# 2. Introduction and Background

## Contents



Word count: 1927

## Introduction and Background

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 analagous 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 for[1].

## Causal Effect Estimation in MR

At its simplest, the relationship between an exposure $X$ and outcome $Y$ (assuming both are continuous variables) can be represented as a linear model:

$$Y = \alpha + \beta X + \epsilon \tag{1}$$

where $\alpha$ represents all non-$X$ determinants of $Y$, $\beta$ is the causal effect of $X$ on $Y$ and $\epsilon$ is an error term. The $\beta$ term is therefore a numerical measure of strength of causal exposure-outcome association, where $\beta = 0$ implies no causal link between exposure and outcome, $\beta > 0$ implies $X$ causes $Y$, and $\beta < 0$ implies $X$ prevents $Y$. To estimate this causal effect using a genetic variant in an IV analysis, three key assumptions must be met[?]:

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.

Typically, MR studies estimate the true causal effect using several genetic instruments, generalised as $k$ instruments numbered $1, 2...k$; the effect estimate derived from the $jth$ instrument is denoted $\hat{\beta}_j$. Each estimate $\hat{\beta}_j$ acknowledges there will be specific effects on the observed values of exposure and outcome given the presence of that specific genetic variable $G_j$ under study, i.e. $\hat{\beta}_j$ is based on the observed exposure $X|G_j$ and outcome $Y|G_j$. These observed values of exposure and outcome can be described by their own linear models:

$$X|G_j = \gamma_0 + \gamma_j G_j + \epsilon_{X_j} \tag{2}$$

$$Y|G_j = \Gamma_0 + \Gamma_j G_j + \epsilon_{Y_j} \tag{3}$$

where, for exposure and outcome respectively:

- $\gamma_0$ and $\Gamma_0$ reflect base values without influence of the genetic variant, i.e. with two non-effect alleles of $G_j$
- $\gamma_j$ and $\Gamma_j$ are coefficients of association with the genetic variant, representing the extent to which the effect allele will perturb the value of $X$ or $Y$ versus the non-effect allele

- $\epsilon_{X_j}$ and $\epsilon_{Y_j}$ are error terms, containing contributions from confounders of the exposure-outcome relationship ($U$ in the causal diagram), and all genetic variants except $G_j$.

It can be shown that a simple causal effect estimate for the exposure on the outcome can be obtained from a single genetic instrument by the Wald method, dividing the coefficient of gene:outcome association by the coefficient of gene:exposure association, i.e.:

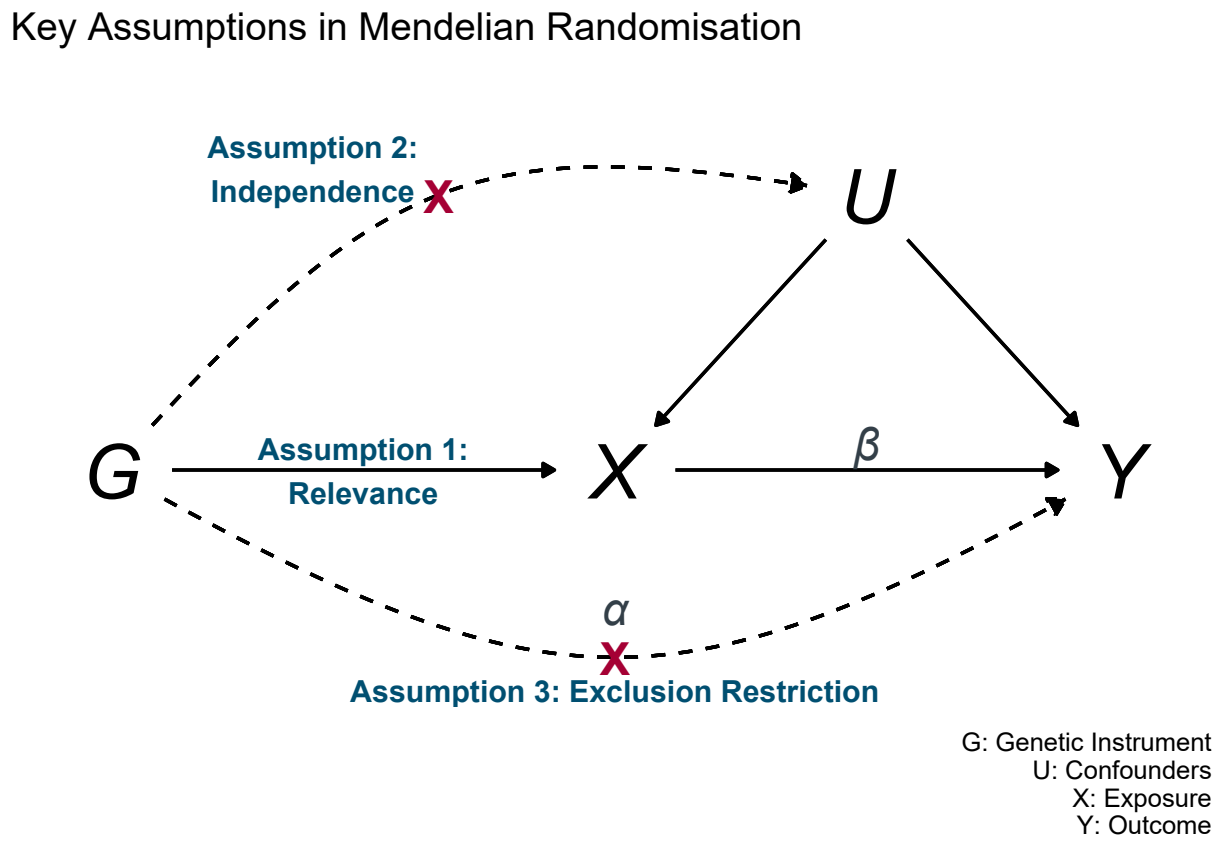


Figure 1: Causal diagram illustrating the relationships between genetic instrument, exposure, outcome and confounders of the exposure-outcome relationship. Crosses represent violations of assumptions that may lead to invalid inference. Adapted from Burgess et al 2016[2]

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} \tag{4}$$

Each instrument may be valid or invalid, depending on whether it meets all the above assumptions. The overall causal effect estimate $\hat{\beta}$ from any given MR method will typically seek to pool the effect estimates of all $k$ instruments in such a way as to minimise the effects of any invalid instruments included, e.g. by removing or down-weighting the contribution of genetic instruments which violate one or more assumptions. This is equivalent to plotting all coefficients of gene:outcome association ($\bar{\Gamma}$) versus all coefficients of gene:exposure association ($\bar{\gamma}$) for the set of $k$ instruments, then using the gradient of a regression line through the points as the causal effect estimate $\hat{\beta}$; picking an MR methodology would be analagous to choosing the method to draw the line of best fit - see Figure 2. For binary outcomes, the causal effect estimate can be converted to an odds ratio (OR) through exponentiation, i.e.:

$$OR = e^{\hat{\beta}} \tag{5}$$

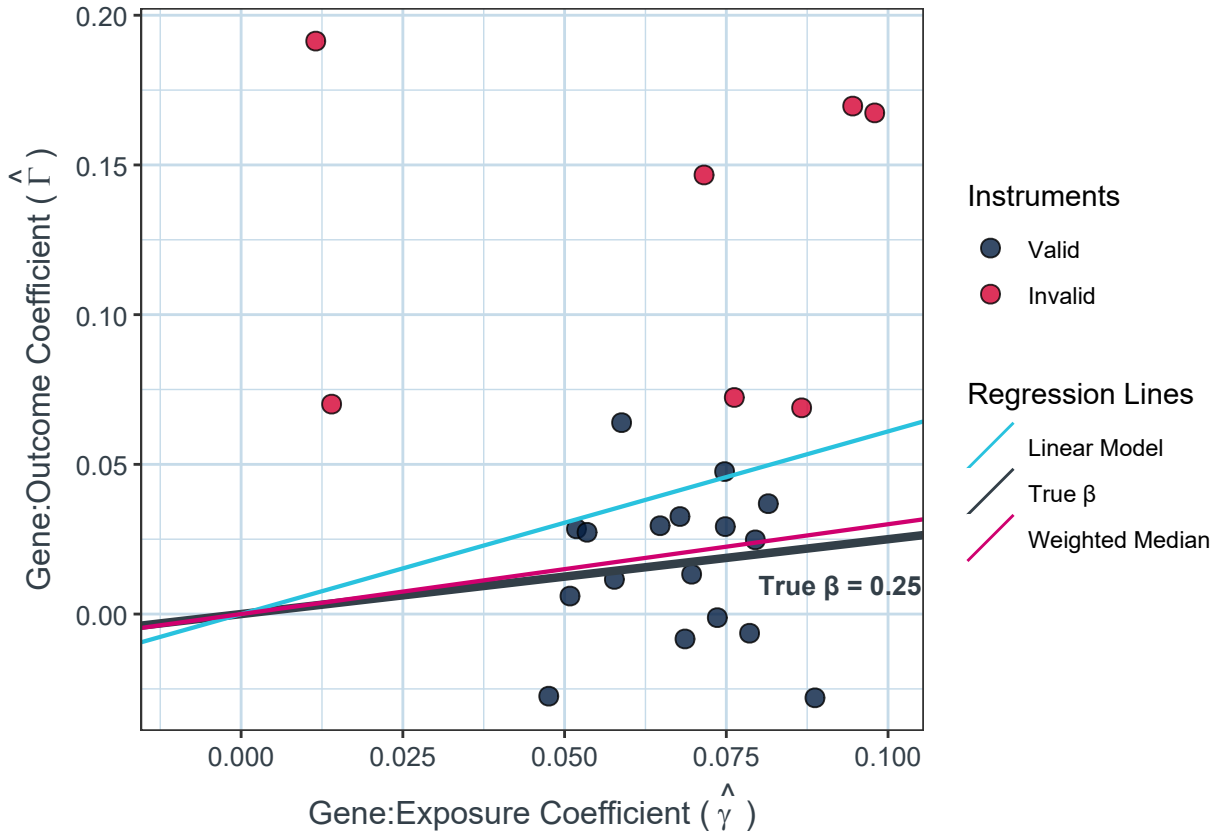


Figure 2: Simulated

## Violations to Assumptions

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 the sample size under investigation[3]. An $F$-statistic of $ $10 is generally considered to represent a strong enough gene:exposure association to consider a genetic instrument for use[?].

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 violations of these two additional assumptions, including the weighted median estimator, described below[4].

Threats to the independence assumption will vary depending on the population, exposure and outcome being studied. The main methods to avoid violations to this assumption relate to appropriate selection of populations studied to avoid confounding due to ancestry or population stratification. For example, in two-sample MR studies, gene-exposure and gene-outcome coefficients are estimated from two separate GWAS studies; it is common practice to select GWAS studies performed in similar population groups (e.g. both in Western Europeans). This practice helps avoid spurious exposure-outcome associations being generated by confounding due to underlying differences in allele frequency, baseline disease risks etc between different populations[3].

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 outcome[5]. Where pleiotropic effects are in both positive and negative directions with a mean of zero - "balanced pleiotropy" - then they only add noise to causal effect estimation[?]. By contrast, "directional pleiotropy", where the mean of pleiotropic effects is non-zero, may introduce bias away from the null, and inflate Type I error rate (false positives). the causal pathway acts directly between gene and outcome, this may be referred to variously as "direct genetic effects" or "uncorrelated pleiotropy"

## Weighted Median Estimator

A common approach to produce exposure-outcome causal effect estimates robust to violations of the exclusion restriction assumption is the Weighted Median Estimator (WME) method, proposed by Bowden et al[4].

In WME analysis, several genetic instruments are used to estimate the exposure-outcome causal effect $\hat{\beta}$. 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 instruments[4]. (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. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. Epidemiology (Cambridge, Mass) [Internet]. 2016 Nov [cited 2024 Oct 22];28(1):30. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5133381/

3. 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/

4. 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/

5. 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/