# Mendelian Randomization

2025-01-08

## Background

Mendelian randomization (MR) is used when it is either not possible or unethical to conduct randomized controlled trials. In essence MR uses genetic variants, such as single-nucleotide polymorphisms (SNPs), to estimate the causal effect on an outcome. The genetic variants are used since they are randomly allocated to individuals before any exposure or outcome, which eliminates many confounding factors. In the setting of MR genetic variants are also called instruments/instrumental variables. In this document they are called genetic instrument.

There are three core assumptions for the genetic instruments (Bowden, Smith, and Burgess 2015; Sekula et al. 2016). These are:

1. It must be reproducibly and strongly associated with the exposure
2. It must not be associated with any known confounding factors
3. It must not be associated with the outcome except through the exposure (Bowden, Smith, and Burgess 2015; Sekula et al. 2016).

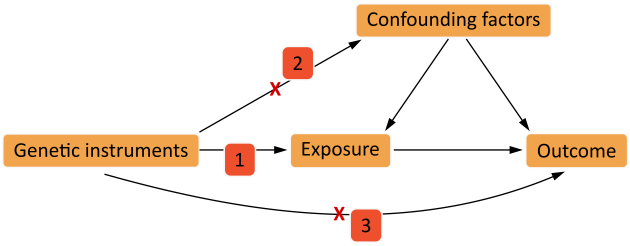


Figure 1: A schematic overview of MR analysis and its assumptions: **Assumption 1:** The instrument is reproducibly and strongly associated with the exposure. **Assumption 2:** The instrument is not associated with any known confounding factors. **Assumption 3:** The instrument is not associated with the outcome except through the exposure (Bowden, Smith, and Burgess 2015; Sekula et al. 2016).

The first assumption can be investigated and proven either right or wrong, but the second and third assumption cannot. Therefore some statistical methods for MR have taken this into consideration, but they have other assumptions:

4. **InSIDE assumption:** Instrument Strength Independent of Direct Effects: The pleiotropic effects must be distributed independently of the instrument strength (Bowden, Smith, and Burgess 2015; Bowden et al. 2016; Hartwig, Smith, and Bowden 2017).
5. **ZEMPA assumption:** ZEro Modal Pleiotropy Assumption: Across all instruments the most frequent value of the bias term is 0 → the most common causal effect estimate is a consistent estimate of the true causal effect, even if the majority of instrumetns are invalid (Hartwig, Smith, and Bowden 2017).

There are several steps of an MR analysis:

1. Define the presumed causal association of interest
2. Choose at least one genetic variant to use as the genetic instrument
    - The stronger the genetic instrument is associated with the exposure, the stronger the power is. If the association is not strong enough to reach the required power, more can be used, but this

increases the risk of bias, since it increases the risk of including invalid instruments.
3. Evaluate core assumptions and discuss their applicability
4. Carry out the statistical MR analysis
    - The statistical MR analysis depends on the available data → is the data individual level or summary? Is it a one sample or two sample MR analysis? Are there one or more genetic instruments?
5. Interpret and discuss results (Sekula et al. 2016).

As mentioned above some of the confounding factors of randomized controlled studies are avoided in MR analysis, but MR still has some limitations and possible confounding factors, these are:

- **Low statistical power:** MR studies can have low power if the sample size is too small or the genetic instrument does not account for enough of the investigated association (Smith and Hemani 2014).
- **Reverse causation** the genetic instrument may cause the disease outcome, which causes the biomarker or the causal direction may be in the opposite direcetion (Smith and Hemani 2014).
- **Population stratification:** The allele frequencies and disease exposure rates may vary between different subgroups of the population, which may lead to confounding (Sekula et al. 2016).
- **Reintroduced confounding through pleiotropy:** The gentic instrument is associated with more than one apparently unrelated trait or disease - it influences more than one post-transcriptional process, which leads to confounding (Sekula et al. 2016; Smith and Hemani 2014).
- **Linkage disequilibrium (LD) induced confounding:** The non-random association between different genetic variants on the same chromosomes, which leads to confounding if LD leads to the association of genetic instruments related to more than one post-transcriptional process (Sekula et al. 2016; Smith and Hemani 2014).
- **Canalization/developmental compensation:** Compensatory developmental processes buffers the effect of genetic instruments that leads to potentially disruptive influences (Sekula et al. 2016).
- **Lack of genetic instruments to proxy for modifiable exposure interest:** If there is no reliable genetic instrument associated with the exposure, the MR cannot be performed (Smith and Hemani 2014).
- **Complexity of associations:** If there is not enough biological knowledge regarding the genetic instrument, the exposure, the outcome and the associations of these, wrong conclusions may be drawn from the analysis (Smith and Hemani 2014).
- **Weak instruments:** If the association between the genetic instrument and the exposure is weak, the exposure-outcome association may be biased and therefore the genetic instrument may give misleading results (Sekula et al. 2016).

Therefore it is best to use a genetic instrument with a functional association to the exposure/a genetic instrument that is located in genes with biologic functions that are well known so it is more simple to assess whether or not the assumptions are met (Sekula et al. 2016).

There are several types of MR studies/analyses some of these are one-sample MR, two-sample MR, network MR and bidirectional MR. This document will focus on the two-sample MR analysis using summary data. Here two individual samples are used where one has the measurements for the exposure, and the other has the measurements for the outcome. Both samples have the same genetic instruments (Boehm and Zhou 2022). When performing a two-sample MR study one has to consider two additional assumptions:

- The two samples must represent the same underlying population.
- There is no overlap in participants between the two samples (Davies, Holmes, and Smith 2018).

When a presumed causal association of interest has been chosen the rest of the MR analysis can be performed in R using the Two Sample MR analysis package (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017). The Two Sample MR analysis package uses the following statistical methods:

- Wald Method
- Inverse Variance Weighting (IVW)
- Weighted Median Estimator
- MR-Egger

- Simple mode
- Weighted mode (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017).

Each method has its own weaknesses and strengths. (They may be performed all at once to see if they 'agree' but one should be aware if the assumptions for the specific method is met.) An overview of some of the limitations of these methods can be seen in the figure below.

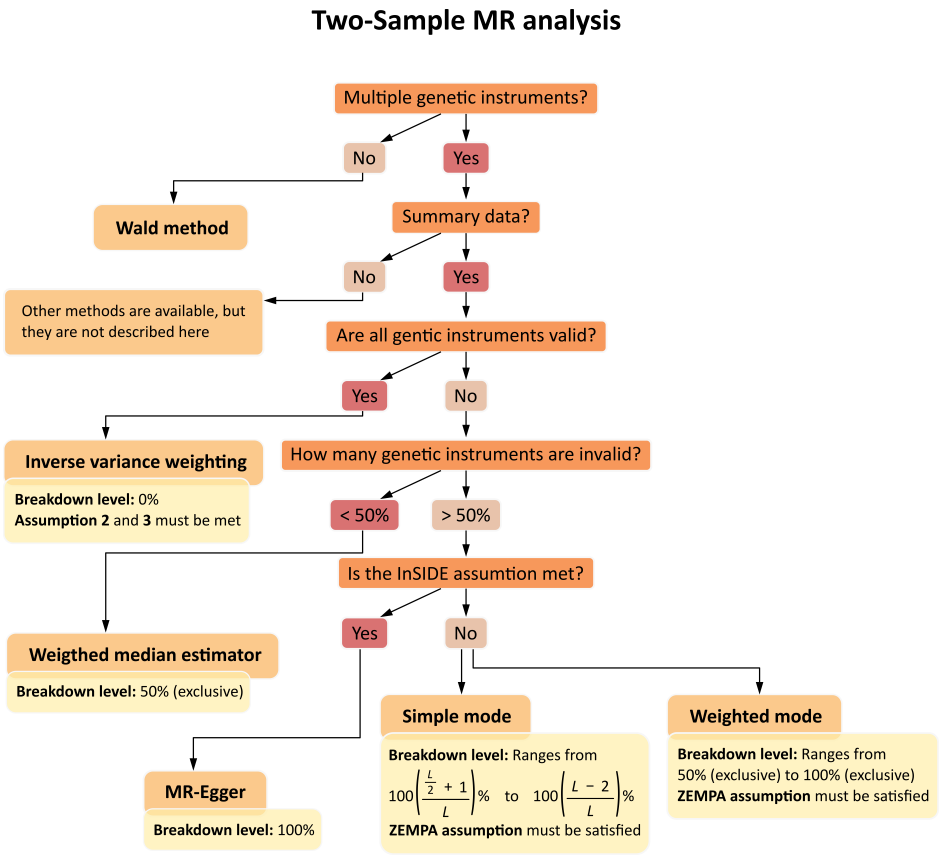


Figure 2: An overview of some of the statistical methods for two-sample MR analysis: **Breakdown level:** The percent of variants that can be invalid, while still providing a consistent estimate (Boehm and Zhou 2022; Hartwig, Smith, and Bowden 2017). **Assumption 1:** The instrument is reproducibly and strongly associated with the exposure. **Assumption 2:** The instrument is not associated with any known confounding factors. **Assumption 3:** The instrument is not associated with the outcome except through the exposure (Bowden, Smith, and Burgess 2015; Sekula et al. 2016). **InSIDE assumption:** Instrument Strength Independent of Direct Effects: The pleiotropic effects must be distributed independently of the instrument strength (Bowden, Smith, and Burgess 2015; Bowden et al. 2016; Hartwig, Smith, and Bowden 2017). **ZEMPA assumption:** ZEro Modal Pleiotropy Assumption: Across all instruments the most frequent value of the bias term is 0 $\rightarrow$ the most common causal effect estimate is a consistent estimate of the true causal effect, even if the majority of instrumetns are invalid (Hartwig, Smith, and Bowden 2017). L: the number of genetic instruments.

## The statistical analyses

To create an overview of the logical flow in the following statistical analyses Figure 1 can be expanded to include the notations given to the variables and assumptions as seen in Figure 3.

All of the statistical methods start by defining two linear regressions. One regression of the exposure on the genetic instruments and one of the outcome on the genetic instruments. This can be done by defining the following: If:

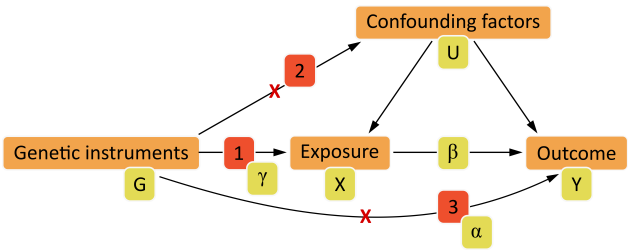


Figure 3: A schematic overview of MR analysis and its assumptions expanded: The yellow boxes indicate the notation given to the assumption or variable in the statistical methods. **Assumption 1:** The instrument is reproducibly and strongly associated with the exposure. **Assumption 2:** The instrument is not associated with any known confounding factors. **Assumption 3:** The instrument is not associated with the outcome except through the exposure (Bowden, Smith, and Burgess 2015; Sekula et al. 2016).

- $G_j$ is the jth genetic instrument out of L variables ($j = 1, ..., L$),
- $X$ is the exposure,
- $Y$ is the outcome
- $U$ is the confounding variables,
- $\Gamma_j$ is the genetic association with the outcome
- $\gamma_j$ is the genetic association with the exposure
- $\beta$ is the causal effect of the exposure on the outcome

Then for the ${G_j}^{th}$ instrument the linear regressions are:

$$X_{G_j} = \gamma_0 + \gamma_j G_j + \varepsilon_{Xj} \tag{1}$$

$$Y_{G_j} = \Gamma_0 + (\beta\gamma_j + \alpha_j)G_j + \varepsilon_{Yj} = \Gamma_0 + \Gamma_j G_j + \varepsilon_{Yj} \tag{2}$$

Since the setting is a two-sample setting the error terms $\varepsilon_{Xj}$ and $\varepsilon_{Yj}$ are independent and it is assumed that they are normally distributed and contain contributions from the confounders and all genetic instruments except $G_j$. Assumption 1 says that $\gamma_j \neq 0$ for all $j$. Assumption 2 and 3 say that the genetic associations with the outcome are equal to the genetic associations with the exposure multiplied by the causal effect of the exposure on the outcome:

$$\Gamma_j = \beta\gamma_j + \alpha_j \tag{3}$$

Where $\alpha_j = 0$ (Bowden et al. 2016)

$\beta\gamma_j$ is the effect of $G_j$ on $Y$ through $X$, where we want to estimate $\beta$ which is the causal effect of $X$ on $Y$ and $\alpha_j$ represents the association between $G_j$ and $Y$ not through the exposure of interest caused by horizontal pleiotropy (Hartwig, Smith, and Bowden 2017).

The different methods calculate the estimate of the causal effect of $X$ on $Y$ in different ways.

### The Wald method

In the Wald method the slope estimates/regression coefficients from the two linear regressions, shown above are used to calculate the Wald ratio (Boehm and Zhou 2022):

$$\hat{\beta}_{wald} = \frac{\hat{\Gamma_j}}{\hat{\gamma_j}} \tag{4}$$

This method can be used on GWAS summary statistics however it can only be used if there is one genetic instrument (Boehm and Zhou 2022; Bowden, Smith, and Burgess 2015; Sekula et al. 2016).

### The Inverse Variance Weighting Method

The inverse variance weighting (IVW) method consists of a weighted linear regression of the genetic associations with the outcome on the genetic associations with the exposure is performed, this is weighted by the inverse-variance of the genetic associations with the outcome ($\sigma_{Yj}^{-2}$). Here the intercept is constrained to equal zero (Bowden, Smith, and Burgess 2015). Each genetic instrument's effect is weighted by the inverse of the variance of the ratio estimator. The complete causal effect is then the sum of the weighted genetic instrument's causal effects (Boehm and Zhou 2022).

$$\hat{\beta}_{IVW} = \frac{\sum_j \hat{\gamma_j}^2 \sigma_{Yj}^{-2} \hat{\beta}_j}{\sum_j \hat{\gamma_j}^2 \sigma_{Yj}^{-2}} \tag{5}$$

Since

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\gamma_j} \tag{6}$$

This can also be written as:

$$\hat{\beta}_{IVW} = \frac{\sum_j \hat{\gamma_j} \sigma_{Yj}^{-2} \hat{\Gamma}_j}{\sum_j \hat{\gamma_j}^2 \sigma_{Yj}^{-2}} \tag{7}$$

This method can be used on several genetic instruments and GWAS summary statistics (Boehm and Zhou 2022). The breakdown level of IVW is 0%, meaning that all the genetic instruments must be valid for the method to provide a consistent estimate (Bowden et al. 2016; Hartwig, Smith, and Bowden 2017). If some of the genetic instruments are invalid, and directional pleiotropy is present, the estimates given by IVW suffers from bias and increased Type 1 error rates (Bowden et al. 2016).

When the genetic instrument is invalid and $\alpha \neq 0$ the ratio estimate is then equal to the true causal effect $\beta$ plus the bias term $\frac{\alpha_j}{\gamma_j}$ and therefore IVW will tend toward:

$$\beta + \frac{\sum_{j_1}^{J} \gamma_j \sigma_{Yj}^{-2} \alpha_j}{\sum_{j=1}^{J} \hat{\gamma_j}^2 \sigma_{Yj}^{-2}} = \beta + Bias(\alpha, \gamma) \tag{8}$$

(Bowden, Smith, and Burgess 2015)

### The Weighted Median Estimator Method

It is assumed that the genetic variants are uncorrelated, all genetic instruments are valid, the relationships between the variables are linear without heterogeneity or effect modification. Remember equations 1-4 and equation 6. Simply put, here the median of all the causal effect estimates using the genetic instruments is found.

Before attempting the weighted median estimator method it is helpful to understand the simple median estimator. If $\hat{\beta}_j$ is the jth ordered ratio estimate, where all the calculated ratio estimates are arranged from smallest to largest, and the total number of genetic instruments are odd ($J = 2k + 1$) then the simple median estimator is the middle ratio estimate $\hat{\beta}_{k+1}$, where $k = \frac{J-1}{2}$. If the total number of genetic instruments is even $J = 2k$ then the simple median estimator is interpolated between the two middle estimates $\frac{1}{2}(\hat{\beta}_k + \hat{\beta}_{k+1})$ where $k = \frac{J}{2}$. The simple median estimator has a breakdown level of 50% (exclusively) (Bowden et al. 2016).

The simple median estimator is inefficient when the individual estimates varies a lot. The weighted median estimator accounts for this. $w_j$ is the weight given to the jth ordered ratio estimate. The weighted median estimator is the median of a distribution having estimate $\hat{\beta}_j$ as its $p_j = 100(s_j - \frac{w_j}{2})$th percentile, where $s_j$ is the sum of weights up to and including the weight of the jth ordered estimate ($s_j = \sum_{k=1}^{j} w_k$). The weights are standardized so the sum of weights $s_J$ is 1. The inverse of the variance of the ratio estimates can be used as weights, giving:

$$w_j^{'} = \frac{\hat{\gamma}_j^2}{\sigma_{Y_j}^2} \tag{9}$$

Bowden et al. (2016) uses only the first-order term from the delta expansion. The standardized weights are then

$$w_j = \frac{w_j^{'}}{\sum_j w_j^{'}} \tag{10}$$

(Bowden et al. 2016).

The weighted median gives a consistent estimate if at least 50% of the weight comes from valid genetic instruments, where it is assumed that no single genetic instrument contributes more than 50% of the weight, so it has a breakdown level of 50% (exclusively) (Bowden et al. 2016).

**The MR-Egger Regression Method**

The weighted linear regression is performed without the intercept being constrained to zero. Here the intercept represents the average pleiotropic effect across the genetic instruments. The MR-Egger test, then consists of assessing whether or not the intercept differs from zero. If this is the case it indicates the presence of directional pleiotropy. If the InSIDE assumption is satisfied the slope coefficient from the MR-egger regression is a consistent estimate of the causal effect. The InSIDE assumption is violated if the pleiotropic effects act through a confounder of the exposure (Bowden, Smith, and Burgess 2015).

If a genetic instrument is not valid because it has a direct effect on the outcome (assumption 3 is not satisfied), then $\alpha \neq 0$ and the ratio estimate based on the genetic instrument j is (in an infinite sample) then equal to the true causal effect $\beta$ plus the bias term $\frac{\alpha_j}{\gamma_j}$:

$$\beta_j = \beta + \frac{\alpha_j}{\gamma_j} \tag{11}$$

The IVW estimate is the slope of the best fitting lines through the data points, that also passes through the origin. When the InSIDE assumption is satisfied, $\hat{\alpha}_j$ is independent of $\hat{\gamma}_j$, therefore the bias of the ratio estimate $\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$ is inversely proportional to $\gamma_j$

When regression of the $\hat{\Gamma}_j$ coefficients on the $\hat{\gamma}_j$ and the intercept is not constrained to zero the following linear model is found/made, which performs Egger regression:

$$\hat{\Gamma}_j = \beta_{0E} + \beta_E \hat{\gamma}_j \tag{12}$$

If the intercept term ($\beta_{0E}$) is not zero it indicates the presence of directional pleiotroy (this is the Egger's test), therefore the estimated value of ($\beta_{0E}$) can be used as an estimate of the average pleiotropic effect. $\hat{\beta}_E$ is a bias-reduced estimate, since (under model 12) the following is true for the slope coefficient from Egger regression:

$$\hat{\beta}_E = \frac{cov(\hat{\Gamma}, \hat{\gamma})}{var(\hat{\gamma})} = \hat{\beta} + \frac{cov(\hat{\alpha}, \hat{\gamma})}{var(\hat{\gamma})} \tag{13}$$

Because of the InSIDE assumption the following is then true:

$$cov(\hat{\alpha}, \hat{\gamma}) \xrightarrow{N \to \infty} cov(\alpha, \gamma) \xrightarrow{J \to \infty} 0 \tag{14}$$

Meaning that $\hat{\beta}_E$ is a consistent estimate of $\beta$ (the causal effect) (Bowden, Smith, and Burgess 2015).

If the InSIDE assumption is satisfied the breakdown level of the MR-Egger method is 100% (Bowden et al. 2016; Hartwig, Smith, and Bowden 2017). However the MR-Egger often has low statistical power for effect estimation, and it has been suggested that it should primarily be used to reveal pleiotropy, as a sensitivity analysis, in addition to the other methods (Haycock et al. 2016; Burgess and Thompson 2017).

**The Simple Mode and Weighted Mode Method**

When calculating the mode-based estimate, the mode of the smoothed empirical density function of all $\hat{\beta}_j$s is used as the causal effect estimate, but before doing this a few things must be defined.

It is assumed that all L genetic variants are independent of each other.

Remember $\beta\gamma_j$ is the effect of $G_j$ on Y through X, where we want to estimate $\beta$ which is the causal effect of X on Y and $\alpha_j$ represents the association between $G_j$ and Y not through the exposure of interest caused by horizontal pleiotropy.

When $\alpha_j \neq 0$ then $\hat{\beta}_j = \beta + b_j$ where $b_j = \frac{\alpha_j}{\gamma_j}$ also called the bias term. For the simple mode and weighted mode to consistently estimate the true causal effect, the ZEMPA assumption must be satisfied, which says that across all the instruments, the most frequent value (the mode) of $b_j$ is 0 (Hartwig, Smith, and Bowden 2017).

If $k \in \{1, 2, ..., L\}$ represents the number of unique values of $b_j$ among L variants, then if all $b_j$ terms are identical then $k = 1$, but if they are all unique then $k = L$. Additionally $n_1, n_2, ..., n_k$ represents the number of genetic instruments that have identical non-zero value of $b_j$, $n_1$ having the smallest non-zero value and $n_k$ having the largest non-zero value, $n_0$ is then the number of genetic instruments with $b_j = 0$ and these are the valid instruments. The number of variants is then:

$$L = n_0 + n_1 + n_2 + ... + n_k \tag{15}$$

When ZEMPA is satisfied, the following must then be true:

$$n_0 > \max(n_1, ..., n_k) \tag{16}$$

Meaning that the number of valid genetic instruments is larger than any other group of genetic instruments with the same bias term (Hartwig, Smith, and Bowden 2017).

The breakdown level of the simple mode method ranges from:

$$100\left(\frac{\frac{L}{2}+1}{L}\right)\% \text{ to } 100\left(\frac{L-2}{L}\right) \tag{17}$$

Where the lower limit is the situation where there are some valid genetic instruments, but all the invalid genetic instruments estimate the same causal effect parameter - they all have the same bias term, meaning that the ZEMPA assumption is satisfied when up to, but not including, half of the instruments are invalid. The upper limit is the situation where all invalid instruments have different bias terms, so they estimate different causal effect parameters ($n_1 = n_2 = ... = n_k = 1$). Here ZEMPA is satisfied if only two of the genetic instruments are valid and the remainder invalid. When the larges number of identical estimates comes from invalid instruments ZEMPA is violated (Hartwig, Smith, and Bowden 2017).

In the weighted mode method ZEMPA is satisfied if the weights associated with the valid genetic instruments are the largest among all $k$ subsets of instruments, giving:

$$w_0 > max(w_1, ..., w_k) \tag{18}$$

The breakdown level of the weighted mode method ranges from 50% (exclusive) to 100% (exclusive). As the number of genetic instruments increases the lower and upper limits of the breakdown level tend toward 50% and 100% (Hartwig, Smith, and Bowden 2017).

As mentioned in the beginning of this section the mode of the smoothed empirical density function of all $\hat{\beta}_j$s is used as the causal effect estimate. The standardized weights for the weighted mode estimate can be found using:

$$w_j = \frac{\sigma_{Rj}^{-2}}{\sum_{j=1}^{L} \sigma_{Rj}^{-2}} \tag{19}$$

Where $\sigma_{Rj}^{-2}$ is the standard error of $\hat{\beta}_j$ and can be calculated by:

$$\sigma_{Rj} = \sqrt{\frac{\sigma_{Yj}^2}{\hat{\gamma}_j^2} + \frac{\hat{\Gamma}_j^2 \sigma_{Xj}^2}{\hat{\gamma}_j^4}} \tag{20}$$

For the simple mode estimate $w_1 = w_2 = ... = w_L = 1/L$

The normal kernel density function of the $\hat{\beta}_j$s is:

$$f(x) = \frac{1}{h\sqrt{2\pi}} \sum_{j=1}^{L} w_j \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\beta}_j}{h}\right)^2\right] \tag{21}$$

$h$ is the smoothing bandwidth parameter and it regulates a bias-variance trade-off of the mode-based method, meaning that higher $h$ leads to higher precision and higher bias. $h$ can be found by:

$$h = \phi s \tag{22}$$

$\phi$ is a tuning parameter which allow increased or decreased bandwidth. $s$ is the default bandwidth chosen according to a criteria. Hartwig, Smith, and Bowden (2017) uses the modified Silverman's bandwidth rule:

$$s = \frac{0.9 \min(sd(\hat{\beta}_j), 1.4826 mad(\hat{\beta}_j))}{L^{\frac{1}{5}}} \tag{23}$$

Here $sd(\hat{\beta}_j)$ is the standard deviation and $mad(\hat{\beta}_j)$ is the median absolute deviation from the $L\hat{\beta}_j$s.

The causal effect estimate obtained using this method ($\hat{\beta}_M$) is the value of $x$ that maximizes $f(x)$ $\left(f\left(\hat{\beta}_M\right) = \max[f(x)]\right)$ (Hartwig, Smith, and Bowden 2017).

This theoretical and mathematical background is useful when performing the MR analysis, since it helps when choosing which method to use. However several of the methods can be performed at once and the results compared to see if they "agree", but it is important to remember the assumptions behind the methods and discuss if they are met, since all the methods may produce an estimate, but not all may be reliable in your situation.

## Interpreting and understanding MR analysis results in research articles

One thing is understanding the mechanics behind the MR studies, another thing is understanding how research articles present the results from their MR analyses. This section will try to provide some insight into this.

The first thing to remember is that all of the methods give an estimate of $\beta$. If this estimate is significant and positive it indicates that there is a causal relationship between the exposure and the outcome.

A scatter plot can be presented, with all the genetic instruments (SNPs), where the genetic instrument-outcome association is on the Y-axis and the genetic instrument-exposure association is on the X-axis. In the plot the 95% confidence interval for the exposure and outcome association, for each genetic instrument is also shown as horizontal and vertical lines from each data point representing a genetic instrument. For the IVW method a line is then added which joins the data points to the origin. In this way the plot visualizes the used genetic instruments and it may reveal if some of the genetic instruments perhaps are be outliers. If outliers are present, it may affect the estimated $\beta$ value given by the different methods, where the MR-Egger method is more influenced by outliers than the simple mode and weighted mode (Bowden and Holmes 2019; Burgess and Thompson 2017). If a line representing the MR-Egger regression is included in the scatter plot, this can be used to see if directional pleiotropy is present in the data. However the InSIDE assumption must be met for this sensitivity analysis to hold. If the InSIDE assumption is satisfied and the intercept of MR-Egger

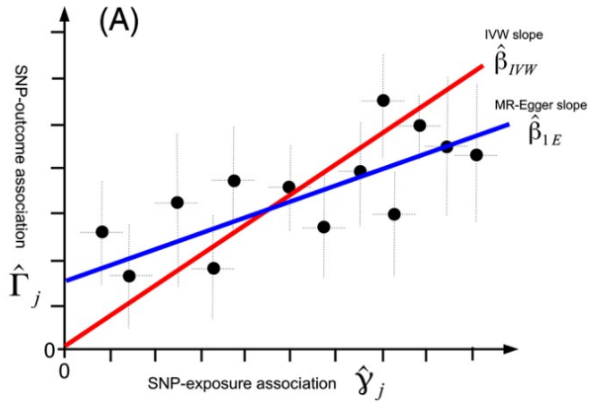


Figure 4: An example of a scatter plot where the data point for the genetic instruments are plotted according to their association with the exposure and outcome, the 95% confidence intervals for these associations are also shown with the horizontal and vertical lines protruding from each data point. The lines represent the ratio estimate found using either the IVW method or the MR-Egger method. The figure is from Bowden and Holmes (2019).

regression differs from zero it could indicate the presence of directional pleiotropy (Bowden and Holmes 2019). In Figure 4 an example of such a scatter plot can be seen, the scatter plot is from Bowden and Holmes (2019).

A funnel plot can also be presented with the data points from each individual genetic instrument, where the causal effect estimates are on the X-axis and the square-root precision is on the Y-axis. The funnel plot will be symmetrical if there is either no pleiotropy or balanced pleiotropy. If the InSIDE assumption holds the MR-Egger estimate in an asymmetrical funnel plot can then be interpreted as the value that a symmetrical would have produced (Bowden and Holmes 2019). An example of such a funnel plot can be seen in Figure 5, the funnel plot is from Bowden and Holmes (2019).

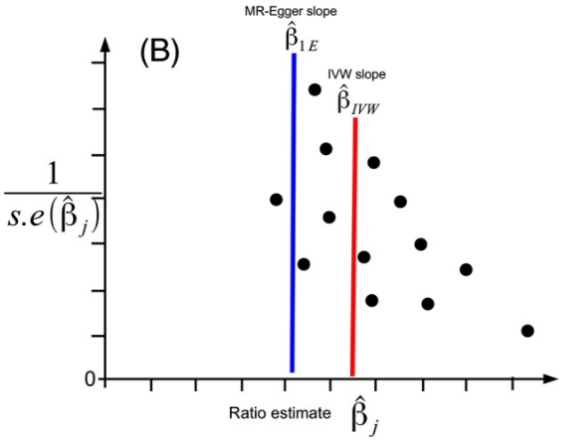


Figure 5: An example of a funnel plot where the data point for the genetic instruments are plotted according to their causal effect estimates and the square-root precision. Since the funnel plot is not symmetrical it indicates the presence of directional pleiotropy. The figure is from Bowden and Holmes (2019).

Additionally Q and Q' statistics can also be shown, which can also help interpreting the likelihood of statistical heterogeneity around the IVW and MR-Egger estimates due to horizontal pleiotropy, especially if the Q-Q' difference is great. If this is the case it may be relevant to check if an outlier analysis has been performed, to examine if one genetic instrument drives the heterogeneity (Bowden and Holmes 2019).

Some papers also present a forest plot like the ones seen in meta-analyses, either over the combined odds

ratio from the different methods they have used, or showing all the odds ratios (Wald ratios) from the genetic instruments used. The last type of forest plot, may be a result of a leave-one-out analysis, or just to show the individual Wald ratios.

The paper by Davies, Holmes, and Smith (2018) also explains how to be critical of results from MR studies and provides a list of questions one can ask oneself when reading research papers concerning MR studies.

## Using the Two Sample MR analysis package

The package MRCIEU/TwoSampleMR can be used to perform an MR analysis (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017; Hermani, n.d.c). This is done in four steps:

1. Selecting the instruments for the exposure
2. Acquiring the instruments from the IEU GWAS database for the outcome
3. Harmonizing the effect sizes for the instruments on the exposure and outcomes to be for the same reference allele.
4. Performing the MR analysis (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017; Hermani, n.d.c).

### Selecting the instruments for the exposure

The R commands used in this step are:

- `extract_instruments()`
- `read_exposure data()`
- `library(MRInstruments)`
  - `data(gwas_catalog)`
  - `data(aries_mqtl)`
  - `data(gtex_eqtl)`
  - `data(proteomic_qtls)`
  - `data(metab_qtls)`
- `clump_data()`

To perform the analysis this package need the information concerning the instruments in a data frame, where each line has the information for one variant for one exposure. Different types information concerning the instrument is needed as seen in the table below.

Table 1: An overview of the information that can be used in the two-sample mendelian analysis. *This information is necessary for the analysis.

| Type of information | name/label |
|---|---|
| The rs ID* | `SNP` |
| The effect size (is the trait binary, use log(OR))* | `beta` |
| The standard error of the effect size* | `se` |
| The allele of the SNP which has the effect `beta`* | `effect_allele` |
| The non-effect allele | `other_allele` |
| The effect allele frequency | `eaf` |
| The name of the phenotype the SNP has an effect on | `phenotype` |
| Physical position of the variant (chromosome) | `chr` |
| Physical position of variant (position) | `position` |
| Sample size for estimating the effect size | `samplesize` |
| The number of cases | `ncase` |
| The number of controls | `ncontrol` |
| The P-value for the SNP's association with the exposure | `pval` |
| The units in which the effects are presented | `units` |

| Type of information | name/label |
|---|---|
| The gene or other annotation for the SNP | `gene` |

This information/exposure data can either be acquired from a text file using the `read_exposure_data` function, where the file has a header with column names according to the labels in the table above. If the text file does not have column names according to the labels in the table above, they must be defined, so the data is read correctly. This can be done in the `read_exposure_data`:

```
#library(TwoSampleMR)
#name_of_file <- system.file("place of file")
#exposure_phenotype_exp_dat <- read_exposure_data(
#     filename = name_of_file,
#     sep = ",", #if it is a comma seperating the fields
#     snp_col = "name of snp column in file"
#     beta_col = "name of effect size column in file"
#     se_col = "name of se column in file"
#     effect_allele_col = "name of effect allele column in file"
#     other_allele_col = "name of non-effect allele column in file"
#     eaf_col = "name of effect allele frequency column in file"
#     pval = "name of the pval column in file"
#     units_col = "name of the units column in file"
#     gene_col = "name of the gene column in file"
#     samplesize_col = "name of the sample size column in file"
#)
```

In case the `phenotype` column is not in the imported file, is will be assumed that the phenotype's name is "exposure", which is entered in the exposure column, but it can be renamed by

```
#exposure_phenotype_exp_dat$exposure <- "phenotype"
```

If data is acquired from a previous file, the `read_exposure_data` converts the data to a data frame with standardized column names

The exposure data can also be acquired from a previous data frame, the `format_data()` function will convert the data to the correct format.

GWAS databases can also be browsed for instruments through the `MRInstruments` package. The GWAS catalog can be found/browsed the following way:

```
#library(TwoSampleMR)
#library(MRInstruments)
#data(gwas_catalog)
#head(gwas_catalog)
```

A list of studies is then shown. A study can then be chosen from which the instruments will be found and the data defined in a variable named `exposure_phenotype_exp_dat` as shown below:

```
#exposure_phenotype_gwas <-
#   subset(gwas_catalog,
#          grepl("Authorname", Author) &
#            Phenotype == "exposure_phenotype")
#exposure_phenotype_exp_dat <- format_data(exposure_phenotype_gwas)
```

The IEU GWAS database can also be accessed to define instruments for an exposure.

When the data has been acquired it is important to ensure that the instruments (for exposure) are independent. This can be done using the `clump_data()` function.

```
#exposure_phenotype_dat <- clumping_data(exposure_phenotype_exp_data_from_study
#str(exposure_phenotype_dat))
```

*A more thorough review of this step can be found here.* (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017; Hermani, n.d.a)

### Acquiring the instruments from the IEU GWAS database for the outcome

The R commands used in this step are:

- `extract_outcome_data()`
- `read_outcome_data()`
- (`available_outcomes()`)

Now the instruments associated with the exposure trait has been identified. The next step is to extract these genetic variants from the outcome data. Remember that we are looking for the same genetic instruments as we have just found, but now associated with the outcome, and from a different study population.

Here the IEU GWAS database can be browsed again. Details about the available GWASs can be found the following way:

```
#library(TwoSampleMR)
#ao <- available_outcomes()
#head(ao)
```

The `available_outcomes()` function gives a table of the available studies (each with a unique ID) in the database. The ID of the study is used, when extracting SNPs from a specific study the following way:

```
#exposure_phenotype_exp_dat <- extract_instruments(outcomes = "study ID")
#head(exposure_phenotype_exp_dat)
```

```
#ao[grepl("outcome trait", ao$trait), ]
```

This gives a list of possible GWAS studies, the wanted SNPs from the selected study can then be acquired by:

```
#outcome_phenotype_out_dat <- extract_outcome_data(
#    snps = exposure_phenotype_exp_dat$SNP,
#    outcomes = 'study ID'
#)
```

In `the extract_outcome_data()` function the `snps` argument only require an array of rsIDs and the `outcomes` argument can be a vector of outcomes.

If a specifically requested SNP is not present in the outcome GWAS then a SNP (proxy) that is in linkage disequilibrium with the requested SNP (target) is searched for instead.

Local GWAS summary data can also be used

*A more thorough review of this step can be found here.* (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017; Hermani, n.d.d)

### Harmonizing the exposure and outcome data

The R commands used in this step are:

- `harmonise_data()`
- `power_prune()`

Now you have both the exposure and outcome data saved in the two variables `exposure_phenotype_exp_dat` and `outcome_phenotype_out_dat`.The next step is to harmonize the effect of a SNP on the exposure and the effect of that SNP on the outcome, so they correspond to the same allele. This is done the following way:

```
#library(TwoSampleMR)
#dat <- harmonise_data (
#  exposure_dat = exposure_phenotype_exp_dat,
#  outcome_dat = outcome_phenotype_out_dat
#)
```

The output is a new data frame where the exposure and outcome data is combined.

There three ways to harmonize data:

1. Assume that all alleles are presented on the forward strand.
2. Try to infer the forward stand alleles using allele frequency information.
3. Correct the strand for non-palindromic SNPs, but drop palindromic SNPs.

The `harmonise_data` function uses option 2 as a default, but by using the `action` argument, this can be modified: `harmonise_data(exposure_dat, outcome_dat, action = 3)`

The dataset may contain duplicate exposure-outcome summary sets after data harmonisation. It is recommended that duplicate exposure-outcome summary sets are removed, so only the exposure-outcome combination with the highest expected power remains, this can be done using the `power_prune()` function.

```
# dat <- power_prune(dat, method = 1, dist.outcome = "binary")
```

The method can be set to either 1 or 2. When the method is set to 1, the duplicate exposure-outcome sets with the smaller outcome sample size are removed. If there are still duplicates, they are removed according to the exposure sample size. If there are many SNPs available to instrument an exposure, the outcome GWAS with the better SNP coverage might have better power than the outcome GWAS with the larger sample size. In this case the method is set to 2. The studies are the removed according to instrument strength as well as sample size. Here it is assumed that the SNP-exposure effects correspond to a continuous trait with a normal distribution. If the exposure is binary then the method should be set to 1.

*A more thorough review of this step can be found here.* (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017; Hermani, n.d.b)

### Performing the MR analysis

The R commands used in this step are:

- `mr()`
- `mr_singlesnp()`
- `leaveoneout()`
- `mr_heterogeneity()`
- `mr_steiger()`
- `mr_pleiotropy_test()`

When we have acquired the exposure data and the outcome data and this has been harmonized the MR analysis can be performed, since the effects and standard errors for each instrument SNP present/available for the exposure and outcome traits has now been acquired. The MR analysis is performed using the `mr()` function.

```
#library(TwoSampleMR)
#library(ggplot2)
#res <- mr(dat)
```

The output is a data frame where the causal effect of the exposure on the outcome is calculated using five different MR methods (MR Egger, Weighted median, Inverse variance weighted, Simple mode, Weighted mode). If only some of these MR methods should be performed, they can be specified the following way:

```
#mr(dat, method_list = c("wanted MR analysis1", "wanted MR analysis2))
```

The name of the wanted analysis that should be specified can be found using the `mr_method_list` function, which yields a long list of MR methods and their given names in the package.

A heterogeneity test can be performed using the `mr_heterogeneity()` function, here that should be used can also be specified.

A horizontal pleiotropy test, where the intercept term in the MR Egger regression is used, can also be performed using the `mr_pleiotropy_test()` function.

If several MR estimates using each of the SNPs is desired the `mr_singlesnp()` function can be used. The output of this function is a data frame where the MR analysis has been performed several times for each exposure-outcome combination, where a different single SNP has been used every time. The Wald ratio is used to perform the analysis, but if the fixed effects meta analysis method should be used instead, it can be specified the following way:

```
#res_single <- mr_singlesnp(dat, single_method = "mr_meta_fixed")
```

The `mr_singlesnp()` function also calculates the full MR using all available SNPs using IVW and MR Egger. It can be specified the following way:

```
#res_single <- mr_singlesnp(dat, all_method = "mr_two_sample_ml")
```

Then only the maximum likelihood method for the combined test is performed. The `mr_singlesnp()` function needs to be performed if a forest plot that compares the MR estimates using the different MR methods against the single SNP tests is desired.

To identify if a single SNP is driving the association a leave-one-out analysis can be performed, where the MR is performed leaving out each SNP in turn. To do this use the `mr_leaveoneout()` function:

```
#res_loo <- mr_leaveoneout(dat)
```

The default method is IVW, but the method argument can be used to change this.

*A more thorough review of this step can be found here.* (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017; Hermani, n.d.e)

**Visualizing the results:** *A more thorough review of this step and the different plots can be found here.* (Hemani et al. 2018; Hemani, Tilling, and Davey Smith 2017; Hermani, n.d.e)

**Scatter plot** A scatter plot that depicts the relationship of the SNP effects on the exposure against the SNP effects on the outcome can be made using the `mr_scatter_plot()` function:

```
#library(TwoSampleMR)
#library(ggplot2)
#res <- mr(dat)
#p1 <- mr_scatter_plot(res, dat)
```

This creates a scatter plot for each exposure-outcome test, stored in `p1` To plot the first plot write:

```
#p1[[1]]
```

To see the number of plots stored in p1 use the `length()` function:

```
#length(p1)
```

For each method used in `mr(dat)` a line is drawn corresponding to the estimated causal effect. To limit the number of lines, the desired methods can be defined in the `mr(dat)` function as shown above.

The plot can be saved as either a pdf or png using `ggsave()`:

```
#ggsave(p1[[1]], file = "filename.pdf", width = 8, height = 8)
#OR
#ggsave(p1[[1]], file = "filename.png", width = 8, height = 8)
```

**Forest Plot** To create a forest plot that compares the MR estimates using the different MR methods against the single SNP tests, the `mr_forest_plot()` function can be used:

```
#p2 <- mr_forest_plot(res_single)
#p2[[1]]
```

The plot shows the causal effects as estimated using each of the SNPs on their own and compared against the causal effect as estimated using the methods that use alle the SNPs. If different methods are desired in the plot, they should be specified in the `mr_singlesnp()` function.

**Leave-one-out plot** The `mr_leaveoneout_plot()` can be used to visualize the leave-one-out analysis:

```
#p3 <- mr_leaveoneout_plot(res_loo)
#p3[[1]]
```

If a specific MR analysis is desired it can be specified in the `mr_leaveoneout()` function using `method = "name of MR analysis"`

**Funnel plot:** A funnel plot can be made using the single SNP results from the `mr_singlesnp()` function by using the `mr_funnel_plot()` function:

```
#p4 <- mr_funnel_plot(res_single)
#p4[[1]]
```

**1-to-many forest plot:** A 1-to-many MR analysis investigates the effect of a single exposure on multiple outcomes or multiple exposures on a single outcome.

**The full code/the code in one piece:**

```
#library(TwoSampleMR)

#Acquiring exposure data:

#library(MRInstruments)
#data(gwas_catalog)
#head(gwas_catalog)

#exposure_phenotype_gwas <-
#   subset(gwas_catalog,
#          grepl("Authorname", Author) &
#            Phenotype == "exposure_phenotype")
#exposure_phenotype_exp_dat <- format_data(exposure_phenotype_gwas)

#clump_data(exposure_phenotype_exp_dat)

#If necessary:
#Acquiring outcome data:

#ao <- available_outcomes()
#head(ao)
```

```
#exposure_phenotype_exp_dat <- extract_instruments(outcomes = "study ID")
#head(exposure_phenotype_exp_dat)

#ao[grepl("outcome trait", ao$trait), ]

#outcome_phenotype_out_dat <- extract_outcome_data(
#    snps = exposure_phenotype_exp_dat$SNP,
#    outcomes = 'study ID'
#)

#Harmonizing data:

#dat <- harmonise_data (
#  exposure_dat = exposure_phenotype_exp_dat,
#  outcome_dat = outcome_phenotype_out_dat
#)

#If necessary:
# dat <- power_prune(dat, method = 1, dist.outcome = "binary")

#The MR analysis:

#library(ggplot2)
#res <- mr(dat) OR #res <- mr(dat, method_list = c("wanted MR analysis1", "wanted MR analysis2))

#mr_heterogeneity(dat)

#mr_pleiotropy_test(dat)

#mr_singlesnp(dat)

#res_loo <- mr_leaveoneout(dat)

#Scatter plot:
#p1 <- mr_scatter_plot(res, dat)
#p1[[1]]

#Forest plot:
#p2 <- mr_forest_plot(res_single)
#p2[[1]]

#Leave-one-out plot:
#p3 <- mr_leaveoneout_plot(res_loo)
#p3[[1]]

#Funnel plot:
#p4 <- mr_funnel_plot(res_single)
#p4[[1]]

#Save plots using ggdsave()
```

## Bibliography


Boehm, Frederick J., and Xiang Zhou. 2022. "Statistical Methods for Mendelian Randomization in Genome-Wide Association Studies: A Review." *Computational and Structural Biotechnology Journal* 20 (January): 2338. https://doi.org/10.1016/J.CSBJ.2022.05.015.

Bowden, Jack, and Michael V. Holmes. 2019. "Meta-analysis and Mendelian Randomization: A Review." *Research Synthesis Methods* 10 (December): 486–96. https://doi.org/10.1002/JRSM.1346.

Bowden, Jack, George Davey Smith, and Stephen Burgess. 2015. "Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection Through Egger Regression." *International Journal of Epidemiology* 44 (April): 512–25. https://doi.org/10.1093/IJE/DYV080.

Bowden, Jack, George Davey Smith, Philip C. Haycock, and Stephen Burgess. 2016. "Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator." *Genetic Epidemiology* 40 (May): 304. https://doi.org/10.1002/GEPI.21965.

Burgess, Stephen, and Simon G. Thompson. 2017. "Interpreting Findings from Mendelian Randomization Using the MR-Egger Method." *European Journal of Epidemiology* 32 (May): 377–89. https://doi.org/10.1007/S10654-017-0255-X.

Davies, Neil M., Michael V. Holmes, and George Davey Smith. 2018. "Reading Mendelian Randomisation Studies: A Guide, Glossary, and Checklist for Clinicians." *The BMJ* 362: k601. https://doi.org/10.1136/BMJ.K601.

Hartwig, Fernando Pires, George Davey Smith, and Jack Bowden. 2017. "Robust Inference in Summary Data Mendelian Randomization via the Zero Modal Pleiotropy Assumption." *International Journal of Epidemiology* 46 (December): 1985. https://doi.org/10.1093/IJE/DYX102.

Haycock, Philip C., Stephen Burgess, Kaitlin H. Wade, Jack Bowden, Caroline Relton, and George Davey Smith. 2016. "Best (but Oft-Forgotten) Practices: The Design, Analysis, and Interpretation of Mendelian Randomization Studies." *The American Journal of Clinical Nutrition* 103 (April): 965–78. https://doi.org/10.3945/AJCN.115.118216.

Hemani, G., K. Tilling, and G. Davey Smith. 2017. "Orienting the Causal Relationship Between Imprecisely Measured Traits Using GWAS Summary Data." *PLoS Genetics* 13 (11): e1007081. https://doi.org/10.1371/journal.pgen.1007081.

Hemani, G., J. Zheng, B. Elsworth, K. Wade, D. Baird, V. Haberland, C. Laurin, et al. 2018. "The MR-Base Platform Supports Systematic Causal Inference Across the Human Phenome." *eLife* 7: e34408. https://doi.org/10.7554/eLife.34408.

Hermani, Gibran. n.d.a. "TwoSampleMR 0.6.8: Exposure Data." https://mrcieu.github.io/TwoSampleMR/articles/exposure.html.

———. n.d.b. "TwoSampleMR 0.6.8: Harmonize Data." https://mrcieu.github.io/TwoSampleMR/articles/harmonise.html.

———. n.d.c. "TwoSampleMR 0.6.8: Introduction." https://mrcieu.github.io/TwoSampleMR/articles/introduction.html.

———. n.d.d. "TwoSampleMR 0.6.8: Outcome Data." https://mrcieu.github.io/TwoSampleMR/articles/outcome.html.

———. n.d.e. "TwoSampleMR 0.6.8: Perform MR." https://mrcieu.github.io/TwoSampleMR/articles/perform_mr.html.

Sekula, Peggy, Fabiola Del Greco M, Cristian Pattaro, and Anna Köttgen. 2016. "Mendelian Randomization as an Approach to Assess Causality Using Observational Data." *Journal of the American Society of Nephrology* 27 (11): 3253–65. https://doi.org/10.1681/asn.2016010098.

Smith, George Davey, and Gibran Hemani. 2014. "Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies." *Human Molecular Genetics* 23 (September): R89–98. https://doi.org/10.1093/HMG/DDU328.