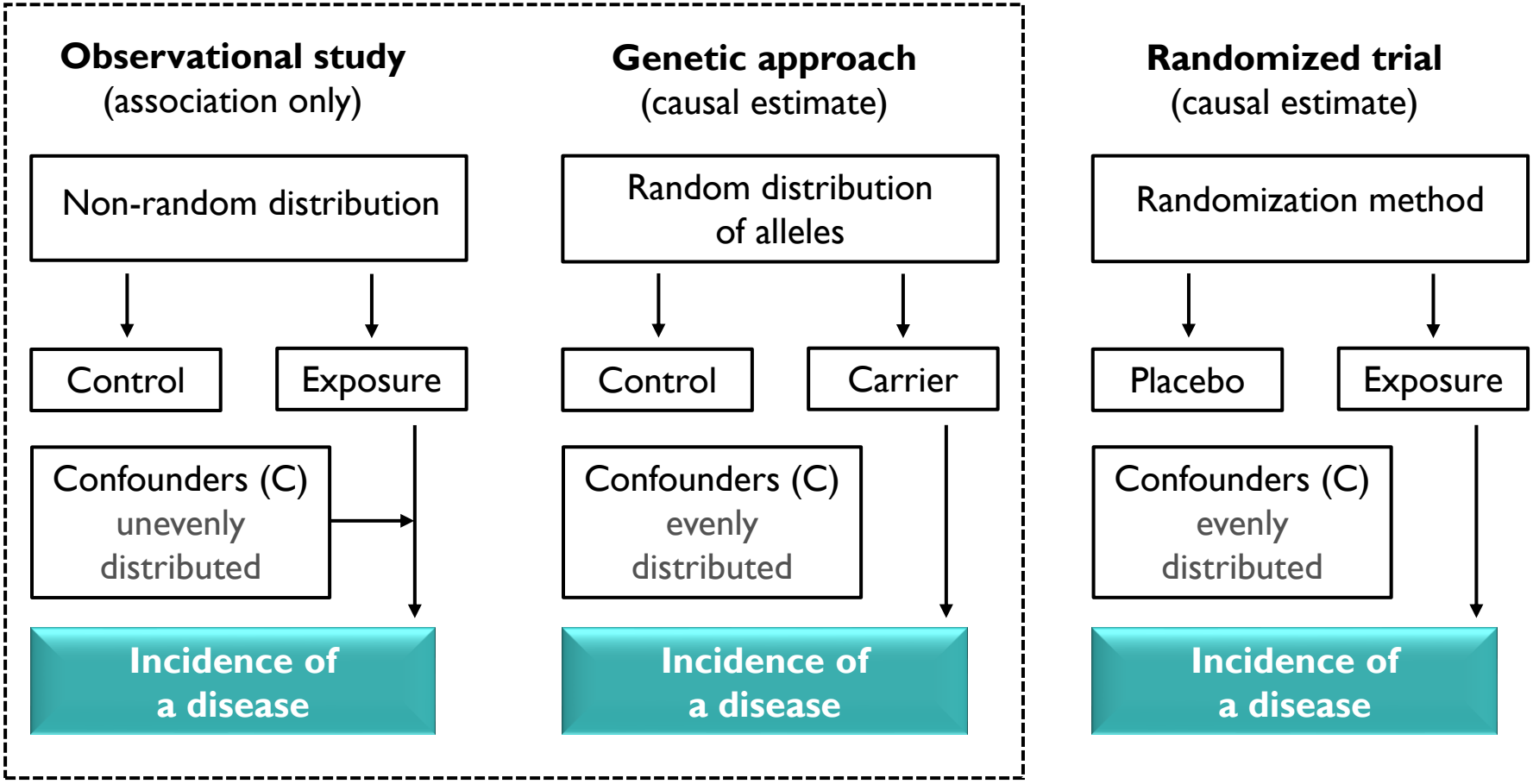
# Genetic Epidemiology




Each polymorphism is allocated approximately randomly at the time of conception in a manner analogous to a long-term RCTs.

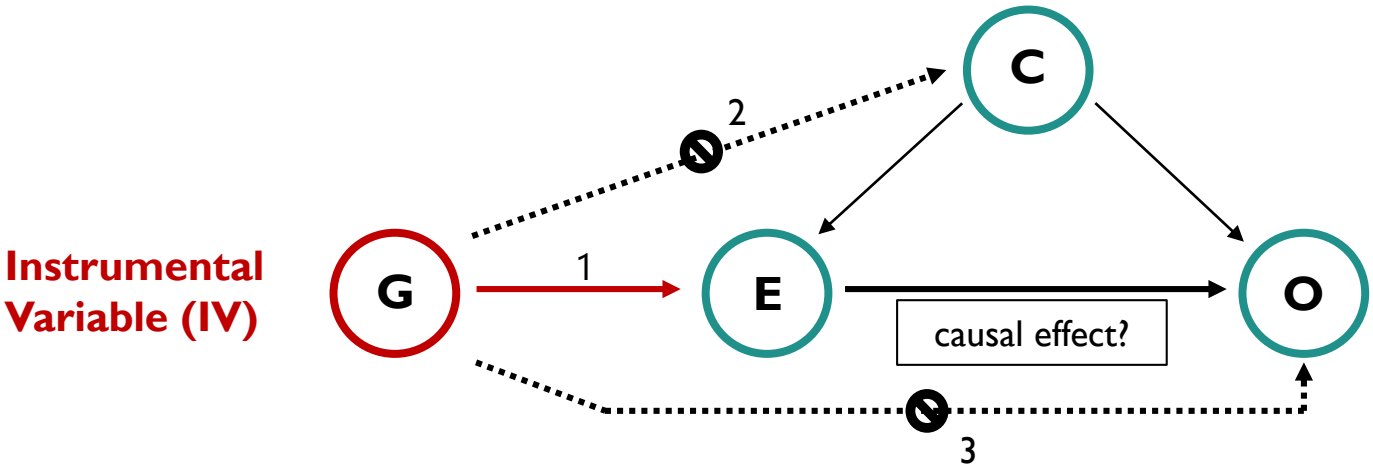
# Genetic Epidemiology

As observational studies, this design offers the opportunity to study a "real life situation"



# Mendelian Randomization approach

In the Mendelian Randomization study design genetic variants following Mendelian inheritance are used as instrumental variables. To use the Mendelian Randomization principle and instrumental variable analysis to draw conclusions on causal effects, there are three key assumptions that must be fulfilled:



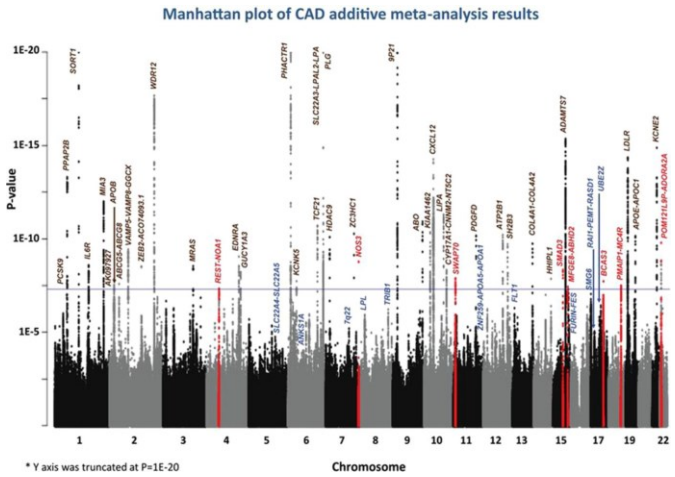

1. RELEVANCE: IV must be reproducibly and strongly associated with the exposure
2. EXCHANGEABILITY: IV must not be associated with confounders
3. EXCLUSION RESTRICTION: IV must be only associated with the outcome through the exposure

# Mendelian Randomization approach: PROs and CONs

| | PROs | | CONs |
|---|---|---|---|
| 1 | Analogous to a RCT | 1 | Lack of suitable genetic variants and potential weak instrument bias |
| 2 | Limits the presence of confounding | 2 | Pleiotropy |
| 3 | Can study exposures that are expensive or difficult to measure | 3 | Linkage disequilibrium |
| 4 | Can assess causality of risk factors for which interventions are not available | 4 | Canalization / developmental compensation |
| 5 | Removes the possibility of reverse causation | 5 | Causal estimates are often overestimated |

# Polygenic Risk Scores

**GWAS Summary Statistics**
estimate the effect size ($\beta$) of the association of variants (SNPs) with a trait of interest



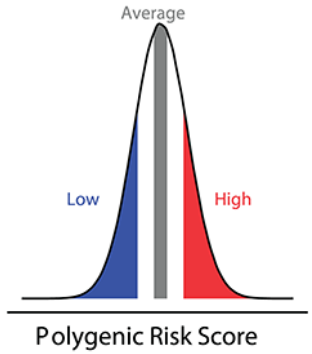

*GWAS Summary Statistics*

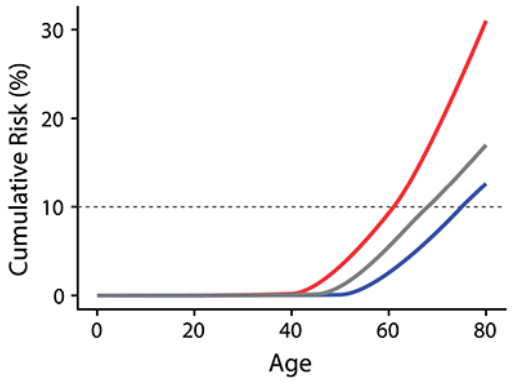
- Select SNPs (e.g. $p<5x10^{-8}$, $r^2<0.2$)
- Sum of the effects of *n* SNPs, based on the estimated SNP effect sizes ($\beta$)

$$PGS=\sum_{j=1}^{n} x_{ij}\hat{\beta}_j$$

*where* $x_{ij}$ is the genotype for the *i*th individual and *j*th SNP (usually encoded as 0, 1 or 2 for the effect allele dosage)

**Polygenic Risk Score**
A collection of variants that when combined into a score are predictive of an individual's genetic predisposition to a trait




*Lambert SA et al, Hum Mol Genet 2019*

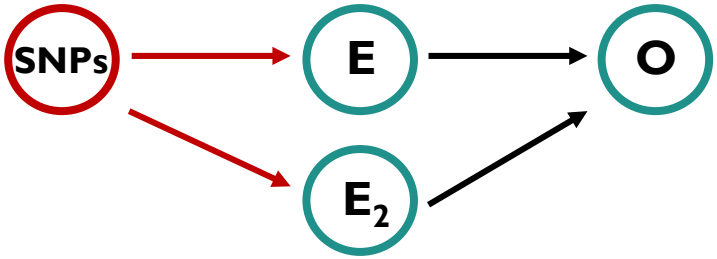
# Mendelian Randomization approach: PROs and CONs

1. Analogous to a RCT
2. Limits the presence of confounding
3. Can study exposures that are expensive or difficult to measure
4. Can assess causality of risk factors for which interventions are not available
5. Removes the possibility of reverse causation

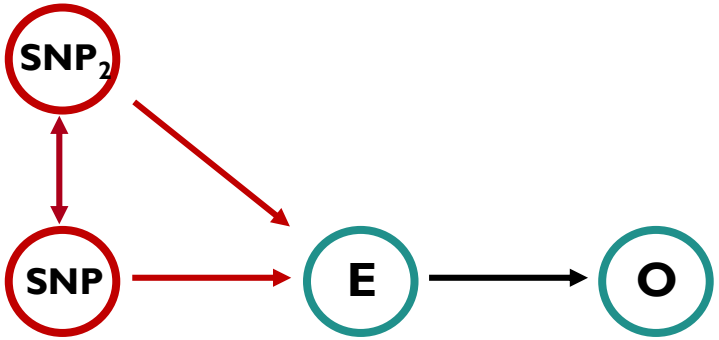
1. Lack of suitable genetic variants and potential weak instrument bias
2. Pleiotropy
3. Linkage disequilibrium (LD)
4. Canalization / developmental compensation
5. Causal estimates are often overestimated

# CONs

**PLEIOTROPY**

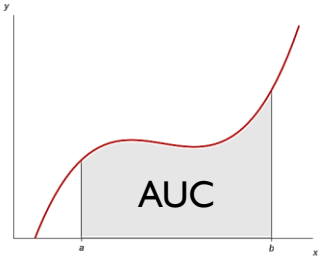



**LD**

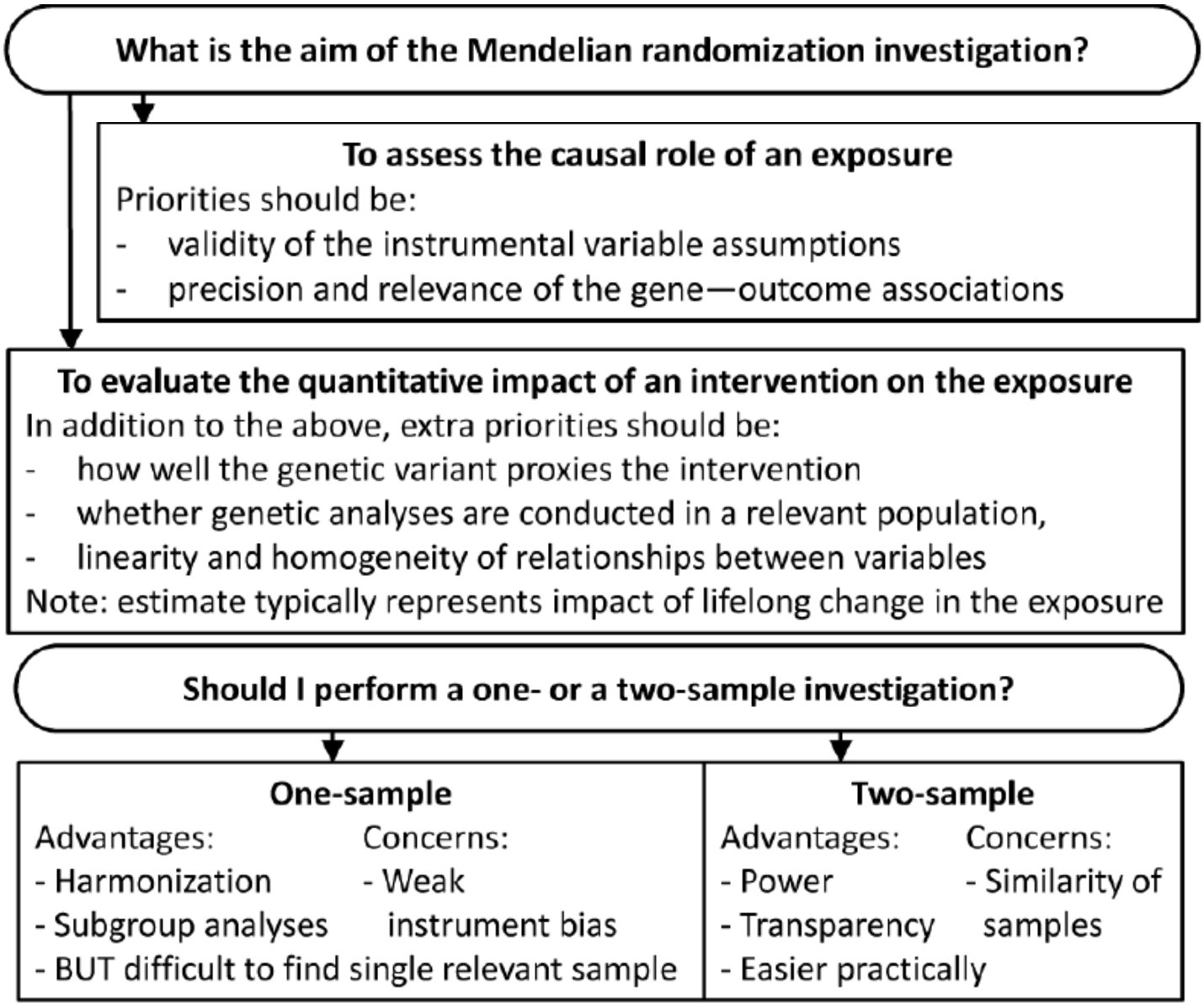



**OVERESTIMATED**

Lifelong exposure



# Key analytic choices in performing a MR analysis




*Burgess S et al, Wellcome Open Res 2020; 4:186*

# Mendelian Randomization Analysis: Data source

**ONE-SAMPLE MR**: genetic variants, exposure, and outcome are measured in the same individuals

**TWO-SAMPLE MR**: variant-exposure associations are estimated in one dataset, and variant-outcome associations are estimated in a second dataset

| IV assumption | One-sample MR | Two-sample MR |
|---|---|---|
| 1° | The partial F statistic and partial r squared, or risk difference | Variants are associated with the risk factor in a large genome-wide study |
| 2° | Covariate balance tests and bias Component plots. Adjusting for principal components of population stratification | Evidence from large genome-wide association studies on the association of the genetic variants used as instruments with other baseline covariates |
| 3° | Biological knowledge, tests of association of the genetic variants and potential alternative mediating pathways | Evidence from large genome-wide association studies that the genetic variants associate with alternative pathways. MR Egger test for pleiotropy, Cook's distance evaluation of outliers |

*Davies NM et al, BMJ 2018; 362 :k601*

# Mendelian Randomization Analysis: Data source

**INDIVIDUAL-LEVEL DATA**: genetic and phenotype (exposure and outcome) measures for each individual in the study

**SUMMARY-LEVEL DATA**: genetic association estimates from regression of the exposure or outcome on a genetic variant; several large consortia have made such estimates publicly available for hundreds of thousands of variants

| Consortium name | Description | Sample size |
|---|---|---|
| BCAC | Breast cancer | **256,123** |
| CARDIoGRAMplusC4D | Coronary artery disease and myocardial infarction | **184,305** |
| CKDGen | Chronic kidney disease | **111,666** |
| DIAGRAM | Diabetes | **159,208** |
| EAGLE | Antenatal and early life and childhood phenotypes | **47,541** |
| EGG | Early growth | **153,781** |
| GIANT | Height, BMI, and other adiposity traits | **693,529** |
| GLGC | Global lipids genetics consortium | **331,368** |
| ISGC | Stroke | **84,961** |
| MAGIC | Glucose and insulin related traits | **224,459** |
| PGC | Psychiatric genetics, alcohol and tobacco, and other related traits | **>500,000** |
| SSGAC | Educational attainment and well-being | **293,723** |

*Davies NM et al, BMJ 2018; 362 :k601*

# Key analytic choices in performing a MR analysis

**How to select genetic variants?**
**What sensitivity and supplementary analyses should I perform?**

## Biologically driven approach

**If there are genetic variants having biological relevance to the exposure...**

... then consider performing an MR analysis using these variants only.

Advantages:

- Instrumental variable assumptions more plausible
- Relevance to intervention often more clear

Concerns:

- Low power
- Results sensitive if locus is pleiotropic

Sensitivity analyses:

- Single locus: colocalization. Multiple loci: assess heterogeneity
- Consider positive and negative control outcomes

## Statistical approach

**If such variants are not available...**

... then consider performing an agnostic polygenic MR analysis.

Advantages:

- Can use robust methods

Concerns:

- Pleiotropy is likely

Sensitivity analyses:

- Assess heterogeneity: statistical test and graphically (e.g. scatter plot)
- Perform a range of robust methods making different assumptions
- Check genetic associations with variables on pleiotropic pathways
- Liberal and conservative choices of variants, leave-one-out analyses
- Conduct relevant subgroup analysis

*Burgess S et al, Wellcome Open Res 2020; 4:186*

# Key analytic choices in performing a MR analysis

## Polyunsaturated fatty acids and risk of anorexia nervosa: A Mendelian randomization study

Table 1: Summary statistics of plasma phospholipid levels of polyunsaturated fatty acids-raising genetic variants.

| | | | | | Effect size estimates for PUFAs[a] | | | | Effect size estimates for anorexia nervosa[b] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| PUFA | SNP | Chr | Effect allele | Other allele | EAF | $\beta^c$ | SE | p | $\beta$ | SE | p |
| Linoleic acid (LA, 18:2n6) | rs10740118 | 10 | G | C | 0.56 | 0.248 | 0.043 | $8.08*10^{-9}$ | 0.024 | 0.014 | 0.076 |
| | rs174547 | 11 | C | T | 0.32 | 1.474 | 0.042 | $4.98*10^{-274}$ | 0.009 | 0.014 | 0.547 |
| | rs16966952 | 16 | G | A | 0.69 | 0.351 | 0.044 | $1.23*10^{-15}$ | 0.030 | 0.015 | 0.037 |
| Arachidonic acid (AA, 20:4n6) | rs174547 | 11 | T | C | 0.68 | 1.691 | 0.025 | $3.00*10^{-971}$ | −0.009 | 0.014 | 0.547 |
| | rs16966952 | 16 | G | A | 0.69 | 0.199 | 0.031 | $2.43*10^{-10}$ | 0.030 | 0.015 | 0.037 |
| α-Linolenic acid (ALA, 18:3n3) | rs174547 | 11 | C | T | 0.33 | 0.016 | 0.001 | $3.47*10^{-64}$ | 0.009 | 0.014 | 0.547 |
| Eicosapentaenoic acid (EPA, 20:5n3) | rs3798713 | 6 | C | G | 0.43 | 0.035 | 0.005 | $1.93*10^{-12}$ | −0.014 | 0.014 | 0.291 |
| | rs174538 | 11 | G | A | 0.72 | 0.083 | 0.005 | $5.37*10^{-58}$ | 0.001 | 0.014 | 0.955 |
| Docosapentaenoic acid (DPA, 22:5n3) | rs780094 | 2 | T | C | 0.41 | 0.017 | 0.003 | $9.04*10^{-9}$ | −0.024 | 0.014 | 0.076 |
| | rs3734398 | 6 | C | T | 0.43 | 0.040 | 0.003 | $9.61*10^{-44}$ | −0.015 | 0.014 | 0.264 |
| | rs174547 | 11 | T | C | 0.67 | 0.075 | 0.003 | $3.79*10^{-154}$ | −0.009 | 0.014 | 0.547 |
| Docosahexaenoic acid (DHA, 22:6n3) | rs2236212 | 6 | G | C | 0.57 | 0.113 | 0.014 | $1.26*10^{-15}$ | 0.013 | 0.014 | 0.355 |

Chr, chromosome; EAF, effect allele frequency; PUFA, polyunsaturated fatty acid; SE, standard error; SNP, single-nucleotide polymorphism

[a] Summary statistics for PUFA from "PLoS Genet 2011;7(7):e1002193. doi: https://doi.org/10.1371/journal.pgen.1002193" and "Circ Cardiovasc Genet 2014;7(3):321–31. doi: https://doi.org/10.1161/CIRCGENETICS.113.000208".

[b] Summary statistics for anorexia nervosa (16,992 cases, 55,525 controls) from "Nat Genet 2019;51(8):1207–14. doi: https://doi.org/10.1038/s41588-019-0439-2".

[c] Expressed as % of total fatty acids.

*Nomura M et al. J Affect Disord 2023; 330:245-248*

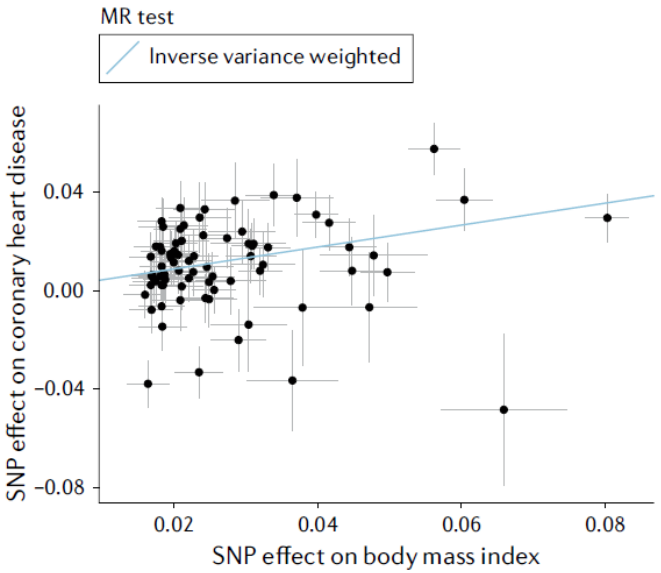
# Mendelian Randomization Analysis: Estimation Methods

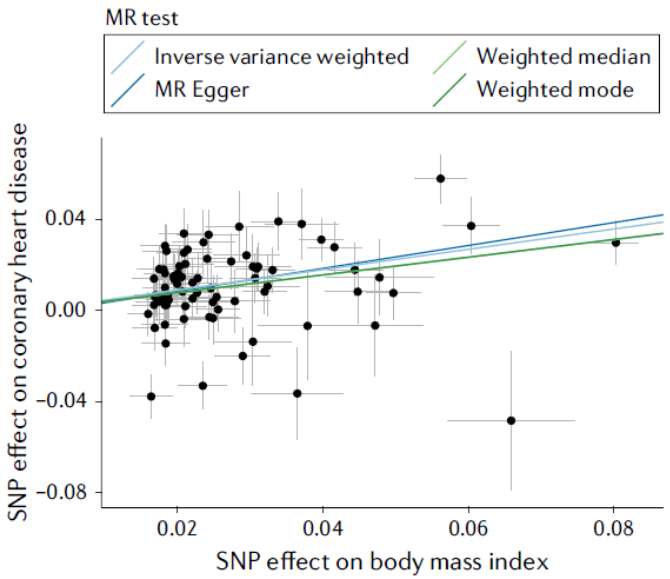
| Category | Core IV assumption relaxed | Individual-level data | Summary data |
|---|---|---|---|
| 'Basic' MR method | None | Wald ratio estimation, 2SLS regression analysis[a] | Wald ratio estimation, IVW[a,37] |
| Weak instrument robust methods | IV1; allows for weak instruments | LIML[26], allele score approaches[26] | MR RAPS[87], debiased IVW[187], MR GRAPPLE[88], NOME adjustment[188], two-sample AR[189] |
| Outlier/variant selection and removal | IV3; allows for balanced/sparse pleiotropy | Weighted median[190] | Weighted median[a,82] |
| Outlier/variant selection and removal | IV3; allows for (some) directional pleiotropy | sisVIVE[70], adaptive LASSO[71], weighted mode[190] | Weighted mode[a,83], MR LASSO[84], Steiger filtering[a,93], Welch-weighted Egger[94], contamination mixture[191], GSMR[79], MR-Clust[192], Bayesian MIMR[193], CIV[72] |
| Outlier/variant adjustment | IV3; allows for balanced pleiotropy | Limited approaches currently available | MR RAPS[87], MRCIP[194] |
| Outlier/variant adjustment | IV3; allows for (some) directional pleiotropy | Limited approaches currently available | MR TRYX[85], MR Robust[84], MR CAUSE[89], MR PRESSO[86], MR GRAPPLE[88], MRMix[195], MR-LDP[196], IMRP[197], regularization[198], MR-PATH (see preprint[199]) |
| Estimation adjustment | IV3; allows for balanced pleiotropy | Limited approaches currently available | Debiased IVW[187] |
| Estimation adjustment | IV3; allows for (some) directional pleiotropy | Constrained IVs[72], multivariable MR[73] | MR Egger[90], multivariable MR[73,91], MR Link[200], hJAM[201], GIV[202], Bayesian network analysis[203], BMRE[204], BayesMR[205] |
| Environmental control adjustment | IV3; allows for (some) directional pleiotropy | MR GxE[75,76], MR GENIUS[77] | Limited approaches currently available |

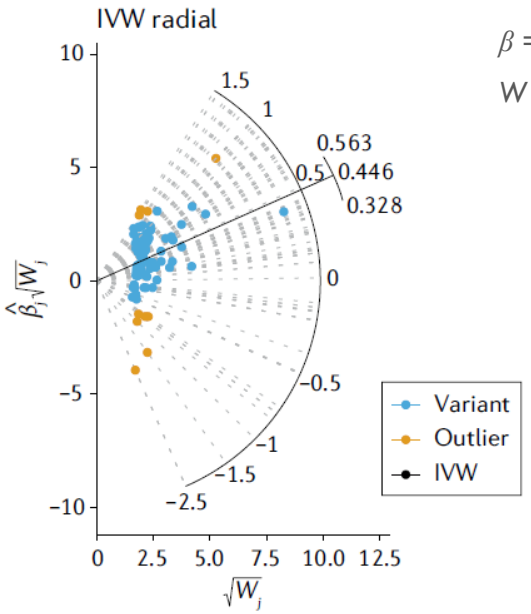
2SLS two-stage least-squares; IVW inverse variance weighted

*Sanderson E et al, Nat Rev Methods Primers 2022; 2:6*

# Mendelian Randomization Analysis: Estimation Methods

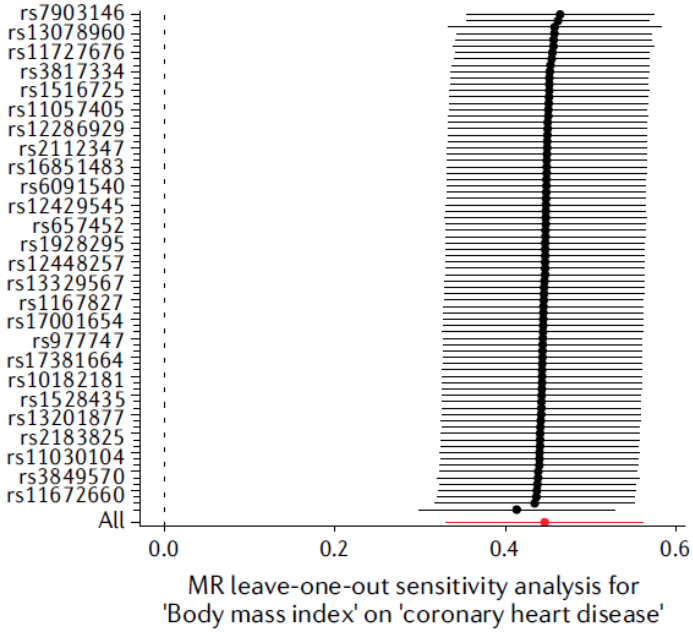



$\beta$ = *effect estimate*

$W$ = *weight given to each point*



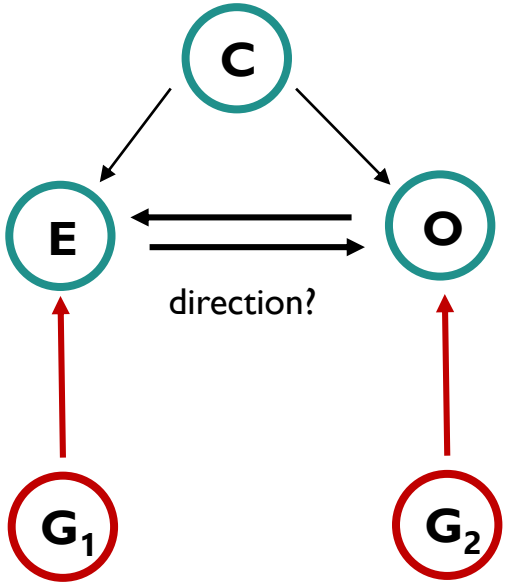
# Mendelian Randomization Analysis: Further extensions

**BIDIRECTIONAL MR**

**MULTIVARIABLE MR**

**MR MEDIATION ANALYSIS**

**NON-LINEAR MR**

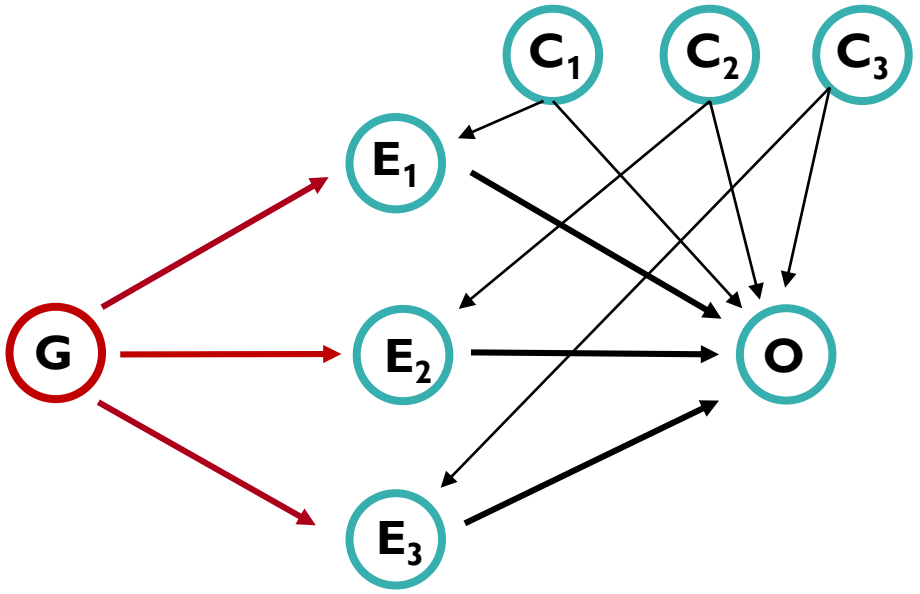



*Sanderson E et al, Nat Rev Methods Primers 2022; 2:6*

# Mendelian Randomization Analysis: Further extensions

**BIDIRECTIONAL MR**

**MULTIVARIABLE MR**




**MR MEDIATION ANALYSIS**

**NON-LINEAR MR**

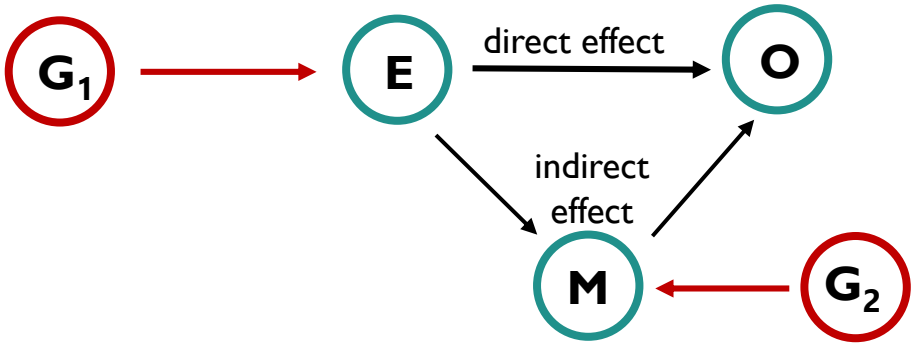
*Sanderson E et al, Nat Rev Methods Primers 2022; 2:6*

# Mendelian Randomization Analysis: Further extensions

**BIDIRECTIONAL MR**

**MULTIVARIABLE MR**

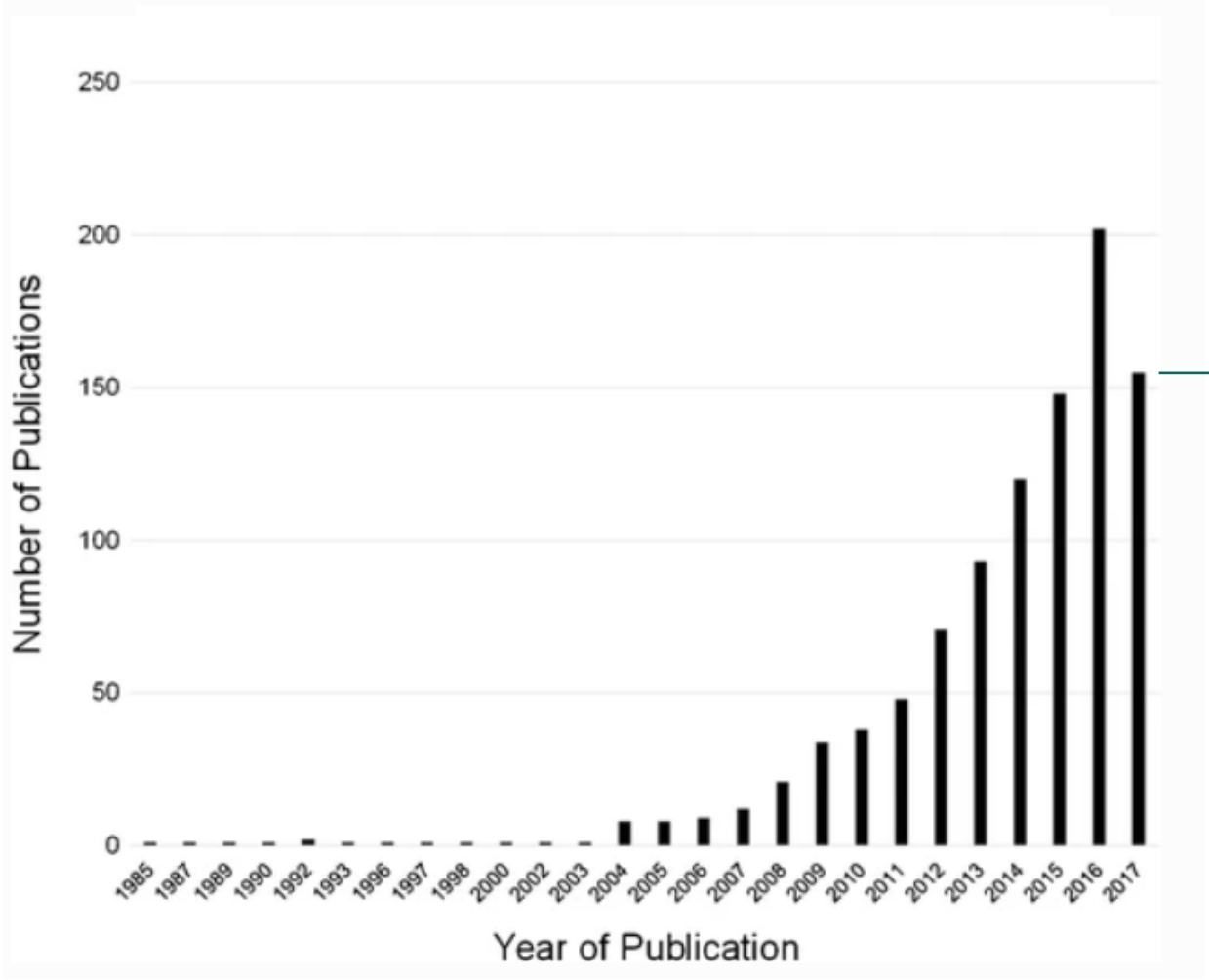
**MR MEDIATION ANALYSIS**




**NON-LINEAR MR**

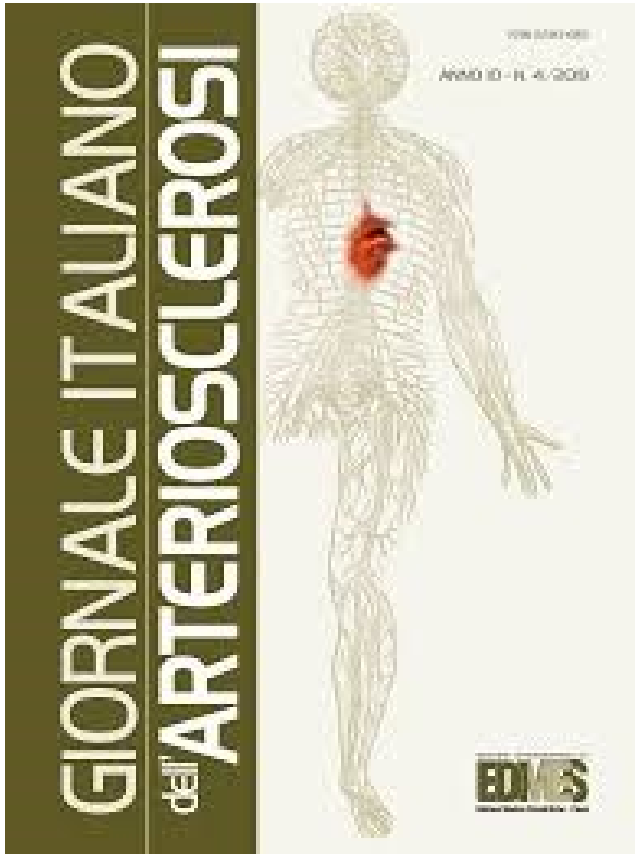
*Sanderson E et al, Nat Rev Methods Primers 2022; 2:6*

# Mendelian Randomization Studies in PubMed




*Savla, J, Neeland, IJ, Curr Cardiovasc Risk Rep 2018; 12:2*

# For more details...




FOCUS

## COS'È UNO STUDIO DI RANDOMIZZAZIONE MENDELIANA E QUALI SONO LE APPLICAZIONI IN AMBITO DI DISLIPIDEMIE

## What is a Mendelian randomization study and what are the applications in the field of dyslipidemias

**FEDERICA GALIMBERTI[1,2], ELENA OLMASTRONI[1,2]**

*[1]Epidemiology and Preventive Pharmacology Service (SEFAP), Department of Pharmacological and Biomolecular Sciences, University of Milan, Milan, Italy;*
*[2]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Worts Causeway, CB1 8RN Cambridge, United Kingdom*

*ANNO II N. 2/2020*

SPRING 2023

Spring Meeting Giovani Ricercatori

SID SIIA SIMI SIPREC SISA


*Federica Galimberti*

*IRCCS MultiMedica*
*Sesto San Giovanni (MI)*

*federica.galimberti@multimedica.it*

IRCCS Istituto di Ricovero e Cura a Carattere Scientifico MultiMedica

*WORKSHOP: "Come approcciarsi ai test statistici: road to Mendelian Randomization"*

# Lavoro a gruppi

**IDENTIFICARE GLI ELEMENTI ESSENZIALI DI UNO STUDIO DI RANDOMIZZAZIONE MENDELIANA:**

- ✓ **Hypothesis to be tested** → Exposure(s) and Outcome(s)
- ✓ **Instrumental Variable(s)** → Single or multiple genetic variants
- ✓ **Individual- or Summary-level data**
- ✓ **One- or Two-sample MR**
- ✓ **Estimation Method(s)**