

2. Introduction and Background

Contents

Introduction and Background	1
Causal Effect Estimation in MR	2
Weighted Median Estimator	3

Word count: 1231

Introduction and Background

α

Epidemiology is the study of determinants and distribution of disease across populations; a common epidemiological study aim is therefore to seek evidence as to whether a given exposure (e.g. cigarette smoking) may cause a given outcome (e.g. lung cancer)[?]. Logistics limit feasibility of delivering experimental interventions across large groups of people, so insights regarding associations between exposures and outcomes at scale are typically gleaned from observational data of people in the population of interest. Comparing rates of a particular health outcome between individuals with different levels of a particular exposure may highlight potential links, e.g. higher cancer incidence in those who smoke more would be consistent with a potential causal role for cigarettes in carcinogenesis[?].

However, correlation does not by itself prove causation. A key challenge in epidemiology is accounting for so-called “confounding” factors; these are other variables associated with both the exposure and the outcome of interest which represent an alternative causal explanation for any exposure:outcome links observed[?]. If those who smoke also drink more alcohol than non-smokers, then an observed link between smoking and increased cancer risk could plausibly be caused by increased alcohol exposure, either in part or entirely. Another potential issue with observational data is “reverse causation”, where the presumed outcome is in fact a cause of the exposure; this might be the case if a cancer diagnosis drove individuals to drink and smoke more, and data were collected without accounting for relative timings of each of these factors.

Mendelian randomisation (MR) is a methodology intended to support causal inference from observational data. It applies the principles of instrumental variable (IV) analysis to genetic data, in essence performing a type of natural experiment often likened to a randomised-controlled trial (RCT)[?].

In a properly conducted RCT, causality can be inferred due to a randomisation process being used as an “instrument” to allocate different levels of exposures to different experimental groups. If groups are randomly allocated, any confounding variables which might otherwise influence exposure:outcome relationships should be evenly distributed between groups, whether these confounders are known or not. As such, there should be no systematic differences between individuals from different groups in the exposure of interest - that is, there should be no bias[?]. Statistical methods can quantify the probability that any observed outcome differences could have occurred by chance, and thereafter any outcome differences can be interpreted as caused by exposure differences. As allocation and receipt of exposures is known to precede outcome measurements, reverse causality is impossible.

In MR, naturally occurring genetic variants - “genetic instruments” – are chosen based on their known association to an exposure of interest. In theory, provided that the assumptions of IV analysis (detailed

below) are met, random assignment of genetic variants from parents to offspring during meiosis can create a form of natural randomisation analogous to that performed for an RCT – both measured and unmeasured confounders should be distributed evenly between the groups created, allowing valid causal inference after other sources of bias and random variation are accounted for¹.

Causal Effect Estimation in MR

At its simplest, the relationship between an exposure X and outcome Y can be represented as a linear model:

$$Y = \beta X + \epsilon \quad (1)$$

where β is the causal effect of X on Y and ϵ is an error term. To estimate this causal effect using a genetic variant in an IV analysis, three key assumptions must be met²:

1. Relevance – the genetic variant must be associated with the exposure of interest
2. Independence – the genetic variant is independent of confounders of the relationship between exposure and outcome
3. Exclusion restriction – the genetic variant must not be associated with the outcome except via the exposure

These assumptions are represented graphically in Figure 1.

Key Assumptions in Mendelian Randomisation

Adapted from Burgess et al 2016

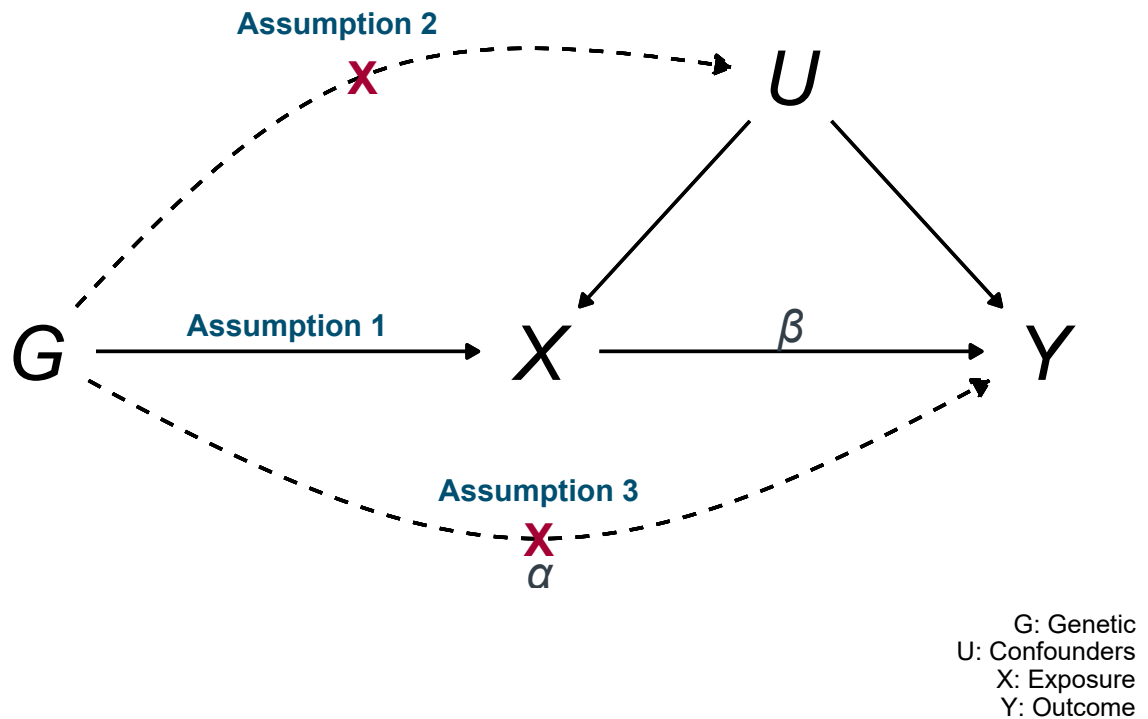


Figure 1: Causal Diagram Illustrating...

In practice, only the relevance assumption can be directly tested and proven. Typically, genetic variants for MR studies are selected as instruments based on Genome Wide Association Studies (GWAS), which quantify the associations between genetic Single Nucleotide Polymorphisms (SNPs) and various phenotypes. Association between a genetic variant and a phenotype representing an exposure of interest can be partly assured by selection using an appropriate genome-wide significance level (e.g. $p < 10^{-8}$). Separate statistical testing can also quantify the gene:exposure relationship. Commonly used measures include the r^2 statistic, which represents the proportion of variance in the exposure explained by the genotype, and the related F -statistic, which additionally accounts for the sample size under investigation⁷.

The assumptions of independence and exclusion restriction depend on all possible confounders of the exposure-outcome association, both measured and unmeasured; as such, these can never be proven absolutely. Various methods have been proposed to quantify and account for [@ *REF*].

instruments to an outcome of interest, genetic data can be used to investigate causal links between exposures and outcomes². by the Wald ratio, i.e. by dividing the co-efficient of gene-outcome association by the co-efficient of gene-exposure, giving a numerical measure of strength of causal exposure-outcome association³.

If these assumptions are satisfied, the “causal effect estimate” of the exposure on the outcome can be estimated by comparing measures of gene:exposure association with those of gene:outcome association as detailed below.

Threats to the independence assumption will vary depending on the population, exposure and outcome being studied[@ *REF*]. Exclusion restriction is a particularly universal issue in MR, due to so-called (horizontal) genetic pleiotropy, where a single genetic variant may have multiple “pleiotropic” effects – i.e. it may influence several traits simultaneously. Such pleiotropic effects may be unknown and open unmeasured causal pathways between a genetic instrument and the outcome, separate to the path involving the exposure of interest, thus potentially biasing MR estimates of the association between exposure and outcome⁴.

Weighted Median Estimator

Although not possible to prove exclusion restriction for any MR study, several methods attempt to produce exposure-outcome causal effect estimates which are robust to violations of this assumption. A common approach is the Weighted Median Estimator (WME) method, proposed by Bowden et al⁵.

In WME analysis, several genetic instruments are used to estimate the exposure-outcome causal effect. Each instrument is known to be associated with the exposure of interest, but an unknown proportion of these instruments may be invalid due to pleiotropic genetic effects. Any instrument linked to an outcome via multiple pleiotropic causal pathways will exhibit a less consistent gene-outcome association than a relationship mediated by a single pathway; this results in larger variance in causal estimates derived from invalid/pleiotropic genetic instruments versus estimates from valid instruments.

WME therefore assigns a weight to each genetic instrument’s estimate of the causal effect according to the inverse of the variance of the estimate; these weighted effect estimates are used to construct a cumulative distribution function for probability of true causal effect size across the range of estimated values. The 50th percentile of this distribution can then be taken as a “weighted median estimate” of the true causal effect, theoretically producing consistent causal estimates even if up to 50% of the included information comes from invalid instruments⁵. (Fig?)

1. Davies NM, Holmes MV, Smith GD. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ* [Internet]. 2018 Jul [cited 2025 Jan 7];362:k601. Available from: <https://www.bmj.com/content/362/bmj.k601>
2. Richmond RC, Smith GD. Mendelian Randomization: Concepts and Scope. *Cold Spring Harbor Perspectives in Medicine* [Internet]. 2022 Jan [cited 2024 Oct 22];12(1):a040501. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8725623/>

3. Burgess S, Foley CN, Allara E, Staley JR, Howson JMM. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications* [Internet]. 2020 Jan [cited 2025 Jan 7];11(1):376. Available from: <https://www.nature.com/articles/s41467-019-14156-4>
4. Hemani G, Bowden J, Smith GD. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics* [Internet]. 2018 May [cited 2024 Oct 23];27(R2):R195. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC6061876/>
5. Bowden J, Smith GD, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* [Internet]. 2016 Apr [cited 2024 Oct 22];40(4):304. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC4849733/>