

Article

Not peer-reviewed version

From GWAS to Causal Inference: A Beginner's Guide to Mendelian Randomization with Code Examples

[Ahmed M Salih](#)^{*}, Roman Roy, Yuhe Wang, Irene Treccani, Andre Altmann, Zahra Raisi-Estabragh, Gloria Menegaz

Posted Date: 9 April 2026

doi: 10.20944/preprints202604.0553.v1

Keywords: Mendelian randomization; guide; GWAS



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

From GWAS to Causal Inference: A Beginner's Guide to Mendelian Randomization with Code Examples

Ahmed M Salih ^{1,2,3,*}, Roman Roy ^{2,4}, Yuhe Wang ^{5,6}, Irene Treccani ⁷, Andre Altmann ⁸, Zahra Raisi-Estabragh ^{2,4} and Gloria Menegaz ⁷

¹ British Heart Foundation (BHF) Leicester Centre of Research Excellence, University of Leicester, UK

² William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, Charterhouse Square, London, UK

³ PRIME Lab, Scientific Research Center, University of Zakho, Kurdistan Region, Iraq

⁴ Barts Heart Centre, St Bartholomew's Hospital, Barts Health NHS Trust, London, EC1A 7BE, UK

⁵ Diabetes Research Centre, University of Leicester, Leicester, UK

⁶ National Institute for Health Research Leicester Biomedical Research Centre, University of Leicester and University Hospitals of Leicester NHS Trust, Leicester, UK

⁷ Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy

⁸ Department of Medical Physics and Biomedical Engineering, University College London, London, UK

* Correspondence: ams133@leicester.ac.uk

Abstract

Background: Mendelian randomization (MR) is a powerful approach for assessing causal relationships between risk factors and health outcomes using genetic variants as instrumental variables (IVs). The increasing availability of large genome-wide association study (GWAS) summary statistics from resources such as UK Biobank, FinnGen, and other population-based cohorts has made MR analyses more accessible than ever. However, many available guidelines and tutorials remain highly technical, requiring advanced knowledge of statistical genetics and R programming. **Objective:** This paper aims to provide a clear, step-by-step guide for conducting MR analyses using GWAS summary statistics, designed specifically for non-technical researchers. **Methods:** We outline a structured workflow covering key stages of MR analysis, including dataset selection, quality control, IVs selection, harmonization, and causal estimation. The workflow integrates online tools for quality control and demonstrates the use of commonly applied R packages such as TwoSampleMR. Each step is illustrated with example code and practical guidance to promote reproducibility. **Results and Conclusion:** The proposed workflow supports the process of conducting MR analyses, bridging the gap between theoretical guidelines and hands-on implementation. By offering an accessible and reproducible framework, this tutorial aims to help applied researchers, clinicians, and early-career scientists confidently perform MR analyses and interpret causal findings using publicly available GWAS summary data.

Keywords: Mendelian randomization; guide; GWAS

1. Introduction

In recent years, the availability of large-scale genome-wide association study (GWAS) summary statistics has transformed the landscape of genetic epidemiology. Large biobanks and consortia such as UK Biobank [1], FinnGen [2], and many national and international cohorts have made publicly accessible summary-level associations for hundreds to thousands of traits and diseases. These resources permit investigators to explore causal relationships between traits in a cost-effective and scalable way, without needing individual-level data.

Mendelian randomization (MR) is a method that leverages genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure (modifiable risk factor) on an outcome, while mitigating confounding and reverse causation [3]. The method is grounded in Mendel's laws of inheritance and IVs theory, and has been widely used in epidemiology and human genetics to interrogate

causal hypotheses. With the proliferation of GWAS summary data, MR has become an increasingly popular tool to translate genomic associations into causal inference in observational research.

Rather than serving as a standalone analytical tool, MR is best viewed as a method for interrogating mechanistic hypotheses and strengthening causal inference when used alongside traditional observational analyses. It enables triangulation of evidence by combining genetic data with epidemiological reasoning. This tutorial aims not only to guide readers through conducting MR analyses but also to support critical appraisal of studies that apply these methods.

Despite its appeal, conducting MR analyses is not trivial in practice. A growing number of papers describe MR approaches, guidelines, and applications, but many of them assume that the reader already possesses a strong statistical background in causal inference and fluency in R or equivalent programming environments. Several authoritative resources provide comprehensive guidance on Mendelian randomization. For example, Burgess et al. [4] outline best practices for conducting MR analyses, while Davies et al. [5] offer an accessible framework for interpreting MR findings in clinical contexts. Likewise, the recent Nature Methods primer [3] summarizes MR concepts and assumptions at a theoretical level. These resources are designed for readers with stronger statistical backgrounds and serve a different purpose from the present tutorial, which focuses on practical, step-by-step implementation for beginners.

In short, while numerous resources exist, there remains a gap: a clear, step-by-step guide that (i) walks readers through exposure and outcome selection, quality control, harmonization, and instrument choice, (ii) shows code in commonly used R packages, and (iii) caters to users without advanced statistical training. Our aim in this manuscript is to fill that gap by providing a hands-on, accessible tutorial for performing MR from start to end. In doing so, we hope to broaden uptake, promote reproducibility, and reduce barriers for applied researchers interested in harnessing GWAS summary data for causal inference.

2. Mendelian Randomization

MR is an analytical approach used to strengthen causal inference in observational research. In observational settings, conventional association analyses often struggle to distinguish correlation from causation because both the exposure and the outcome may be influenced by shared environmental, behavioural, or biological factors. To overcome this limitation, causal inference frameworks rely on IVs that influence the exposure of interest but are not affected by confounders and do not directly affect the outcome through any alternative pathways. Genetic variants provide a natural source of such instruments.

The rationale for MR is based on Mendel's laws of inheritance. Genetic variants are randomly allocated at conception and remain fixed throughout life. This random allocation mimics key characteristics of a randomized controlled trial because genotype is assigned prior to disease onset and is largely unaffected by external confounding factors. As a result, MR can reduce bias from confounding and reverse causation [4]. MR differs fundamentally from classical observational studies. Traditional analyses estimate associations between exposures and outcomes but cannot reliably infer causality, particularly when unmeasured or unknown confounders exist. In contrast, MR leverages genetic proxies for exposures. For example, variants that increase or decrease circulating cholesterol levels to approximate a causal contrast under a set of core assumptions.

Three assumptions underpin the validity of MR. First, the genetic variant (instrument) must be robustly associated with the exposure (*relevance*). Second, it must be independent of confounders of the exposure-outcome relationship (*independence*). Third, it must influence the outcome only through the exposure and not via any alternative biological pathways (*exclusion restriction*) [5]. In contemporary practice, multiple independent single-nucleotide polymorphisms (SNPs) identified through GWAS are combined to strengthen the instrument and increase statistical power. When these assumptions are reasonably satisfied, MR provides a powerful and accessible tool for evaluating potential causal relationships in epidemiology and clinical research.

MR can be implemented using either a one-sample or a two-sample design. In one-sample MR, the genetic variants, exposure, and outcome are all measured in the same group of individuals. This allows flexible modelling but can introduce bias if weak instruments are present, because the genetic associations with the exposure and the outcome are estimated in the same sample. In contrast, two-sample MR uses summary statistics from two independent GWAS: one for the exposure and one for the outcome. This design avoids weak instrument bias related to sample overlap, is often more powerful due to larger sample sizes, and is the most commonly used approach in modern MR applications. However, it relies on the assumption that both GWAS are drawn from comparable populations to avoid bias introduced by differences in ancestry or linkage disequilibrium structure.

3. MR Assumption Examples

To provide additional intuition for readers who may be unfamiliar with causal inference, we give simple examples illustrating when each of the three core MR assumptions is satisfied (valid examples) and when it is violated (invalid examples):

1. Relevance: The instrument must be associated with the exposure.
 - **Valid example:** Genetic variants in the *FTO* locus are strongly associated with body mass index (BMI), making them suitable instruments for studying the causal effect of BMI on health outcomes.
 - **Invalid example:** A genetic variant that shows no measurable association with the exposure (e.g., a SNP not associated with cholesterol levels) would violate the relevance assumption because it provides no useful variation in the exposure.
2. Independence: The instrument must not be associated with confounders.
 - **Valid example:** A SNP that influences circulating LDL cholesterol but is not associated with socioeconomic status, diet, smoking, or other lifestyle factors satisfies the independence assumption.
 - **Invalid example:** A SNP whose allele frequency differs systematically across ancestral groups, where ancestry is also related to the outcome, violates independence due to population stratification acting as a confounder.
3. Exclusion restriction: The instrument must affect the outcome only through the exposure.
 - **Valid example:** A variant in the *HMGCR* gene affects cardiovascular disease risk solely through its impact on LDL cholesterol levels, consistent with exclusion restriction.
 - **Invalid example:** A pleiotropic SNP that influences both BMI and blood pressure through separate biological pathways violates the exclusion restriction assumption because it affects the outcome through mechanisms other than the exposure.

4. Workflow

The MR workflow consists of several key stages, from selecting appropriate GWAS datasets to performing causal inference. The overall process is summarized in Figure 1. The first step involves selecting GWAS summary statistics for the exposure and the outcome, ensuring that both datasets originate from similar populations to minimize bias. Depending on the presence or absence of sample overlap, either a one-sample or two-sample MR design is applied. IVs are then identified from the exposure GWAS using genome-wide significance thresholds and linkage disequilibrium (LD) pruning. The selected IVs are matched across the exposure and outcome datasets, and proxy variants are identified if necessary. Next, harmonization is performed to align effect alleles between datasets, followed by the main MR analysis using appropriate statistical methods and R packages. Finally, causal estimates are interpreted alongside sensitivity analyses to assess the validity and robustness of the findings.

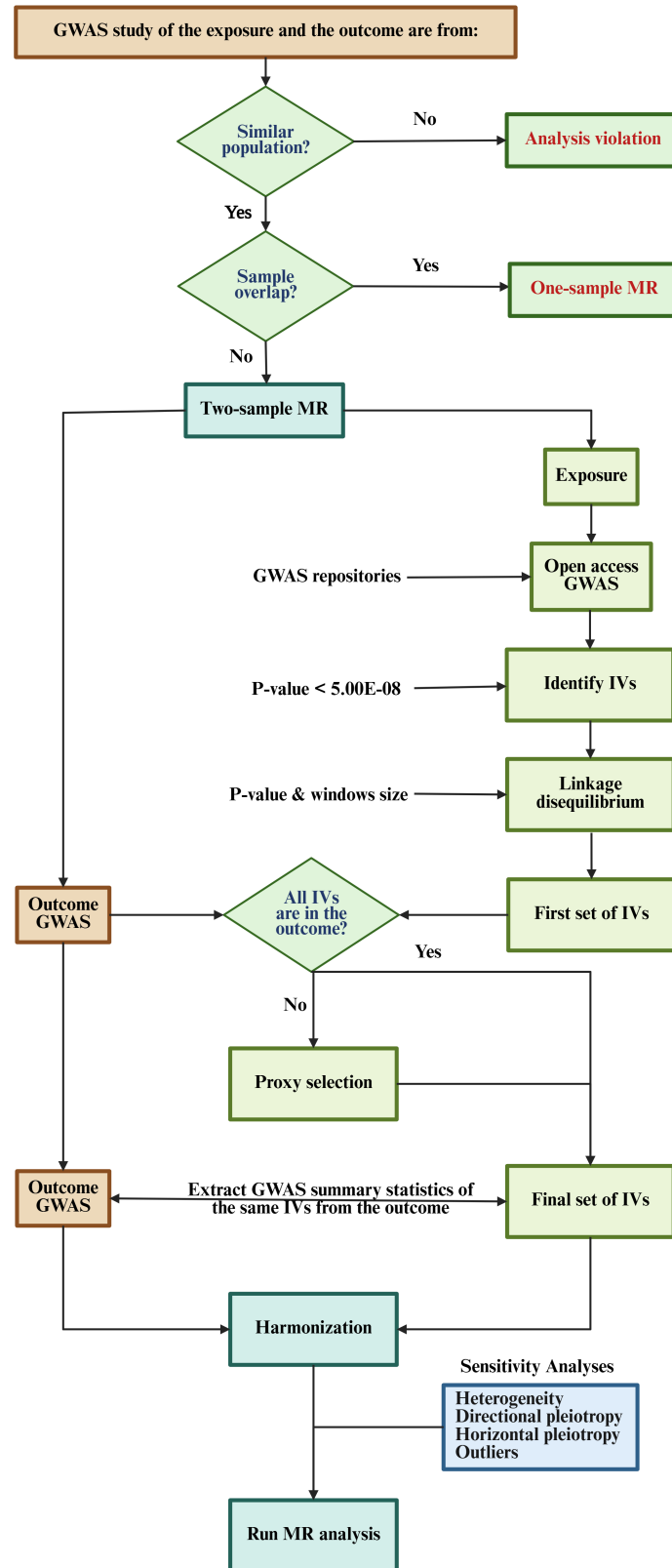


Figure 1. Main steps of Mendelian Randomization workflow

4.1. Exposure and Outcome Population

Using populations with comparable ancestral backgrounds in two-sample MR is a critical requirement, often referred to as the “same-population assumption,” for obtaining unbiased causal estimates. When genetic associations with the exposure and outcome are drawn from populations with differing

ancestries or structures, the core assumptions of MR can be violated, leading to unreliable results. The primary concern with population mismatch is the difference in LD patterns between the exposure and outcome samples [6]. LD refers to the non-random association of alleles at different loci. When a SNP is selected as an instrument based on its strong association with the exposure in one population, this association relies on the local LD structure.

If this structure differs in the outcome population (e.g., due to different ancestral backgrounds), the instrument SNP may not be as strongly associated with the causal variant in the second sample. This violation can affect the validity of the IV-exposure association in the outcome sample, biasing the causal estimate.

Population differences can also lead to the violation of the MR assumption that the genetic instrument is independent of unmeasured confounders (IV independence). This occurs through population stratification, where differences in allele frequencies and phenotype distributions across sub-populations (due to non-random mating or ancestral structure) create a correlation between the genetic variant and the outcome that is not mediated by the exposure [7]. Although standard GWAS attempts to control for this using principal components, it may not always be sufficient, especially with disparate populations.

Empirical evidence and theoretical work show that a violation of the same-population assumption in two-sample MR generally leads to a bias in the causal estimate towards the null (zero). This is also known as attenuation bias. This bias is directly linked to the degree of genetic distance between the study populations, with greater distance resulting in more shrinkage of the effect estimate. While this conservative bias does not typically increase the rate of false positives (i.e., finding a causal effect where none exists), it can severely reduce the statistical power to detect a true causal effect [5].

In summary, while leveraging large, genetically diverse datasets is valuable, researchers conducting two-sample MR must be mindful that using samples from distinct populations risks introducing bias through differences in LD and population stratification, often resulting in an attenuated causal estimate.

A related consideration is the use of multi-ancestry GWAS. Because linkage disequilibrium patterns and allele frequencies differ across ancestral groups, instruments identified in one population may not tag the same causal variants in another. This can violate the same-population assumption and attenuate causal estimates. For this reason, two-sample MR typically requires exposure and outcome GWAS derived from broadly similar ancestries.

4.2. Exposure and Outcome Selection

The first step in conducting an MR analysis is to identify the exposure of interest. Once the exposure is defined, it is essential to determine whether GWAS summary statistics for that trait are publicly available through GWAS repositories. These repositories vary in scope and focus. Some specialize in specific phenotypes or organ systems, while others aggregate results from multiple studies covering a wide range of traits and health outcomes. Additionally, certain repositories are dedicated to particular cohorts or ethnic groups, allowing for population-specific analyses. Table 1 summarizes major repositories that provide access to GWAS summary statistics across different populations, phenotypes, and study cohorts. In some cases, these summary data can be obtained directly from the corresponding authors of the original GWAS publications.

Table 1. GWAS repositories. GWAS: Genome-wide association study; MRI: Magnetic resonance imaging; IDPs: Image derived phenotypes; CVDs: Cardiovascular diseases; CMR: Cardiovascular Magnetic Resonance; UKBB: United Kingdom Biobank

Repository	Phenotypes	Ethnicity	Link
China Kadoorie Biobank PheWeb	Multi-phenotype	Chinese	https://pheweb.ckbiobank.org/ CKB

Table 1. Cont.

Repository	Phenotypes	Ethnicity	Link
BioBank Japan PheWeb	Multi-phenotype	Japanese	https://pheweb.jp/ BJP
GWAS Catalog	Multi-phenotype	Multi-ethnicity from different studies	https://www.ebi.ac.uk/gwas/ GC
GWAS Atlas	Multi-phenotype	Multi-ethnicity from different studies	https://atlas.ctglab.nl/ GA
GWAS UKBB (nealelab)	Multi-phenotype	White European	https://www.nealelab.is/uk-biobank-GUKBB
FinnGen	Multi-phenotype	White European	https://www.finnngen.fi/en/access_results FinnGen
OpenGWAS	Multi-phenotype	Multi-ethnicity from different studies	https://opengwas.io/ OG
Broad Institute Cardiovascular Disease Knowledge Portal	CVDs and CMR IDPs	Multi-ethnicity from different studies	https://hugeamp.org/downloads.html BICDKP
Michigan Genomics Initiative PheWeb	Multi-phenotype	Multi-ethnicity	https://pheweb.org/MGI/ MGIP
NIAGADS Data Sharing Service	Multi-phenotype	Multi-ethnicity from different studies	https://dss.niagads.org/ NIAGADS
UKBB Brain MRI GWAS	Brain MRI IDPs	White European	https://open.win.ox.ac.uk/ukbiobank/big40/pheweb33k/UKBBMG
UKBB Plasma Proteomic GWAS	2,923 proteins	White European	https://metabolomips.org/ukbbpgwas/ UKBPPG
Genetic factors for osteoporosis consortium	Osteoporosis	Multi-ethnicity from different studies	http://www.gefos.org/ GEFOS
Early Growth Genetics Consortium	Early growth factors	Multi-ethnicity from different studies	https://egg-consortium.org/ GCEG
UKB imputed data	Multi-phenotype	White European	https://yanglab.westlake.edu.cn/data/ukb_fastgwa/imp/ fastGWA
Centre For Cancer Genetic Epidemiology	Cancer	Multi-ethnicity from different studies	https://www.ccge.medschl.cam.ac.uk/breast-cancer-association-consortium-bcac/data-data-access/summary-results CCGE
National Cancer Institute (NIH)	Cancer	Multi-ethnicity	https://exploregwas.cancer.gov/#/downloads NCI
Genetic Investigation of ANthropometric Traits	Anthropometric Traits	Multi-ethnicity from different studies	https://giant-consortium.web.broadinstitute.org/index.php/GIANT_consortium_data_files GIANT
Common Metabolic Diseases	Metabolic Diseases	Multi-ethnicity from different studies	https://hugeamp.org/datasets.html CMD
Disease-centric Mendelian randomization database	Multi-phenotype	Multi-ethnicity from different studies	http://www.inbirg.com/DMRdb/#/base DMRdb
University of Bristol	Multi-phenotype	Multi-ethnicity from different studies	https://data.bris.ac.uk/data/dataset?level=top UoB

Table 1. Cont.

Repository	Phenotypes	Ethnicity	Link
Dryad	Multi-phenotype	Multi-ethnicity from different studies	https://datadryad.org/ DRYAD
The Global Lipids Genetics Consortium	Lipid traits	Multi-ethnicity from different studies	https://csg.sph.umich.edu/willer/public/glgc-lipids2021/ GLGC

At this stage, it is essential to carefully examine the available repositories, identify the GWAS summary statistics corresponding to the exposure of interest, and review key study characteristics such as the number of genome-wide significant associations, total sample size, study cohort, and the ethnicity of participants. The same process is repeated for the outcome of interest, ensuring high quality GWAS summary statistics are available.

GWAS summary statistics may be provided in various formats; however, the core elements of these files are generally consistent. Each dataset should include the SNP identifier (usually beginning with the prefix *rs*), chromosome number, genomic position, effect size (e.g., beta coefficient or odds ratio), standard error, effect and alternative alleles, effect allele frequency, and the associated *p*-value, which indicates the statistical significance of the variant based on standard GWAS thresholds.

5. Illustrative Case Study

For this tutorial, we will use birth weight [8] as the exposure of interest and the outcome is left ventricular mass [9], the left heart chamber's muscle. After identifying the exposure of interest with a sufficient number of significant associations and confirming the public availability of the corresponding GWAS summary statistics, we demonstrate how to load these data into the *R* environment and conduct MR analysis using the <https://mrcieu.github.io/TwoSampleMR/index.html> *TwoSampleMR* package [10]. This package is among the most widely used tools for MR analyses and supports multiple analytical approaches. Figure 2 illustrates that *TwoSampleMR* has received the highest number of GitHub stars compared to other MR packages, reflecting its popularity and utility within the research community. Detailed installation instructions are available <https://mrcieu.github.io/TwoSampleMR/index.html> here, depending on the operating system used.

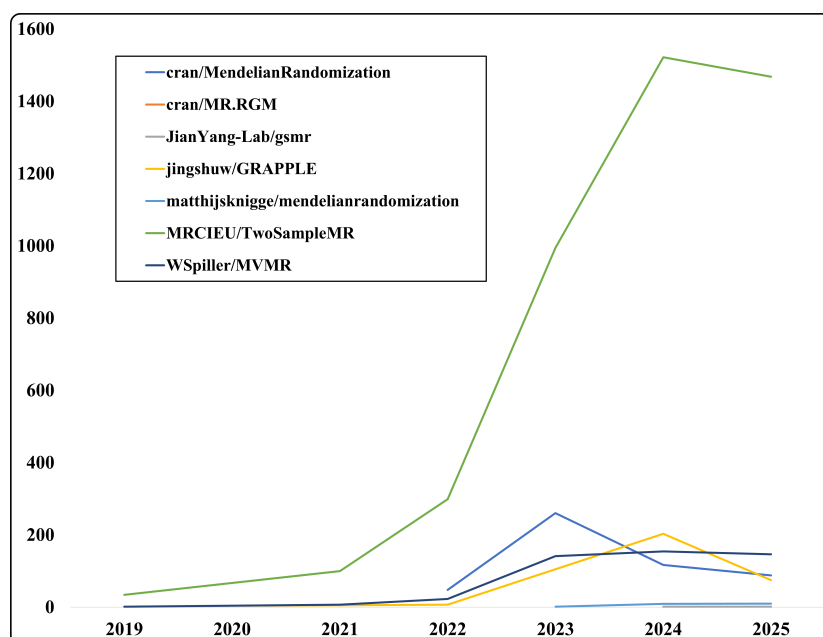


Figure 2. Star history of seven R packages that perform MR analysis for the past six years generated using data from star-history.com

To maintain readability, only key illustrative code snippets are shown in the main text (labeled as “**Code X**”). Full, executable scripts are provided in the Supplementary Material, where all listings are labeled with the prefix “**Supp Code X**” to clearly distinguish them. Numbering of code listings continues seamlessly from the main text into the Supplementary Files, ensuring unique references and preventing hyperlink conflicts.

5.1. Power Calculation

Power calculation is used to determine the sample size needed to reliably detect true effect between an exposure and an outcome. It helps to answer the research questions and decide whether to accept or reject the null hypothesis [11]. There have been many proposed methods to conduct power calculation for two-sample MR. The online tool <https://shiny.cnsgenomics.com/mRnd/> mRnd [12] is a user-friendly power and sample-size calculator tailored for MR studies. The interface presents clear fields for input (e.g. R^2 , outcome type, effect size, alpha) and lets non-technical users easily switch between continuous and binary outcomes. It guides users through the essential parameters without needing advanced statistical software.

In two-sample MR, statistical power is determined primarily by the sample sizes of both the exposure and outcome GWAS, because the precision of the SNP-exposure and SNP-outcome associations jointly influences the detectable causal effect. Importantly, when using publicly available summary statistics, these sample sizes are fixed and cannot be modified by the analyst; power estimation therefore serves mainly to assess feasibility and interpretability rather than to guide design choices.

Note 1

Although MR has no universal power threshold, very low estimated power (e.g., below 50%) suggests limited ability to detect even moderate causal effects.

5.2. Reading GWAS File (Exposure)

First we have to load some libraries that we will need to run the analysis. Supp Code S8 and Code S1 shows the required library and the code to read a GWAS summary statistics file in R. The code shows also the columns of the file include chromosome number, genome position, SNP ID, effect allele, other allele, effect allele frequency, effect size represented as beta value, standard error, the p-value which indicates whether it is significant or not and the final column shows the number of samples. Regardless of the software or model used to run the analysis, the aforementioned columns are the same. Number of the rows (SNPs) in the file is 1,624,5523 indicates the number of the genetic variants used and reported in the GWAS. The next step is to read the GWAS summary statistics into the *TwoSampleMR* package as it is shown in Code S1.

Listing 1: Read the GWAS file into the TwoSampleMR package

```
1 library(markdown)
2 library(Cairo)
3 library(TwoSampleMR)
4 exposure_data <- read_exposure_data(
5 filename = 'BW3_EUR_summary_stats.txt.gz',
6 sep = '\t', snp_col = 'rsid', chr_col = 'chr',
7 beta_col = 'beta', pos_col = 'pos.b37.',
8 se_col = 'se', effect_allele_col = 'EA',
9 other_allele_col = 'OA', pval_col = 'p',
10 eaf_col = 'eaf', samplesize_col = 'n')
11 # add a phenotype column
12 exposure_data$exposure = 'Birth weight'
```

The function https://mrcieu.github.io/TwoSampleMR/reference/read_exposure_data.html `read_exposure_data()` involves several parameters that need to be specified including file name (with the path) of the GWAS summary statistics and within the file the columns of SNP, beta value, SE

value, effect allele, alternative allele, p-value and effect allele frequency. Some of these parameters are mandatory to provide because they are necessary for the analysis including the SNPs ID, beta and standard error value, effect and alternative allele, effect allele frequency and p-value.

5.3. Instrumental Variables Selection

After reading the GWAS summary statistics into the R package, the IVs need to be selected. The IVs in biological science and medicine are usually SNPs. Valid IVs based on the MR core assumption are those significantly associated with the exposure based on the standard GWAS p-value threshold (p-value < 5.00E-08). While the conventional GWAS significance threshold is p-value < 5.00E-08, it may occasionally be appropriate to use a slightly less stringent threshold (e.g. p-value < 5.00E-05), particularly when the exposure has few strongly associated variants. Such decisions should be justified carefully, as relaxing the threshold increases the risk of including weaker or potentially pleiotropic instruments.

To select the significant SNPs, Code S2 shows the command might be applied:

Listing 2: Select significant IVs based on the standard GWAS p-value threshold

```
1 sig_SNPs <- exposure_data[exposure_data$pval.exposure < 5e-8, ]
```

The number of IVs is 2,278 which means this is the number of the SNPs that are significantly associated with the exposure. Although these SNPs pass the GWAS significance threshold, they are not necessarily independent signals, as many may be in LD with one another. Therefore, LD clumping is required to ensure that only independent instruments are taken forward for MR analysis.

5.4. Linkage Disequilibrium Clumping

LD clumping is a process used to select independent genetic variants (SNPs) that are not strongly correlated with each other. In this step, one SNP from a group of correlated SNPs (in high LD) is selected, usually the one most strongly associated with the exposure. This ensures that the selected SNPs represent independent genetic signals. LD clumping is important before MR because using correlated SNPs can bias causal estimates and inflate statistical significance. By keeping only independent SNPs, the analysis provides more reliable and interpretable results.

The function https://mrcieu.github.io/TwoSampleMR/reference/clump_data.html `clump_data` is the part of the *TwoSampleMR* package which run LD clumping. The clumping function includes default values for the window size and r2 threshold, which reflect commonly used settings in GWAS analyses. These defaults are designed to balance the removal of correlated variants with the retention of true independent signals, and they provide a reasonable starting point for most MR applications unless there is a specific rationale for choosing stricter or more permissive parameters. Code S3 show how to run LD clumping locally (online in Supp Code S10) locally. After running the code (LD clumping), 61 independent IVs are left to take it forward for the MR analysis.

Listing 3: Local Linkage disequilibrium clumping

```
1 # LD clumping locally using default parameters
2 clumped <- clump_data(
3   dat = sig_SNPs,
4   clump_r2 = 0.001,
5   pop = 'EUR',
6   bfile = '/Documents/Genome_ref/EUR',
7   plink_bin = '/R/x86_64-pc-linux-gnu-library/4.3/plinkbinr/bin/plink_Linux')
```

Another way to perform LD clumping when the number of the IVs is not very big is manually using either <https://ldlink.nih.gov/?tab=ldmatrix> LDmatrix or https://www.ensembl.org/Homo_sapiens/Tools/LDLD_ensembl.

Another way to measure LD is through pruning which is similar to clumping. The difference is that pruning does not consider the p-value of SNPs in the GWAS summary statistics when selecting

uncorrelated IVs. Accordingly, for MR analysis clumping is preferred because it keeps the most significant associations (the smallest p-value) [13].

5.5. Instrument Strength

The F-statistic in MR quantifies instrument strength, measuring the association between genetic variants and the exposure. It is used to detect weak instrument bias, which can cause misleading causal estimates [14]. An F-statistic below 10 suggests problematic weakness.

It is calculated primarily via two methods: 1) using the Wald statistic, $F = (\beta/SE)^2$, derived directly from summary data and 2) using the coefficient of determination [15], where R^2 is first calculated from the beta and allele frequency $R^2 = 2\beta^2 * EAF * (1-EAF)$, then $F = R^2 * (N-2) / (1 - R^2)$. Both methods are standard in genetic epidemiology for assessing instrument validity.

The Supp Code S11 shows how to calculate F-statistics using the aforementioned methods. Supp Table S1 shows the F-statistics values for the used IVs using the R^2 method. Beside that, the code will print more details of the instrument strength assessment. Since all the IVs are strong based on the F-statistics (min value is 29.9 and max value is 179.8), it is time to read the outcome and prepare it for the MR analysis.

Note 2

F-statistics should not be used to decide which IVs to exclude; an F-statistic below 10 does not mean a variant must be dropped, but signals that weak-instrument bias should be acknowledged when interpreting the results.

5.6. Reading GWAS File (Outcome)

Supp Code S12 shows how to read the GWAS of the outcome in our example which is left ventricular mass into the package. The function https://mrcieu.github.io/TwoSampleMR/reference/read_outcome_data.html `read_outcome_data()` is the same function of the exposure with the same parameters. After reading the file, the total number of the SNPs is 6984035. We then need to check how many of the IVs from the exposure GWAS are available in the outcome which is illustrated in Supp Code S13. It shows that 59 out of the 62 SNPs from the exposure are also available in the outcome.

5.7. Proxy Selection

LD proxy selection in MR refers to the process of choosing substitute genetic variants (proxies) when the original SNPs of interest are not available in the outcome dataset. These proxy SNPs are selected based on their high LD with the original variants, meaning they are strongly correlated and likely represent the same genetic signal. Using LD proxies helps maintain the power and accuracy of MR analyses by allowing the use of available data while preserving the validity of instrumental variables. It is important to select proxies with high LD (usually $r^2 > 0.8$) to ensure they closely reflect the effect of the original SNPs. This approach reduces bias and improves the reliability of causal inference in MR studies.

To run it globally using online tools, you can use the tool <https://ldlink.nih.gov/?tab=ldproxy> LDproxy. Similar to LD clumping, the windows size, the population and the r^2 need to be input. In our tutorial, we did not apply LD proxy because the vast majority of the IVs in the GWAS of the exposure are available in the GWAS of the outcome.

5.8. Data Harmonization

Harmonizations means matching the genetic variants (SNPs) found in the exposure and the outcome. It aligns them, which means it makes sure that for each SNP, the effect is measured on the same allele in the exposure and the outcome. Specifically, it makes sure the effect allele (the allele for which the effect size is reported) and the alternative allele are the same in both the exposure and outcome data. If the effect alleles do not match between datasets, the effect sizes could be interpreted

in a wrong way. The harmonisation step implemented in the *TwoSampleMR* package automatically removes strand-ambiguous SNPs (A/T or C/G) with allele frequencies close to 0.5, as their orientation cannot be reliably determined. This prevents misalignment of effect alleles and avoids introducing spurious directionality into the MR estimates. Code S4 shows data harmonization step and how to save the results of the harmonization.

Listing 4: Data harmonization and save the results in a text file

```
1 Harmonization <- harmonise_data(
2 exposure_data, Outcome_data, action=2)
3 mr_report(
4 Harmonized, study = 'xxx xxxxx xxx')
5 save_stat = paste0(
6 current_path, '/stat/',fn, '_stat',
7 extension='.csv')
8 write.csv2(Harmonized, file = save_stat,
9 sep=',', '')
```

5.9. Linear Methods

After the data of the exposure and the outcome are harmonized, the data is ready to run the analysis. Standard practice is to conduct main analysis using inverse variance weighted (IVW) and complement this with additional analyses using MR Egger, weighted mode, weighted median and others. Code S5 shows how to run and save the results of MR analysis.

Listing 5: Run and save the results of the MR analysis

```
1 res <- mr(Harmonized)
2 odd = generate_odds_ratios(res)
3 save_res = paste0(current_path,
4 '/report/',fn, '_repor', extension='.csv')
5 write.csv2(odd, file = save\_res, sep=',', '')
```

By default, the package generates an html file that displays the results of the main method and the complementary analyses which is displayed in the Figure 3. Let us break down the results to understand and interpret them:

1. Main and complementary analyses

- **Inverse variance weighted:** This is the primary estimate. With $pval < 0.05$, it suggests a 1-SD increase in genetically predicted birth weight is associated with a 0.1724-SD increase in left ventricular mass.

Note 3

When interpreting MR results, it is important to check whether the exposure and outcome GWAS report effect sizes in original units or in transformed or standardised scales. If either trait has been transformed (for example, Z-scored or inverse-normalised), the MR estimate represents the effect in standard deviation units of the transformed phenotype rather than in its original biological units. In such cases, the causal effect cannot be expressed in natural units (e.g., each kilogram increase in birth weight is associated with 0.17g higher left ventricular mass in adulthood) because the transformation removes the original scale.

- **Weighted Median:** With $pval < 0.05$, this confirms the significant positive causal effect. This method is robust to up to 50% invalid instruments.
- **MR Egger:** This estimate is non-significant ($pval > 0.05$). Its significance here usually confirms the IVW conclusion unless heterogeneity or pleiotropy is detected.
- **Weighted Mode:** Non-significant ($pval > 0.05$). This method is highly robust to invalid instruments but has very low power. Its significance here usually confirms the IVW conclusion.

Results from two sample MR:

method	nsnp	b	se	pval
MR Egger	57	0.0817664	0.2589725	0.7533995
Weighted median	57	0.1877310	0.0930793	0.0437066
Inverse variance weighted	57	0.1723915	0.0729775	0.0181642
Simple mode	57	-0.0375777	0.2546148	0.8831993
Weighted mode	57	-0.0535535	0.1936909	0.7831900

Heterogeneity tests

method	Q	Q_df	Q_pval
MR Egger	98.92629	55	0.0002585
Inverse variance weighted	99.16588	56	0.0003355

Test for directional horizontal pleiotropy

egger_intercept	se	pval
0.0028811	0.0078941	0.7165341

Test that the exposure is upstream of the outcome

snp_r2.exposure	snp_r2.outcome	correct_causal_direction	steiger_pval
0.0202055	0.0065101	TRUE	0

Figure 3. The results of the MR analysis in the main method and the complementary analyses.**Note 4**

If the IVW estimate is significant but the weighted mode is not, this may indicate that some instruments violate MR assumptions. In such cases, more robust estimators (such as the weighted median or weighted mode) should be prioritised, and the IVW result interpreted cautiously. Additional checks for heterogeneity and pleiotropy are recommended before drawing conclusions.

Figure 4 shows a plot of the MR analysis represented by the regression line of the main and the complementary analyses.

2. Heterogeneity Tests: It measures whether the individual SNP effect estimates are consistent with each other. Significant heterogeneity suggests that the genetic variants are likely affecting the outcome through different pathways (pleiotropy).

- **MR Egger:** With $pval < 0.05$, the result is highly significant. This indicates strong heterogeneity across the 57 instruments.
- **IVW:** Similarly, this is highly significant ($pval < 0.05$), confirming that there is substantial inconsistency in the causal estimates derived from the different SNPs.

The strong and consistent evidence for heterogeneity suggests that many of the instruments likely exhibit horizontal pleiotropy (affecting the outcome through pathways other than the intended exposure).

3. Test for Directional Horizontal Pleiotropy: This test uses the intercept from the MR-Egger regression to determine if there is systematic, directional pleiotropy (where the pleiotropic effects tend to push the outcome in a consistent direction).

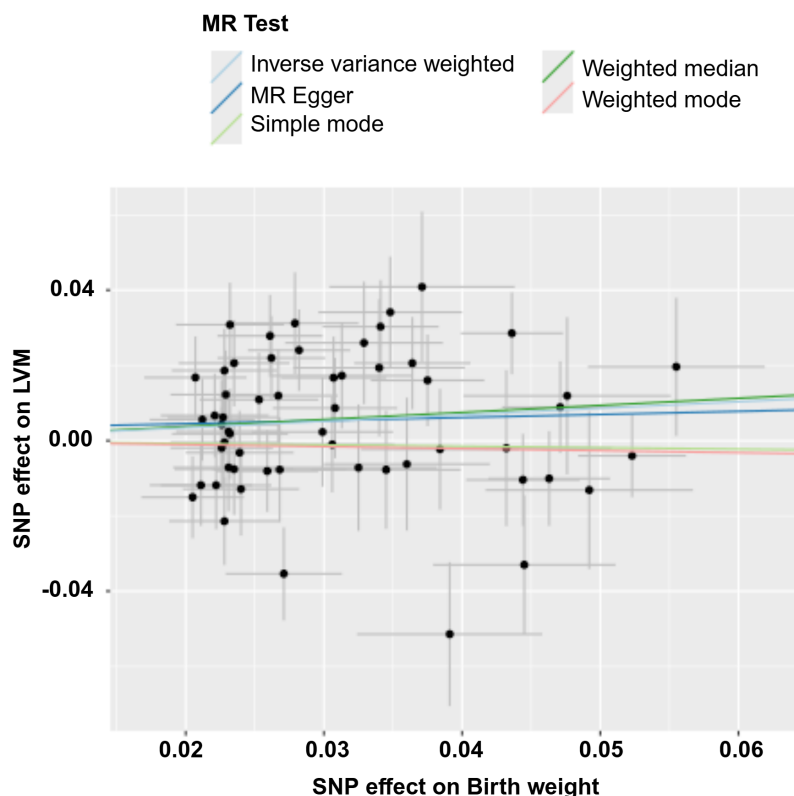


Figure 4. Regression lines of the main and the complementary MR analyses

In our example, the p-value is non-significant, there is no strong evidence for systematic directional horizontal pleiotropy. While individual SNPs may be pleiotropic (as shown by the Heterogeneity tests), their effects appear to be randomly distributed and not biasing the overall result in a single direction.

4. Test that the Exposure is Upstream of the Outcome: The Steiger test checks whether the variance explained in the exposure is greater than the variance explained in the outcome, ensuring the presumed causal direction (exposure \rightarrow outcome) is correct.

In our analysis, the test returned TRUE, indicating that the instruments indeed account for substantially more variance in the exposure. This supports the biological plausibility of the assumed direction and makes reverse causation unlikely.

Note 5

1- Significant pleiotropy does not necessarily invalidate an MR analysis, but it reduces confidence in the IVW estimate and requires greater reliance on pleiotropy-robust methods and cautious interpretation.

2- Where bidirectional effects are biologically plausible, this can be examined using bidirectional MR by reversing the exposure and outcome roles and testing the reciprocal causal direction.

5. Test for Pleiotropy Residual Sum and Outlier: The *TwoSampleMR* includes another significant package which is called Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO) [16]. It evaluates the presence of horizontal pleiotropy in multi-instrument to detect and correct accordingly. The detection is implemented through the *global test* while the correction is conducted using the *outlier test*. Finally, in case it detected and corrected horizontal pleiotropy, then it runs another test which is called *distortion test* to show the distortion in the association before and after outlier removal. Code S6 shows how to run the MR-PRESSO analysis.

Listing 6: Run and save MR_PRESSO analysis in a text file

```
1 MR_PRESSO_run = run_mr_presso(
```

```

2 Harmonized, NbDistribution = 400,
3 SignifThreshold = 0.05)
4 MR_PRESSO = paste0(
5 current_path, '/MR_PRESSO/', fn, '_MR_PRESSO',
6 extension = '.csv')}\
7 capture.output(MR_PRESSO_run, file= MR_PRESSO)

```

The global test pval is < 0.0025 which indicates detection of outliers. The beta value and the p-value before the correction are 0.172 and 0.021. After the correction, the beta value is 0.191 and the pval is 0.0088. The p-value of the distortion is 0.80 which means that the association did not change.

Conclusion: The MR results provide compelling and robust evidence for a causal relationship between the exposure and the outcome. While initial analyses revealed significant heterogeneity (Cochran's Q pval < 0.05), subsequent sensitivity testing through MR-Egger Intercept and MR-PRESSO confirmed that this heterogeneity did not introduce systematic bias. Specifically, the MR-PRESSO outlier-corrected estimate (Beta = 0.191, pval = 0.0088) reinforced the primary IVW finding, while the Steiger Directionality Test (Result: TRUE) successfully ruled out reverse causation. The consistency across the Weighted Median and the corrected IVW models paired with the absence of directional pleiotropy supports the conclusion that the exposure is a significant causal driver of the outcome. These findings suggest that targeting exposure could be an effective strategy for modifying the outcome levels.

6. Possible Overlap

Conventional two-sample MR packages such as *TwoSampleMR* do not correct for weak-instrument bias or winner's curse, both of which tend to become more pronounced when the exposure and outcome GWAS samples overlap, and can distort causal effect estimates [17]. The MRlap package was developed to address these limitations, providing two-sample MR estimates that remain valid even when exposure and outcome GWAS share participants [18]. MRlap combines standard IVW MR with a cross-trait LD score regression (LDSC)-based correction to quantify and adjust for overlap-related bias. The package requires GWAS summary statistics containing SNP identifiers, alleles, sample size, and Z-statistics (or effect size and standard error). Instrument selection is performed using distance and LD-based pruning, rather than clumping as in *TwoSampleMR*, and requires input files for LD-score regression (ld and hm3, available online <https://github.com/n-mounier/MRlap> MRlap).

Note 6

MRlap requires two external resources: LD score files, which contain pre-computed linkage disequilibrium metrics used by LD score regression, and the HapMap 3 (HM3) SNP reference panel, which provides a consistent set of well-imputed common variants for stable LD estimation. These are standard resources in genetic epidemiology and are publicly available.

Code S7 illustrates the basic syntax for running MRlap, in which key parameters MR_threshold (p-value threshold for SNP selection), MR_pruning_distance (genomic distance window for pruning) and MR_pruning_LD (LD r2 threshold) can be defined.

Listing 7: MRlap analysis with the specified parameters

```

1 mr_lap_results <- MRlap(
2 exposure = exposure_file,
3 exposure_name = 'Birthweight',
4 outcome = outcome_file,
5 outcome_name = 'LV_Mass', ld = ld_path,
6 hm3 = hm3_path, MR_threshold = 5e-8,
7 MR_pruning_LD = 0.1, MR_pruning_dist = 10000)

```

MRlap outputs the observed and bias-corrected causal effect estimates with standard error, as well as a test statistic and p-value assessing whether overlap correction significantly alters the estimate.

As it is explained in Figure 5 which lists the results of MRlap analysis, the bias-corrected point estimate was greater than the observed estimate, and the difference was statistically significant ($p=0.020$). These results suggest that the observed association was attenuated by bias, with correction using MRlap increasing the estimate magnitude.

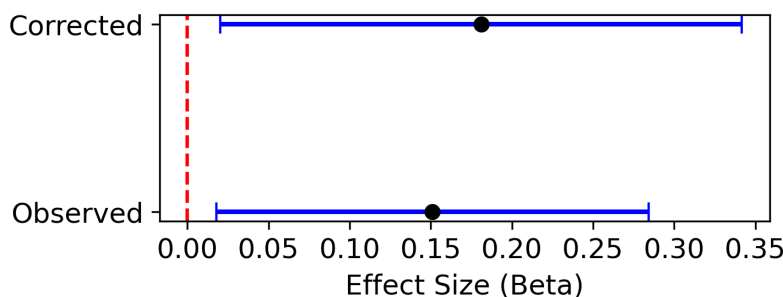


Figure 5. Observed and corrected effect size using MRlap

MRlap also outputs LDSC estimates of SNP-heritability for the exposure and outcome. In our case, h^2 was 0.103 for birth weight and 0.233 for LVM, indicating that common SNPs captured in the GWAS (or tagged through LD) explained 10% and 23% of variance in these traits respectively. It should be noted that h^2 reflects variance explained by common additive variants measured in the GWAS, and is therefore typically lower than total heritability which may include contributions from rarer variants and non-additive genetic effects [19].

7. Multiple Variables MR

Multivariable MR (MVMR) is most appropriate when two or more exposures are genetically correlated or may each plausibly influence the outcome, making univariable MR insufficient to disentangle their independent effects. It is particularly useful when exposures lie on the same biological pathway, when one exposure may mediate the effect of another, or when omitting a relevant exposure would introduce pleiotropy. In practice, MVMR should be considered whenever multiple traits are suspected to contribute jointly to the outcome and valid genetic instruments exist for each exposure.

In this tutorial, we considered the meta-analysis of gestational duration [20] as another variable of interest to explore its genetic effect on LVM alongside birth weight. In addition, our selection of gestational duration is that it is highly associated with birth weight because those with shorter gestational duration are usually with lower birth weight. The GWAS summary statistics can be downloaded from <https://egg-consortium.org/GCEG> (check Table 1). The instructions to run the analysis using the function `mv_multiple()` function within the *TwoSampleMR* package is illustrated https://mrcieu.github.io/TwoSampleMR/articles/perform_mr.html#multivariable-mr here.

First, we need to extract the significant IVs from all exposures and combine them into one dataframe. Thereafter, apply `ld_clumping` to remove those in LD. Combining them in one dataframe before applying `ld_clumping` ensures to exclude those in LD across the exposures. Then, re-extract the summary statistics from the GWAS using the final list of IVs after clumping. Ensure to harmonize the IVs of all exposure to ensure they are on the same allele. The final step is to run the analysis using the `mv_multiple()` function. The results will be presented for each exposure toward the outcome as it is shown in Table 2. Full code is provided in the Supp Code S14.

Table 2. Multivariable Mendelian Randomization analysis

Exposure	Outcome	nsnps	Beta	SE	P_value
Gestational duration	LVM	2	0.0983	0.211	6.55E-01
Birth weight	LVM	44	0.176	0.083	3.29E-02

The results show the number of the IVs, effect size and the p-value to indicate whether the association is significant or not. The results can be interpreted as a 1-unit increase (e.g., one standard deviation) in birth weight is conditionally associated with a 0.176 unit increase in LVM, even after accounting for the effect of gestational duration. This suggests that the genetic liability for higher birth weight contributes to increased LVM independently of the length of gestation. On the other hand, the result of gestational duration should be interpreted with caution. The number of IVs is 2 for gestational duration is the bare minimum for an MVMR analysis. This low instrument count suggests low statistical power and a very high risk of weak instrument bias. While the result is null, it may be a *false null* due to the limited number of instruments.

8. Causal Analysis Using CAUSE

Conventional two-sample MR approaches such as IVW and MR-Egger can produce biased or overly confident causal inferences when horizontal pleiotropy is present, i.e. when genetic variants influence the outcome through pathways other than the exposure. This is particularly problematic for correlated horizontal pleiotropy, where variants act through an unobserved shared heritable factor that affects both traits. In this setting, SNP-exposure and SNP-outcome effects can be correlated even when the true causal effect is zero.

To address this limitation, we applied the Causal Analysis Using Summary Effect estimates (CAUSE) method [21]. CAUSE explicitly models two competing explanations for the observed genetic association:

1. **Causal model:** the exposure has a direct causal effect on the outcome.
2. **Sharing model:** both traits share common genetic influences, but there is no direct causal relationship.

CAUSE compares how well these models explain the data using the expected log predictive density (ELPD). Strong evidence for causality is present when the causal model fits significantly better than the sharing model ($p < 0.05$).

CAUSE requires harmonised GWAS summary statistics with SNP/variant identifiers, effect/other alleles, effect size and standard error (explained in Supp Code S15). After harmonisation, nuisance parameters (shared effects and pleiotropic effects) are estimated using the function `est_cause_params()`, LD pruning is performed (the authors recommend using PLINK for this step), and the main analysis is run with `cause()`. This fits both sharing and causal models, and returns posterior distributions and a comparison of model fit between the two models. Evidence for causality is strongest where the causal model fits better than the sharing model, signified by $p < 0.05$.

In our analysis, CAUSE did not show evidence that the causal model fit better than the sharing model ($p = 0.26$; Table 3). This indicates that the observed association between genetically predicted birth weight and LVM can be explained equally well by shared genetic influences. This result contrasts with the IVW analysis, which suggested a statistically significant positive association. Although the MR-Egger intercept test did not detect directional pleiotropy, this test only evaluates uncorrelated pleiotropy and does not account for correlated pleiotropy due to shared heritable factors. Taken together, these findings suggest that the IVW association may reflect shared genetic architecture between birth weight and LVM rather than a direct causal effect.

Table 3. Results of the CAUSE package analysis

model1	model2	delta_elpd	se_delta_elpd	p
sharing	causal	-0.852	1.343	0.26

The graphical output of CAUSE (Figure 6) provides further insight. Panel A shows the prior and posterior parameter distributions under the sharing and causal models, illustrating that neither model demonstrates a clear advantage in explaining the data. Panel B displays the contribution of each SNP

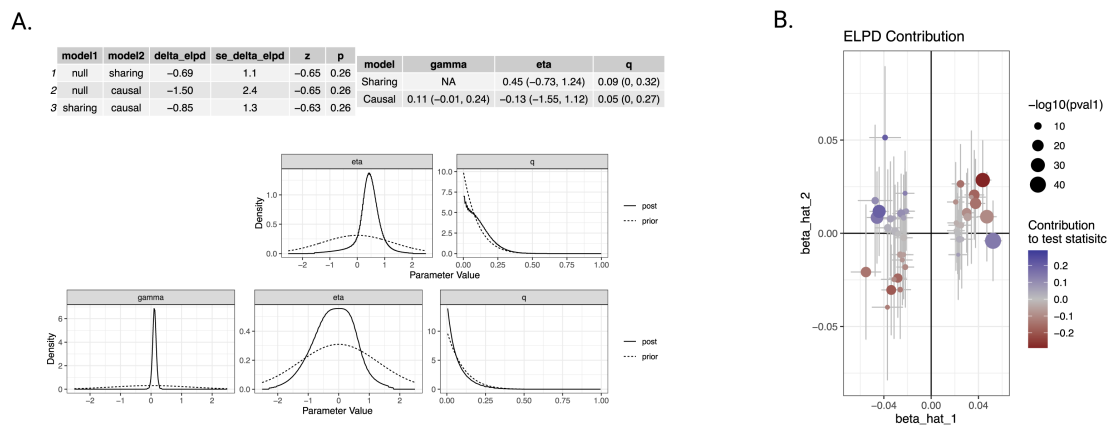


Figure 6. Graphical results of the CAUSE package analysis

to the model comparison statistic. The SNP contributions are distributed in both positive and negative directions, without a consistent pattern favoring the causal model. This visual pattern supports the non-significant statistical comparison and reinforces the interpretation that shared genetic factors may explain the association.

9. Non-Linear Methods

Most MR studies assumed a linear relationship, providing a single “population average” causal effect. However, biological systems are rarely strictly linear; many exposures, such as body mass index or alcohol consumption, exhibit threshold or J-shaped effects. To address this complexity, the R packages *nlnmr* [22] and *SUMnlnmr* were developed to move beyond linearity and map the dose-response curve of causal relationships.

The <https://rdrr.io/github/jrs95/nlnmr/> *nlnmr* package was designed to fill the gap in software capable of detecting non-linear causal associations using individual-level participant data. When a researcher has access to raw, person-level genetic, exposure, and outcome data, *nlnmr* provides a robust framework for estimation. The package primarily employs Piecewise Linear Regression and Fractional Polynomials. Piecewise Linear Regression divides the exposure distribution into strata (e.g., deciles) and calculates the localized average causal effect (LACE) [22] within each group. This allows researchers to see if the causal effect varies, for instance, if an increase in a biomarker is harmful only at high levels but neutral at low levels. The Fractional Polynomials approach fits a flexible mathematical curve to the data, allowing for the detection of non-linear shapes like U-curves or plateaus without the rigidity of a standard polynomial. By providing these tools, *nlnmr* allows scientists to test the “exclusion restriction” and “onotonicity” assumptions in a more granular way, ensuring that the estimated causal effects are not obscured by the oversimplification of a straight line.

While *nlnmr* is powerful, it faces a significant real-world hurdle: the “Individual-Level Data Gap.” In modern epidemiology, high-powered studies require massive sample sizes, often necessitating meta-analyses across multiple international cohorts. Due to strict data privacy regulations (such as GDPR), many institutions cannot share raw, individual-level participant data. This limitation often meant that non-linear MR was restricted to single-center studies or required complex, often impossible, data-sharing agreements.

To overcome these barriers, *SUMnlnmr* was developed as the natural successor and companion to *nlnmr*. The primary innovation of <https://github.com/amymariemason/SUMnlnmr> *SUMnlnmr* is its ability to perform non-linear MR using stratified summary-level data. Instead of requiring the raw dataset, *SUMnlnmr* allows each participating center to perform initial calculations locally. Each site stratifies its data by the “instrument-free” exposure and generates summary statistics (means and

standard errors) for each stratum. These small, non-identifiable summary files are then sent to a central hub where *SUMnlmr* aggregates them.

In our case, we aim to investigate whether and how birth weight influences LVM in a non-linear association. We investigated the potential non-linear causal effect of birth weight on LVM using United Kingdom Biobank data. Participants with available birth weight (Data-Field 20022) and LVM (Data-Field 31063) were included. Quality control procedures were applied to genetic data, excluding individuals and variants with high missingness. After filtering, the final analytical sample comprised 1,621 individuals and 54 SNPs (from an initial 3,295 participants). Although genotype imputation represents a possible alternative, variant exclusion was preferred to reduce potential bias from poorly imputed markers.

Note 7

Genotype imputation is a statistical technique used to infer unobserved genetic variants by using reference panels of densely genotyped individuals; it increases genomic coverage and allows analyses to include variants that were not directly genotyped in the study.

The initial step for this method was to stratify the population on the “residual exposure”, that is the residual values from regression of the exposure on the genetic variants (see Supp Code: S16). The exposure should be continuous and not rounded or coarsened into groups, as this may induce bias in the stratification process. The method requires access to individual-level data or stratified summarized data. The impact of the instrument on the exposure is assumed to be homogeneous at different values of the exposure. These stratified summarized data (Supplementary Table S2) can be obtained from individual-level data using the *create_summary_data()* function in the *SUMnlmr* package. The method divides the population by default into 10 groups (deciles). For each group, it estimates how strongly the genetic instrument influences birth weight (bx) and how strongly it influences LVM (by). In addition, it calculates the corresponding standard errors (bxse and byse), as well as the mean exposure level (xmean) and the minimum and maximum exposure values within each group (xmin and xmax). Confidence intervals were obtained using non-parametric bootstrap resampling to account for uncertainty in stratum-specific ratio estimates and to provide robust inference for the overall non-linear exposure–outcome curve. The implementation of the fractional polynomial and piecewise linear methods is performed using, respectively, Supp Code S17 and Supp Code S18.

In the fractional polynomial analysis (Figure 7), the best-fitting model selected power -1. There was no statistical evidence of departure from linearity (fractional polynomial non-linearity $p = 0.83$; quadratic test $p = 0.598$). The global heterogeneity test did not indicate violation of the homogeneity assumption (Cochran’s Q $p = 0.952$). The piecewise linear model similarly showed no evidence of non-linearity (quadratic $p = 0.598$; Cochran’s Q $p = 0.952$). Although LACE estimates varied across deciles (Supplementary Table S3), confidence intervals were wide and consistently included the null, providing no indication of a systematic threshold or J-shaped pattern. Non-parametric bootstrap resampling (1,000 replications) was used to obtain robust confidence intervals.

Overall, these findings provide no statistical evidence supporting a non-linear causal relationship between birth weight and LVM in this dataset. The estimated dose–response curve was compatible with a null or linear effect across the observed exposure range. Other possible reason is that the relatively modest sample size may have limited power to detect subtle non-linear effects.

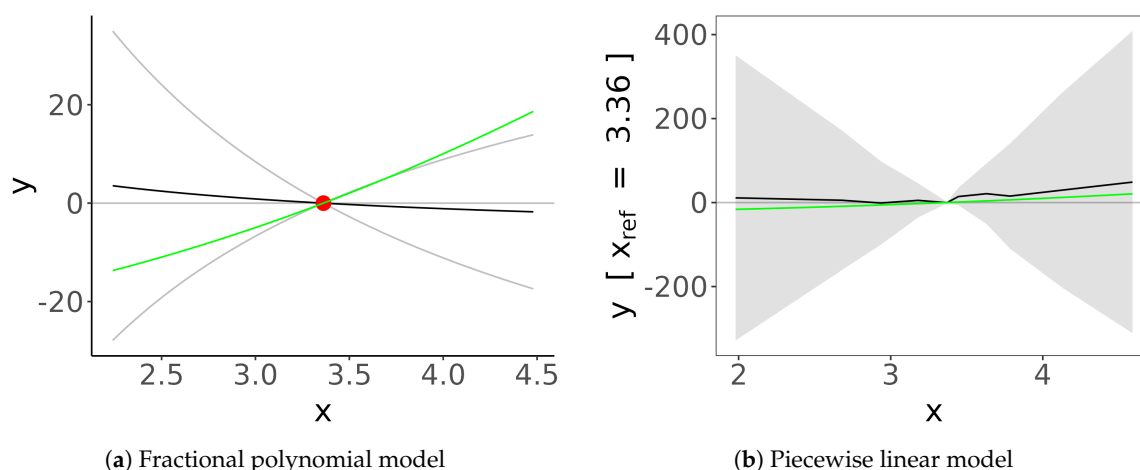


Figure 7. Non-linear Mendelian randomization analysis of Birth Weight on Left Ventricular Mass using SUMnlmr. (a) Fractional polynomial model with 95% confidence intervals. (b) Piecewise linear model based on decile-specific LACE estimates with bootstrap-derived 95% confidence intervals. No evidence of departure from linearity was observed.

10. Multi-Traits Analysis

Standard univariate GWAS analyses do not account for genetic associations across related traits, which can lead to the problem of missing heritability [23]. To address this, MTAG (Multi-Trait Analysis of GWAS) [24] provides a framework to jointly analysing summary statistics from multiple genetically correlated traits. By accounting for both genetic correlations and correlated estimation errors between traits, MTAG significantly increases statistical power compared to single-trait analyses. However, this method relies on the restrictive assumption that a universal variance-covariance matrix of effect sizes is shared for all SNPs across all analysed traits. Despite this strong assumption, MTAG remains a consistent estimator that typically outperforms single-trait predictors and produce robust results even with overlapped samples. The package <https://github.com/JonJala/mtag> MTAG requires GWAS summary statistics including: SNP IDs, alleles (effect and non-effect), sample size, Z-statistics and the effect allele frequency.

Table 4 represents the summary results from the MTAG analysis of the three traits. The columns are defined as follows:

1. #SNPs: The 3.8 million variants that passed quality control in all three files and were included in the analysis.
2. N_{\max} or N_{mean} : The input sample sizes from the original GWAS summary statistics.
3. GWAS mean χ^2 : The original statistic, which measures the average association strength from the single-trait GWAS results.
4. MTAG mean χ^2 : The updated statistic, which measures the average association strength from the MTAG results after borrowing information across traits.
5. GWAS equiv. N_{\max} : The effective sample size estimated from the increase in expected χ^2 which quantifies the gain in statistical power.

Take birth weight (BW) (Row 1) as an example: running MTAG was equivalent to adding the statistical power of recruiting 1,450 (GWAS equiv. $N_{\max} - N_{\max}$) new participants for the BW study. Generally, the higher the genetic correlation, the higher the boost MTAG will provide.

Table 4. Results of Multi-Trait Analysis of GWAS

Trait	#SNPs	N_{\max}	N_{mean}	GWAS mean χ^2	MTAG mean χ^2	GWAS equivalent N_{\max}
Birth weight	3829459	143677	138095	1.276	1.279	145127
LVM	3829459	16923	16923	1.089	1.089	16993
Gestational duration	3829459	84689	84609	1.142	1.148	87897

Overall, these results suggest that while there is some shared genetic architecture among the three traits, the magnitude of genetic correlation is modest. Consequently, the statistical power gain achieved through MTAG in this setting is limited rather than substantial. The primary implication of this analysis is that MTAG provides a small improvement in effective sample size and association strength for birth weight and gestational duration, with negligible improvement for LVM. Given the relatively modest gains and the strong modelling assumptions underlying MTAG, downstream analyses should interpret MTAG-enhanced results cautiously. In practice, MTAG may be particularly useful when traits are strongly genetically correlated, as greater genetic overlap typically leads to larger improvements in power. In the present case, the modest increases indicate limited but measurable cross-trait information sharing.

11. Challenges and Limitations

1. **Sensitivity to GWAS Discovery Parameters.** The reliability of MR estimates is heavily contingent on the quality of the underlying GWAS. Smaller sample sizes often lack the statistical power to identify robust instruments, leading to “weak instrument bias” which can pull results toward the observational association. Furthermore, as sample sizes increase, more genetic variants reach significance; however, including too many variants especially those with marginal effects can introduce noise or include pleiotropic variants that violate core MR assumptions and shift the final causal estimate.
2. **Lack of Ancestral Diversity in Phenotypic Data.** A major bottleneck in MR is the “Eurocentric bias” within genomic databases. Most large-scale GWAS for complex traits, particularly resource-intensive imaging phenotypes like brain morphology or cardiac volumes, are conducted in populations of European descent. This lack of diversity limits the generalizability of findings. Genetic architecture and LD patterns vary across ethnicities; applying instruments discovered in one population to another can lead to biased estimates or a total loss of statistical power.
3. **Assumption of Linearity in Causal Effects.** Most used standard MR frameworks, such as IVW, assume a linear relationship between the exposure and the outcome. However, biological systems often exhibit non-linear associations, such as U-shaped or threshold effects (e.g., the impact of alcohol consumption or BMI on mortality). When these methods are applied to non-linear data, they provide a “population-average” causal estimate that may mask critical nuances, potentially leading to misleading conclusions about the nature of the risk factor.
4. **Constraints of Individual-Level Data Access.** While methods exist to address non-linearity, they typically require individual-level data rather than summary statistics. Obtaining such data is often hindered by stringent data transfer agreements, privacy regulations (like GDPR), or the sheer logistical challenge of harmonizing datasets across different biobanks. Consequently, researchers are frequently restricted to two-sample MR using public summary data, which, despite its convenience, is mathematically limited to estimating linear effects and cannot easily explore complex dose-response relationships.
5. **Influence of Linkage Disequilibrium Settings.** The selection of parameters for clumping and pruning is a critical yet often subjective step that can significantly alter MR outcomes. These methods are used to ensure that genetic instruments are independent by removing variants in LD. If the r^2 threshold is too relaxed, redundant variants may be included, artificially inflating

the precision of the estimate. Conversely, overly stringent pruning can discard valid instruments, reducing statistical power and making the study susceptible to “winner’s curse” bias.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. All codes generated in the tutorial is provided in the supplementary.

Author Contributions: AMS drafted the original version of the manuscript and run the main analysis. RR run the MRlap and CAUSE analyses. YW run the MTAG analysis. IT run the SUMnlmr analysis. RR, YW and IT updated drafts of the article, with important contributions and revisions to their dedicated sections. AMS, RR, YW and IT wrote the statistical code for all used packages. RR, YW and IT contributed equally to the manuscript. GM and ZE revised the manuscript. AA provided statistical advices and interpretation of the results. All authors approved the final version.

Data Availability Statement: All the data used in the tutorial are publicly available and can be accessed through the published papers. The data used in the non-linear association are available at United Kingdom biobank.

Acknowledgments: AMS acknowledges the support of Leicester City Football Club (LCFC) and from British Heart Foundation (RE/24/130031). YW is supported from the Medical Research Council (MR/W007002/1). During the preparation of this manuscript, the authors used the free plan of Chat GPT solely to assist with language clarity, flow, and grammatical accuracy. All scientific content, ideas, and interpretations were developed, reviewed, and approved entirely by the human authors, who take full responsibility for the work.

Conflicts of Interest: AMS declares no conflicts of interest.

References

1. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O’Connell, J.; et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **2018**, *562*, 203–209.
2. Kurki, M.I.; Karjalainen, J.; Palta, P.; Sipil³a, T.P.; Kristiansson, K.; Donner, K.M.; Reeve, M.P.; Laivuori, H.; Aavikko, M.; Kaunisto, M.A.; et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **2023**, *613*, 508–518.
3. Sanderson, E.; Glymour, M.M.; Holmes, M.V.; Kang, H.; Morrison, J.; Munafò, M.R.; Palmer, T.; Schooling, C.M.; Wallace, C.; Zhao, Q.; et al. Mendelian randomization. *Nature reviews Methods primers* **2022**, *2*, 6.
4. Burgess, S.; Smith, G.D.; Davies, N.M.; Dudbridge, F.; Gill, D.; Glymour, M.M.; Hartwig, F.P.; Kutalik, Z.; Holmes, M.V.; Minelli, C.; et al. Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome open research* **2023**, *4*, 186.
5. Davies, N.M.; Holmes, M.V.; Smith, G.D. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *bmj* **2018**, *362*.
6. Li, J.; Morrison, J. Mind the gap: characterizing bias due to population mismatch in two-sample Mendelian randomization. *medRxiv* **2025**, pp. 2025–07.
7. Sanderson, E.; Richardson, T.G.; Hemani, G.; Davey Smith, G. The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *International journal of epidemiology* **2021**, *50*, 1350–1361.
8. Horikoshi, M.; Beaumont, R.N.; Day, F.R.; Warrington, N.M.; Kooijman, M.N.; Fernandez-Tajes, J.; Feenstra, B.; Van Zuydam, N.R.; Gaulton, K.J.; Grarup, N.; et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **2016**, *538*, 248–252.
9. Aung, N.; Vargas, J.D.; Yang, C.; Cabrera, C.P.; Warren, H.R.; Fung, K.; Tzanis, E.; Barnes, M.R.; Rotter, J.I.; Taylor, K.D.; et al. Genome-wide analysis of left ventricular image-derived phenotypes identifies fourteen loci associated with cardiac morphogenesis and heart failure development. *Circulation* **2019**, *140*, 1318–1330.
10. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; et al. The MR-Base platform supports systematic causal inference across the human phenome. *elife* **2018**, *7*, e34408.
11. Jones, S.R.; Carley, S.; Harrison, M. An introduction to power and sample size estimation. *Emergency Medicine Journal* **2003**, *20*, 453–458.
12. Deng, L.; Zhang, H.; Yu, K. Power calculation for the general two-sample Mendelian randomization analysis. *Genetic Epidemiology* **2020**, *44*, 290–299.

13. Marees, A.T.; De Kluiver, H.; Stringer, S.; Vorspan, F.; Curis, E.; Marie-Claire, C.; Derks, E.M. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research* **2018**, *27*, e1608.
14. Burgess, S.; Thompson, S.G.; Collaboration, C.C.G. Avoiding bias from weak instruments in Mendelian randomization studies. *International journal of epidemiology* **2011**, *40*, 755–764.
15. Pierce, B.L.; Burgess, S. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American journal of epidemiology* **2013**, *178*, 1177–1184.
16. Verbanck, M.; Chen, C.Y.; Neale, B.; Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics* **2018**, *50*, 693–698.
17. Burgess, S.; Davies, N.M.; Thompson, S.G. Bias due to participant overlap in two-sample Mendelian randomization. *Genetic epidemiology* **2016**, *40*, 597–608.
18. Mounier, N.; Kutalik, Z. Bias correction for inverse variance weighting Mendelian randomization. *Genetic epidemiology* **2023**, *47*, 314–331.
19. Huang, J.; Kleman, N.; Basu, S.; Shriver, M.D.; Zaidi, A.A. Interpreting SNP heritability in admixed populations. *Genetics* **2025**, p. iyaf100.
20. Liu, X.; Helenius, D.; Skotte, L.; Beaumont, R.N.; Wielscher, M.; Geller, F.; Juodakis, J.; Mahajan, A.; Bradfield, J.P.; Lin, F.T.; et al. Variants in the fetal genome near pro-inflammatory cytokine genes on 2q13 associate with gestational duration. *Nature communications* **2019**, *10*, 3927.
21. Morrison, J.; Knoblauch, N.; Marcus, J.H.; Stephens, M.; He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature genetics* **2020**, *52*, 740–747.
22. Staley, J.R.; Burgess, S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genetic epidemiology* **2017**, *41*, 341–352.
23. Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods* **2013**, *9*, 29.
24. Turley, P.; Walters, R.K.; Maghzian, O.; Okbay, A.; Lee, J.J.; Fontana, M.A.; Nguyen-Viet, T.A.; Wedow, R.; Zacher, M.; Furlotte, N.A.; et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* **2018**, *50*, 229–237.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.