Mendelian randomization: genetic anchors for causal inference in epidemiological studies

George Davey Smith and Gibran Hemani

Presentation by Mykhaylo M. Malakhov

Introduction

Even the most well-designed observational studies are subject to a number of limitations, such as confounding and reverse causation.
Many observational studies have failed to deliver anticipated health benefits when subjected to randomized controlled trials (RCTs).

Mendelian randomization

"use genetic variants as exposure indicators that are not subject to the influences that vitiate conventional study designs"

Overview

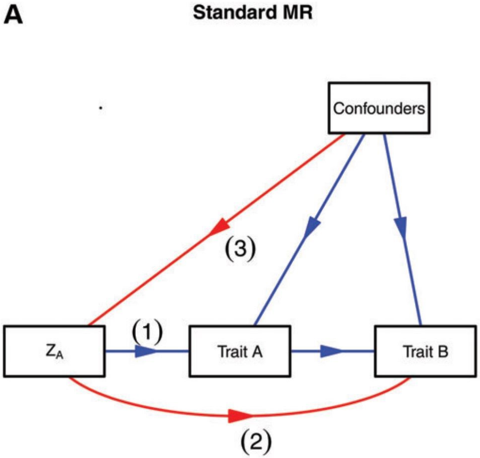
- Basic principles
- Limitations
- Recent developments:
    - Two-sample MR
    - Bidirectional MR
    - Network MR
    - Two-step MR
    - Factorial MR
    - Multiphenotype MR

Genetic variants as instrumental variables

Big idea: If trait A is causing trait B, then any variable that influences trait A should also influence trait B. Hence, we need to find an instrument that is reliably associated with A in a known direction.
Why are genetic variants such great instruments?

1. The direction of causation is clear – always genetic polymorphism to trait
2. Genetic variants are not confounded by social, physiological, and other factors
3. Genetic variants are subject to little measurement error and bias
4. The causal variant is not required, so a proxy marker is sufficient
5. GWAS data are readily available

Schematic representation of MR



IV analysis assumptions and process

An instrument for trait A must be:

1. reliably associated with trait A
2. associated with the outcome (trait B) only through trait A
3. independent of unobserved confounders that influence traits A and B after conditioning on observed confounders

Two-stage least squares (2SLS) regression:

1. A predictor for A is constructed from its instrument
2. The effect of the predictor for A on the outcome B is estimated

Consequences of pleiotropy

Type II pleiotropy: A single process leading to a cascade of events. Necessary for MR. The more common form of pleiotropy.

Type I pleiotropy: A single locus directly influencing multiple phenotypes. Problematic for MR. Potential solutions include using *cis*-variants with respect to the intermediate phenotype under study and applying MR approaches that incorporate more than one independent genetic variant.

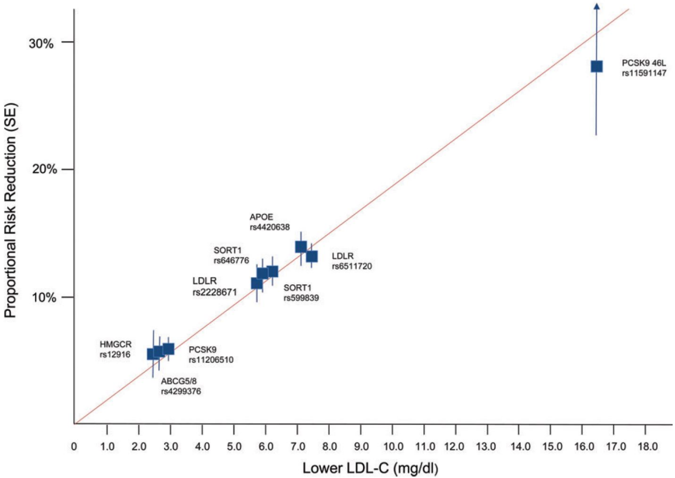
Limitations of MR

Limitation	Potential solutions
Low statistical power	Increase sample size and/or combine genetic variants
Reverse causation	Bi-directional MR
Population stratification	Restrict analyses to ethnically homogeneous groups.
	Perform analysis within a family study context
LD induced confounding	Conduct studies in populations with different LD structures
Canalization	No general approach exists
Complexity of associations	Increased biological understanding of genotype-phenotype links

Use of multiple variants

1. Increasing the number of variants increases the proportion of variance explained by the instrument. Generally the optimal approach is to combine the variants into a weighted allele score.
2. One variant might be pleiotropic or in LD with a variant that affects the outcome. For several independent instruments to result in the same conclusion due to perfectly balanced pleiotropic effects is very unlikely
3. Considering multiple variants can help resolve ambiguities regarding complexity of associations. The particular way through which one variant relates to the intermediate phenotype is unlikely to influence the cumulative evidence

Multiple variants supporting a conclusion



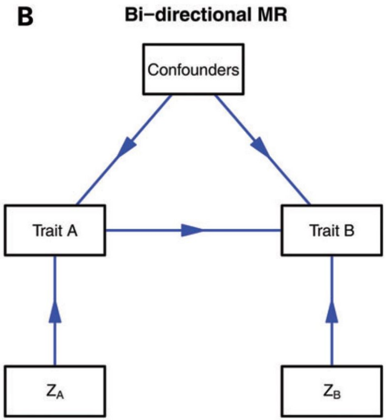
Two-sample MR

“Basic” MR uses a single data set with measurements for the intermediate phenotype and the outcome variable. Two-sample methods have been developed, which are useful in two contexts:

1. Often data sets that relate GWAS data to disease outcomes lack intermediate phenotype data. Two-sample MR allows using independent data sets
2. If MR is performed on the same data set from which instruments are extracted, the MR results can be biased due to the “winner’s curse.” Dividing the data set into separate samples for estimation and testing can mitigate this problem

Bidirectional MR

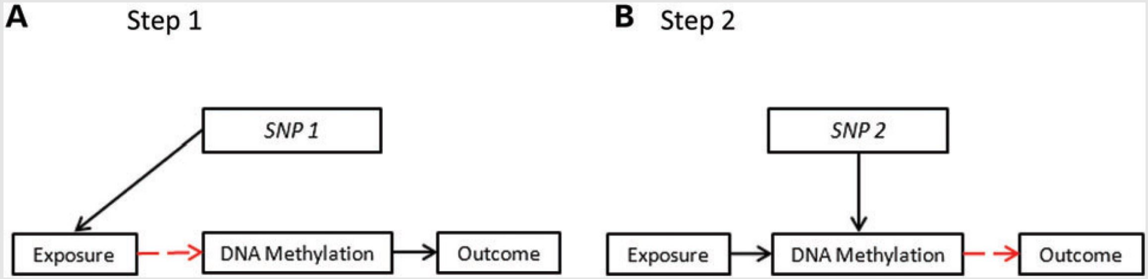
- If A causes B, then z_A will be associated with A, B but z_B will only be associated with B
- z_A and z_B must not be marginally associated
- Caution: *a priori* knowledge of variants' primary influence required
- Can be scaled up to a network of more than two traits



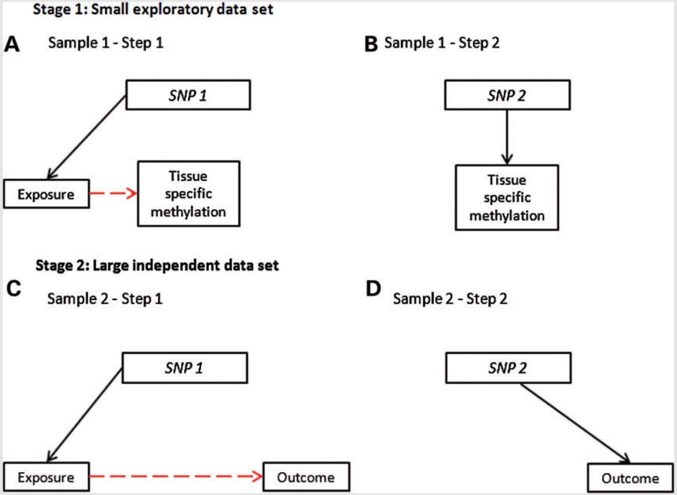
Two-step MR

Is DNA Methylation on the causal pathway between exposure and disease?

Note: SNP 2 must not be related to the exposure



Two-step, two-sample MR


Stage 1: Small exploratory data set
A
Sample 1 - Step 1
SNP 1
Exposure
Tissue specific methylation
B
Sample 1 - Step 2
SNP 2
Tissue specific methylation
Stage 2: Large independent data set
C
Sample 2 - Step 1
SNP 1
Exposure
Outcome
D
Sample 2 - Step 2
SNP 2
Outcome

Factorial MR

- Causes of disease often exhibit above-additive effects
- Example: the combined influence of obesity and heavy alcohol consumption on the risk of liver disease is greater than multiplicative
- **Confounding can be magnified when examining already confounded risk factors**
- Factorial RCTs randomize each treatment independently
- Analogous results can be obtained via MR, using different SNPs for each phenotype

Multiphenotype MR

- Sometimes genetic variants tend to be associated with multiple intermediate phenotypes
- Example: HDL cholesterol and triglycerides are associated with coronary heart disease, but they are also highly inversely correlated with each other
- Factorial MR is not possible since we cannot construct an instrument that purely relates to one phenotype
- Use of regression methods to attempt to separate the effects has shown some results

Hypothesis-free MR

- Blair *et al.* (2013) mined the medical records of 110 million patients, uncovering 2909 associations between Mendelian diseases and complex traits. The majority of these were previously unreported
- Larger GWAS data sets and improved computational capacity now allow constructing instruments for many exposures and mining the available data to obtain evidence regarding their causal effects on various outcomes
- Data mining through split-sample MR can also reveal causal direction within networks of phenotypes
- Other possibilities without advance specification of which exposure or outcome is being examined

Discussion

This article was published in 2014. What promising new advances or possibilities in MR have been proposed since?

The proliferation of machine learning has also illuminated its limitations, e.g. we cannot unscrupulously mine any data set. What are the challenges of large-scale data mining with MR?