Article

Mendelian randomization: Principles and methods

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; International Academy of Ecology and Environmental Sciences, Hong Kong, China
E-mail: wjzhang@iaees.org, zhwj@mail.sysu.edu.cn

Received 8 January 2024; *Accepted 19 February 2024*; *Published online 12 November 2024*; *Published 1 June 2025*

(cc) BY

Abstract

Mendelian randomization (MR) is a methodology for evaluating causality in observational studies. MR tries to find the fact that genotypes are not susceptible to reverse causation and confounding based on Mendel's law of inheritance. MR may provide information on causality in situations where randomized controlled trials are impossible. In present article, the principles and methods of MR were fully discussed.

Keywords Mendelian randomization; causality; causal inference; genetic association; exposure; outcome; confounder.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Causal inference is a research focus in science. In a broad sense, the statistical modeling between known dependent variable(s) and known independent variable(s) is categorized as causal inference. For example, construct a linear regression between the dependent variable, *y*, and the independent variable, *x*. However in many cases we do not know which is dependent variable or independent variable in two variables with causal relationship, i.e., the direction of causality is unknown. For such a case, Zhang (2021a-c) proposed three methodologies for determining causality direction. The first is a statistical simulation and regression methodology for causality inference of linearly correlated (scale or interval) variables (Zhang, 2021b). In this methodology, a statistical simulation and regression method was developed to generate and analyze artificial data of linear correlated variables with known causality. The rule was drawn from the simulation and regression analysis on artificial data. Finally, causality inference of two linearly correlated variables was conducted based on the rule (Zhang, 2021b). The second is a statistical simulation methodology for causality inference of Boolean variables (Zhang, 2021a). In this methodology, the statistical simulation was used to generate artificial data of two Boolean variables with known independent and dependent variables. A law was

drawn from the simulation analysis of the artificial data. For a set of data of two Boolean variables, a randomization method was proposed and used to test the statistical significance of the Boolean correlation measure (point correlation, quartile correlation, or Jaccard correlation, etc.). The causality inference was then conducted to observed data based on the law. Finally, the statistical simulation was used to determine the statistic significance of the causality (Zhang, 2021a). The third is a statistical simulation methodology for causality inference of nominal variables (i.e., categorical variables). In this methodology, a statistical simulation method was developed to generate artificial data of nominal variables with known causality. The law was then drawn from the simulation analysis of the artificial data. For a set of data of two nominal variables, the randomization method was used to test the statistical significance of thnominal correlation measure, and then the statistical simulation was used to determine the causality and its statistic significance of two nominal variables (Zhang, 2021c).

Mendelian randomization (MR) is the methodology for evaluating causality in observational studies. MR has the potential to provide information on causality in many situations where randomized controlled trials are not possible. It is widely used in epidemiological, pharmacological and public health research, especially in evaluating the impact of lifestyle factors, genetic susceptibility and drug targets on disease risk. For example, in the study of cardiovascular disease, cancer and metabolic diseases, MR methods help identify new risk factors and potential intervention targets.

Numerous studies using MR have been conducted in the last years (Zheng et al., 2017). Researhers used MR to report associations between various micronutrients and the risk of various cancers (Fu et al., 2021; Papadimitriou et al., 2021; Yuna et al., 2022). A mendelian randomization study that used genetic variants revealed no significant association between underlying propensities to differing caffeine metabolism and the risk of incident arrhythmia (Kim et al. (2021). Using single-nucleotide polymorphisms associated with micronutrient levels as instrumental variables, Kim et al. (2023) obtained instrumental variables of 14 genetically predicted micronutrient (vitamins and minerals) levels and applied two-sample MR to estimate their causal effects on 22 cancer outcomes from a meta-analysis of the UK Biobank (UKB) and FinnGen cohorts (overall cancer and 21 site-specific cancers). Results showed increased risk of breast cancer with magnesium levels and increased risk of colorectal cancer with vitamin B12 level. Zhu et al. (2024) conducted two-sample bidirectional Mendelian randomization (MR) analyses by using independent genetic variants associated with multiple social isolation phenotypes and with depression as genetic instruments from genome-wide association studies to evaluate the causality between social isolation and onset of depression. In two-sample bidirectional MR, the genetically predicted loneliness and social isolation combined phenotype (LNL-ISO) was positively associated with occurrence of depression. Henry et al. (2022) implemented a Mendelian randomization model that accounted for linkage disequilibrium between instruments and tested the robustness of causal estimates through a multiverse sensitivity analysis that included up to 120 combinations of instrument selection parameters and Mendelian randomization models per protein. The druggability of candidate proteins was surveyed, and mechanism of action and potential on-target side effects were explored with cross-trait Mendelian randomization analysis. Eight circulating proteins were associated with incident HF and showed evidence of a causal relationship. In DePaolo et al. (2023), causal effects of BP on AscAoD were estimated using 2-sample Mendelian randomization using a range of pleiotropy-robust methods. Rasooly et al. (2023) performed Mendelian randomization and colocalization analyses on human proteins to provide putative causal evidence for the role of druggable proteins in the genesis of heart failure. Using a combination of Mendelian randomization proteomics and genetic cis-only colocalization analyses, they identified 10 additional putatively causal genes for heart failure. Deng et al. (2023) performed univariable and multivariable-adjusted MR with adjustments for body mass index (BMI) and physical activity (PA), and found

that frailty was significantly associated with elevated risks of PD.

In present article, the principles and methods of MR will be discussed in detail.

2 Principles of Mendelian Randomization

2.1 Complexity of causality

A common problem in association analysis is that it is difficult to determine whether a variable is a true causal variable, rather than other unobserved factors that affect both the variable and the outcome, thus causing the variable to be associated with the outcome. In situations as evidence-based medicine or when formulating intervention strategies, it is necessary to clarify causality.

This problem is actually related to endogeneity (endogeneity in statistics refers to the correlation between the explanatory variable (x) and the error term in regression analysis), including reverse causation, omitted variable bias due to confounding, measurement error, and bidirectional causality (GWASLab, 2024).

For example, the relationship between obesity and heart disease may have four different scenarios (XM, 2024):

(A) Causality: i.e., true causality. There is a potential direct effect from obesity to heart disease. However, there may be confounding factors, which are other variables that affect both obesity and heart disease.

(B) Confounding: Due to confounding factors, although there is an association between obesity and heart disease, it is not causal relationship. The association may be caused in whole or in part by unidentified confounders that affect both, but there is no direct causal relationship between obesity and heart disease.

(C) Reverse causality: reverse causality, i.e., heart disease may cause obesity. This is the opposite of the situation (A), indicating that causality may be in the opposite direction. Confounding factors still exist and may affect the relationship between heart disease and obesity.

(D) Complex combination: For example, partially confounded, truly causal relationship, and bidirectional associations. The relationship between obesity and heart disease may be complex, involving direct and indirect factors, as well as confounders. There may be a direct association, reverse causation, and the influence of confounders at the same time.

2.2 Mendelian Randomization (MR)

2.2.1 Concept of Mendelian Randomization

Mendelian randomization (MR) is a method that uses genetic variation as the instrumental variable to study the causal relationship between exposure factors and outcomes. It is used to explore the causal relationship between biological factors (such as lifestyle or biomarkers) and outcomes (such as diseases). This method is based on Mendel's law of inheritance and uses the random assignment characteristics of genotype to phenotype in nature to reduce the influence of confounding factors and make causal inference. Because genetic variation is randomly assigned to individuals and is not affected by confounding factors (such as environmental and behavioral factors), it can provide an environment similar to a randomized controlled trial, thereby helping to solve the common confusion problems in observational studies and improve the reliability of causal inference. The core of the Mendelian randomization principle is to use Mendel's second law, or the law of independent assortment, that is, all DNA is randomly and independently combined during meiosis (the principle of random allele assignment): the parents with two pairs (or more pairs) of relative traits are hybridized, when gametes are produced in the next generation, while alleles are separated, genes on non-homologous chromosomes show free combination, which is similar to random grouping in randomized controlled trials. Therefore, Mendelian randomization is a randomized controlled trial based on Mendel's second law (GWASLab, 2024).

The core of the Mendelian randomization method is to use genetic data as instrumental variables (usually single nucleotide polymorphisms, or SNPs) to evaluate the relationship between an exposure factor (such as obesity, blood pressure) and an outcome (such as heart disease).

If a genetic variant is associated with the exposure factor of interest (*X*) and is not associated with the outcome (*Y*), then the genetic variant can be used as an instrumental variable (IV) (*Z*) to represent the exposure factor, thereby inferring the causal effect of the exposure factor on the outcome. In observational studies, the causal direction is often unclear, that is, whether *X* causes *Y* or whether *Y* causes *X*. The MR method uses genetic variants as instrumental variables to better help researchers determine the causal direction, thereby solving the common confounding and reverse causal problems in observational studies and drawing conclusions that are closer to causality.

2.2.2 Three basic assumptions of MR

In MR analysis, in order to ensure the validity and reliability of the results, three basic assumptions need to be met:

(1) Association hypothesis: The selected genetic variation *Z* must be reliably associated with the exposure factor *X* under study (the instrumental variable *Z* must have a strong association with the exposure factor *X*). This means that genetic variation should directly influence exposure factors. For example, genetic variants found to be significantly associated with exposure factors in genome-wide association studies (GWAS) can serve as instrumental variables.

(2) Independence assumption: The selected genetic variation *Z* has nothing to do with any known or unknown confounding factors *U* (the genetic variation cannot be related to any possible confounding factors, that is, the instrumental variable *Z* has nothing to do with any confounding factors). This means that the assignment of genetic variants should be random and undisturbed by other factors that influence exposure and outcome.

(3) Exclusion restriction hypothesis: Genetic variation *Z* affects outcome *Y* only through exposure factor *X*. This means that the effect of a gene on a disease outcome must be through the exposure factor, rather than the gene itself directly affecting the outcome.

In addition, it is also required to meet (GWASLab, 2024):

(4) There is no genetic assortative mating. For example, people often marry people with similar educational and economic levels.

(5) For all individuals, the direction of influence of instrumental variables on exposure factors is the same. For example, potential epistatic effects and GxE gene-environment interactions may influence this hypothesis.

The association between variables *X* and *Y* will definitely be affected by the potential confounding factor *U*, but there is no potential confounding factor between the instrumental variable *Z* and variable *X*, and between variable *Z* and variable *Y*. The association between variables *X* and *Y* cannot be obtained through direct observation, because variable *X* cannot be measured directly; but *Z* is measurable, and the direct association between *Z* and *X* is known or measurable, and it exists independently of other factors. Genetic variants (such as SNPs) that are strongly associated with exposure factors and follow the Mendelian randomization assumptions should be selected as instrumental variables, and the frequency of the variant, the strength of association with the exposure, and the biological interpretability need to be considered.

2.2.3 Comparison of Mendelian Randomization (MR) and Randomized Controlled Trial (RCT)

Generally speaking, the gold standard for clarifying causal relationships is the randomized controlled trial (RCT) (Zhang, 2024b), which means that the subjects are randomly divided into a control group and an experimental group to study the impact of a certain factor. However, in reality, it is very difficult to complete a randomized controlled trial, which requires a lot of manpower and material resources. Sometimes, due to

ethical issues, it is almost impossible to study a certain factor. At this time, other methods must be used, and Mendelian randomization is an alternative.

Mendelian randomization (MR) and randomized controlled trial (RCT) have similarities and differences in purpose, method, data type, cost, applicability, etc. In general, MR is directly related to RCT, and the two have high similarities.

Randomized controlled trials are experimental design methods that study the effects of intervention measures by randomly assigning participants to the experimental group and the control group (random allocation to the two groups). It is a more direct experimental design suitable for directly evaluating the impact of a specific intervention on the results. Randomized controlled trials reduce bias by randomly allocating key elements in experimental design, and control known and unknown confounders (confounders are expected to be equally distributed in the two groups) through random allocation. Intervention measures are fully controllable. In a randomized controlled trial, prospectively collected experimental data are used.

Mendelian randomization is an indirect method that relies on genetic data and uses the random assignment properties of genotype to phenotype (wild-type allele, variants) to explore potential causal relationships. Mendelian randomization is based on Mendel's law of inheritance and uses the random assignment of genetic variants (random segregation of genetic variants or alleles influencing exposure) as natural randomization. Mendelian randomization relies on the randomness of genetic variants to reduce the influence of confounding factors. The results are uncontrollable and determined by genetic variants. In Mendelian randomization, existing genetic and disease data are mainly used.

RCT directly changes weight through intervention measures (supplements) and observes the results, which is suitable for evaluating the effect of specific interventions. However, this design may require more resources, including time and cost, and may face ethical issues.

MR uses genetic variants to indirectly evaluate the relationship between exposure X (such as weight) and outcome Y (such as the risk of heart disease). Genetic variants are randomly assigned at birth and are not affected by acquired lifestyle and environmental factors, which helps to exclude confounding factors and reverse causality. In this way, MR can provide evidence that is closer to randomized controlled trial than observational studies. This is because in randomized controlled trials, interventions are randomly assigned to exclude the influence of confounding factors. Similarly, MR uses genetic variation as a natural randomization tool, providing a unique way to understand potential causal relationships. MR studies are able to reveal potential causal relationships between outcome Y (such as rare diseases) and exposure X (modifiable risk factors), while exploring such relationships in RCTs may require large sample sizes (Zhang, 2024a, b) and long-term follow-up. However, the interpretation of MR analysis results needs to be cautious because genetic effects may involve multiple biological pathways.

Both designs have advantages and limitations, and the choice of which design depends on the purpose of the study, available resources, and data. Ideally, the combination of the two methods can provide more comprehensive and reliable conclusions.

For example, to study the effect of weight gain on the risk of heart disease, studies can be designed using randomized controlled trial (RCT) and Mendelian randomization (MR), respectively. The following are the specific steps and characteristics of two designs:

(1) Randomized controlled trial (RCT)

Purpose: Directly evaluate the effect of weight gain (through supplements) on the risk of heart disease.

Participant allocation: Randomly divide participants into two groups, one receiving supplement treatment (experimental group) and the other receiving placebo (control group).

Experimental design: Ensure that the two groups of participants have similar health status and background variables before starting the experiment to reduce the influence of confounding factors.
Intervention measures: The experimental group takes supplements at a specific dose and time, while the control group takes a placebo.
Data collection: The weight of participants is measured regularly and their heart disease incidence is tracked.
Result analysis: Compare the incidence of heart disease in the two groups and evaluate the relationship between weight gain and heart disease risk.
(2) Mendelian randomization (MR)
Purpose: Use genetic variants as instrumental variables to indirectly evaluate the effect of weight gain on the risk of heart disease.
Genetic variation selection: Select genetic variants related to weight (such as specific SNPs) as instrumental variables.
Participant allocation: "Randomized" grouping based on whether the participant carries the allele that increases weight.
Data collection: Collect the genetic data of the participants, measure their weight, and track the incidence of heart disease.
Result analysis: Use statistical methods to analyze the differences in the incidence of heart disease among participants carrying different alleles, so as to infer the causal relationship between weight and heart disease risk.

2.2.4 Instrumental variable for MR

The statistical essence of Mendelian randomization is to use the instrumental variable to study causality, which is often used in economic research. In simple terms, an instrumental variable is a variable that is related to the exposure factor X but is unrelated to the ignored confounding factors and the outcome Y. In genetics, the instrumental variable is the gene. The advantages of using genotype as an instrumental variable are: (1) In genetic correlation, the direction of causality is determined, and genetic diversity leads to different phenotypes, but the opposite is not true. (2) In general, the environmental exposure factors we measure are more or less related to behavioral, social, psychological and other factors, which causes bias. However, genetic variation is not affected by these confounding factors. (3) Relatively speaking, the measurement error between genetic variation and its effect is small. (4) It is not necessary to find a causal SNP - a SNP in linkage disequilibrium with the causal SNP can meet the hypothesis. (5) Currently, GWAS data is relatively easy to obtain (GWASLab, 2024).

Some instrumental variables commonly used in MR research and their application methods are as follows:
(I) Single nucleotide polymorphism (SNP)
Single Nucleotide Polymorphism (SNP) refers to the variation of a single nucleotide in the DNA sequence. They are a key component of genetic diversity and are closely linked to genetic diseases in humans. In MR analysis, SNPs that are significantly associated with exposure factors are selected as instrumental variables. These SNPs should meet the condition of frequency greater than 1% and have a reliable association with the exposure factor under study. In Mendelian randomization studies, these SNPs (single nucleotide polymorphisms) should meet the condition that the frequency is greater than 1%, which means that the frequency of SNP variants selected as instrumental variables should exceed 1% in the population. This condition is based on the following considerations: (1) Genetic diversity: SNP is the variant form of a single nucleotide in the DNA sequence and is one of the main sources of human genetic diversity. Within a certain population, a specific SNP can be present with varying frequencies. (2) The importance of variant frequency: When conducting genetic correlation studies, we usually focus on those genetic variants that are relatively

common in the population. If the frequency of a SNP is very low (e.g., less than 1%), it may not be representative of the general genetic trend in the population and may not be statistically significant. (3) Representativeness and reliability of the study: Selecting SNPs with a frequency greater than 1% as instrumental variables can help ensure that the research results have broader representativeness and reliability. Such SNPs are more likely to reflect genetic trends prevalent in the population, making the research results more generalizable and applicable.

(II) Genome-wide association studies (GWAS)

Genome-Wide Association Studies (GWAS) data is a general term for a series of data that reflects the impact of SNPs on phenotypes. GWAS data provides association information between genotype and phenotype and is the basis of MR analysis. Through GWAS data, researchers can find SNPs that are significantly associated with specific phenotypes (such as diseases, biomarkers).

Commonly used genome-wide association analysis databases for MR are as follows:

(a) OpenGWAS (https://gwas.mrcieu.ac.uk/datasets/) is the simplest, most used and public comprehensive genome association study (GWAS) database, which brings together a large number of GWAS results, including gene correlation information for various phenotypes (such as diseases, physiological characteristics, behavioral characteristics, etc.). Finding genetic variants (SNPs) associated with specific phenotypes, you can find SNPs that are associated with the phenotype of interest (for example, the exposure variable and outcome variable).

(b) The GWAS Catalog (https://www.ebi.ac.uk/gwas/) is compiled by the European Bioinformatics Institute (EBI). It provides a consistent, searchable, visual and freely available for downloading of SNP traits-linked database that can be easily integrated with other resources and accessible to scientists, clinicians and other users around the world (Hemani et al., 2017). In this site, all eligible published GWAS studies were identified through literature searches and evaluated by staff, who then extracted traits, significant SNP-trait associations, and sample metadata addressed in the literature. The aim is to curate eligible studies based on literature availability within 1-2 months of article publication, with data published weekly. Submissions of unpublished GWAS data will also be accepted after 2020. Published GWAS data can be searched and viewed through the search page of the website, downloaded directly, or called through API. Aggregated data from GWAS can also be downloaded via FTP.

(c) UK Biobank (http://www.nealelab.is/uk-biobank; https://docs.google.com/spreadsheets/d/1kvPoupSzsSFBNSztMzl04xMoSC3Kcx3CrjVf4yBmESU/edit?pli=1#gid=227859291) collects extensive data from approximately 500,000 UK residents between the ages of 50 and 70, including genotypes, clinical measurements, questionnaires, biological samples, etc. The UK Biobank project adopts a long-term tracking design to track the health and disease status of participants, so as to better understand the relationship between genes and diseases, and the interaction between genes and the environment. The free version of the UK Biobank data is updated to August 2018, and this part of the data can be downloaded and used for free. If you want to get more updated data, you need to pay.

(d) FinnGen (https://www.finngen.fi/en; https://storage.googleapis.com/finngen-public-data-r9/summary_stats/R9_manifest.tsv) combines genomic information with digital healthcare data, summarizing GWAS and PheWAS results for multiple diseases, including genomic data from more than 1.4 million Finnish participants and electronic health records, combining participants' electronic health record data; the diseases are extensive, including cardiovascular, metabolic, cancer, mental and many other diseases. The FinnGen database allows free downloading of large sample size and phenotype-rich GWAS summary statistics.

(e) All of Us has identified more than 1 billion genetic variants, including more than 275 million previously unreported genetic variants, of which more than 3.9 million have coding consequences. Using the link between genomic data and longitudinal electronic health records, 3,724 genetic variants associated with 117 diseases were evaluated and found to have high replication rates in participants of European ancestry and participants of African ancestry (All of Us Research Program Genomics Investigators, 2024).

(f) The 1000 Genomes Project (https://www.internationalgenome.org/) created the largest public catalog of human variation and genotype data. With the end of the project, the EMBL-EBI data center received funding from the Wellcome Trust and created the IGSR (International Genome Sample Resource) website to ensure the accessibility of the 1000 Genomes Project data. In addition to European data, the 1000G data also includes a lot of Asian (including Chinese) data. These data can be queried online or downloaded for use.

(g) PAN-UKB (https://pan.ukbb.broadinstitute.org/) is a multi-ethnic study across six continents, which has conducted multi-ancestry analysis on 7,228 phenotypes and a total of 16,131 genome-wide association studies. It can be downloaded and used for free.

(h) PGC (Psychiatric Genomics Consortium; https://www.med.unc.edu/pgc/) is the largest psychiatric biology survey website ever, providing GWAS summary data for mental illnesses such as depression, bipolar disorder, and schizophrenia. Multiple institutions have collaborated to complete multiple large-sample GWAS meta-analyses.

(i) SSGAC (Social Science Genetic Association Consortium; https://www.thessgac.org/) provides GWAS data on behavioral genetics-related phenotypes such as education level, economic and political orientation, and personality.

(j) CTGLAB (Complex Trait Genetics Lab; https://ctg.cncr.nl/) explores the genetic and environmental causes of individual differences in brain-related health and disease. The website integrates knowledge from different fields (genetics, neuroscience, bioinformatics, biology, machine learning), uses and develops analytical tools to analyze and understand genomic data of complex traits, and links them with neuroscience to prove causal relationships in laboratory experiments.

(k) CNCR CTGLAB (https://ctg.cncr.nl/software/summary_statistics) provides GWAS summary statistics for more than 100 common diseases and human phenotypes, including GWAS data for metabolic, cardiovascular, immune, tumor and other diseases.

(III) Linkage disequilibrium (LD)

Linkage disequilibrium (LD) refers to the non-independent segregation of genetic variation. Adjacent gene variants on the same chromosome can be inherited together, and if the allele frequencies are similar, they will cause correlation between them. Linkage disequilibrium describes the situation where two or more SNPs frequently appear together in the same population. When selecting SNPs as instrumental variables, it is necessary to consider linkage disequilibrium and avoid selecting SNPs with high correlation due to linkage disequilibrium to reduce bias.

(IV) SNP-Clumping

This is a technique to reduce linkage disequilibrium problems by clustering SNPs that are highly associated with each other in the same gene region and selecting only one representative. When performing multi-SNP MR analysis, the SNP-Clumping technique is used to select representative SNPs. In multi-SNP MR analysis, in order to reduce linkage disequilibrium problems, SNPs that are highly associated with each other in the same gene region can be clumped together and only one representative can be selected. The process is as follows: First, all SNPs are ranked based on *p*-values (Zhang, 2022c) and the SNP with the smallest *p*-value is selected. All SNPs associated with it are removed around the SNP (e.g., within a fixed genomic window, such as 800kb) because they may be statistically associated with the SNP with the smallest *p*-value in the region.

This process is repeated until all SNPs have been considered or removed.

(V) Sensitivity analysis

Statistical methods are used to evaluate the stability and reliability of models, estimates, or results. In MR analysis, sensitivity analyses such as Leave-One-Out analysis and heterogeneity tests (Zhang, 2024a) are performed to ensure the stability of the results and reduce bias.

(VI) Horizontal pleiotropy and vertical pleiotropy

Pleiotropy refers to the effect of a single genetic variant on multiple phenotypes. Horizontal pleiotropy refers to the effect of a single nucleotide polymorphism on phenotypes of multiple biological pathways, while vertical pleiotropy refers to the effect of a single nucleotide polymorphism on multiple mediating pathways through an exposure factor. In MR analysis, pleiotropy needs to be identified and considered, especially horizontal pleiotropy, as it may violate the exclusion restriction assumption.

2.2.5 General steps of MR study

The following are the general steps of Mendelian randomization analysis. Through this process, a causal analysis closer to randomized controlled trials can be obtained in observational studies with confounding factors. For a specific outcome variable, there are many exposure factors that directly or indirectly affect it. Before conducting MR analysis, it is unknown which specific exposure factor will have a direct impact on the outcome variable. Therefore, when doing MR-related analysis, two preparations need to be made. One is psychological preparation, to learn to accept the fact that the results are not good; the other is to be prepared for both hands, to change the exposure or the outcome when the results are not good.

(1) Select genetic variants related to exposure factors as instrumental variables. At the beginning of the MR study, we need to select genetic variants, usually SNPs, that are strongly correlated with the exposure factors (such as weight) studied from the public GWAS database (such as the UK Biobank dataset, OpenGWAS). These genetic variants will be used as instrumental variables.

(2) Ensure the validity of the instrumental variable, that is, the instrumental variable meets the three basic assumptions of MR.

(a) Find SNPs that are strongly correlated with the exposure factors and ensure the association. It is necessary to ensure that the SNPs meet the premise assumptions. To extract strongly correlated SNPs, SNPs with $p<5*10^{-8}$ are generally selected to meet the association hypothesis.

(b) Remove SNPs with strong linkage disequilibrium to ensure independence. The linkage disequilibrium coefficient refers to whether the frequency of occurrence of two or more genotypes on different alleles is significantly different from their true frequency in the population. The linkage disequilibrium coefficient r refers to the degree of linkage disequilibrium between two loci in a set of study sequences, while the r^2 value refers to the proportion of linkage disequilibrium between the two loci, that is, the explained variance of linkage disequilibrium.

(c) Finally, calculate the F statistic and eliminate weak instrumental variables with $F<10$ or 100. The F statistic is used to compare the variance, goodness of fit, and significance of regression models between two or more groups. It is used to remove weak instrumental variables to ensure the reliability and accuracy of the results, and $F>10$ is required.

(3) Use appropriate statistical methods for causal inference. With effective instrumental variables, a variety of statistical methods can be used to estimate the causal effect of exposure factors on outcome variables. Commonly used methods include two-stage least squares (2SLS), inverse-variance weighting (IVW), weighted median method, model selection method (Egger), etc. See below for details.

(4) Conduct sensitivity analysis to evaluate the stability of the results. A sensitivity analysis method such as Leave-One-Out analysis can be used to examine the impact of a single SNP on the overall estimate by

removing one SNP in turn and re-performing the MR estimate. Methods based on Egger regression can also be used to detect whether there is a problem of horizontal pleiotropy.

(a) Heterogeneity test. The purpose is to see whether different genetic variants (SNPs) have a consistent impact on the results. If there is heterogeneity in the results, it means that the exposure factors may have inconsistent effects on the outcome variable. At this time, a random effects model needs to be used to estimate the causal effect of the exposure factor on the outcome variable and determine whether it still has a statistically significant effect. Inconsistent effects found in heterogeneity tests may indicate that different genetic variants of the exposure factor may affect the occurrence of the outcome variable in different ways. In other words, there may be some specific genetic variants that have a special impact on the expression or function of the exposure factor, thereby affecting the occurrence of the outcome variable. Random effects models were used to account for these heterogeneities and re-estimate the causal effects of exposure factors on outcome variables. In layman's terms, think of this process as a team: even though the performance of each team member (the team here can be imagined as genetic variation) may be different, their impact as a team (the team here can be imagined as the exposure factor) on the outcome (the outcome here can be imagined as the outcome variable) The impact is still significant.

(b) Horizontal pleiotropy test. The purpose is to see whether some genetic variants (SNPs) have an undue influence on the relationship between exposure factors and outcome variables. Being affected by horizontal pleiotropy means that the SNPs of the exposure factor may have too large or too small effect on the outcome variable. This may be due to the fact that these SNPs affect the occurrence of outcome variables through other unknown pathways in addition to affecting exposure factors. This has important implications for causal inference. Ideally, we would like all SNPs to affect the occurrence of the outcome variable solely by affecting the exposure factor. However, if horizontal pleiotropy exists, then these SNPs may affect the occurrence of the outcome variable through other pathways, which may cause us to overestimate or underestimate the true impact of exposure factors on the outcome variable.

(c) One-by-one elimination test. The purpose is to see whether the results will change significantly if a certain genetic variant (SNP) is removed. The results are more stable, meaning they are less likely to be affected by any one specific SNP. In a one-by-one elimination test, each SNP is removed in turn and then it is seen whether the result variable will change significantly. If the result variable does not change significantly after eliminating any SNP, it means that the stability of the result is high. This means that the results are unlikely to be affected by any one specific SNP, but are determined by all SNPs together.

(d) Reverse MR analysis. Not only analyzes the impact of genes on a specific outcome (disease), but also analyzes the impact of disease on genes. By comparing the results obtained in these two directions, the causal relationship between genes and diseases can be more accurately evaluated and the impact of confounding factors can be reduced.

(5) Interpret the results. Based on the results of MR analysis, explain the causal relationship between exposure factors and outcome variables. If the MR analysis shows a significant causal relationship, it can be concluded that the exposure factor is important for the outcome variable. If a causal relationship is not significant, further investigation into other potential explanations may be warranted, such as whether there are unaccounted for confounding factors.

3 Methods of Mendelian Randomization

If a gene variant Z (instrumental variable) is a causal variable of an exposure factor X and has no direct causal relationship with the outcome Y, the causal relationship between X and Y should be explored. The association

between gene variant *Z* and outcome *Y* can only be observed through the causal relationship between *X* and *Y* (*X*→*Y*) (GWASLab, 2024).

3.1 Univariate MR

To evaluate the causal relationship between a specific exposure and a specific outcome, one or more SNPs that are strongly associated with the specific exposure can be used as instrumental variables, which is univariable MR (Univariable MR). For example, researchers may use SNPs associated with an exposure factor to evaluate whether the exposure factor is associated with an increased risk of disease.

In univariate MR, the effect of *X* on *Y* can be estimated using the two-stage least squares method (2SLS). Using the instrumental variable, the effect of exposure on the outcome is estimated in two stages (GWASLab, 2024; XM, 2024).

In the first stage, the instrumental variable is used to predict exposure, that is, *X* is regressed on the instrumental variable

$$X = \alpha + \gamma Z$$

In the second stage, the predicted exposure is used to estimate the impact on the results, that is, *Y* is regressed on the predicted value of *X* in the first stage

$$Y = \beta_0 + \beta_1 \hat{X}$$

After merging, it can be transformed into *Y* to directly regress the instrumental variable

$$Y = \mu + \rho Z$$

The coefficient $\beta_{2\text{sls}}$ we are concerned about is actually equivalent to the ratio of the covariance of the two segments

$$\beta_{2\text{sls}} = \frac{Cov_{y,z}}{Cov_{x,z}}$$

Different statistical methods should be used according to the type of outcome variable. Linear regression is used for continuous results, and logistic regression is used for binary results.

One-sample MR (One Sample Mendelian Randomization) assumes that a genetic variant *Z* is associated with a specific phenotypic exposure *X*, then the genetic variant *Z* should also be associated with the outcome *Y* of the phenotypic feature, and is calculated using the 2SLS statistical analysis method to provide a basis for causal inference (Hernán and Robins, 2006; Palmer et al., 2024). One-sample MR is easy to operate, which does not require external data, and is limited to a single sample, that is, the exposure and the outcome come from the same sample. The selection range of the instrumental variable is relatively limited, and the causal relationship only comes from the same data set, which is easily affected by weak instrumental bias and horizontal pleiotropy (Wang, 2023).

3.2 Inverse-Variance Weighted (IVW)

Inverse-Variance Weighted (Zhang, 2024a) is a common method in MR analysis, which is used to integrate the effects of multiple instrumental variables (such as multiple SNPs) and requires that all instrumental

variables are valid. Based on the effect estimate and variance of each instrumental variable, the effect estimate weight is given, and the instrumental variable with a small variance has a large weight value.

First, for each SNP, its effect estimate on the exposure and outcome is calculated.

Then, each SNP is weighted by the variance of the effect estimate.

Finally, the total causal effect estimate is calculated by weighted average (Burgess et al., 2013).

$$\beta_{\mathrm{IVW}} = \frac{\sum \frac{\beta_{\mathrm{exposure}} \times \beta_{\mathrm{outcome}}}{\mathrm{Var}(\beta_{\mathrm{outcome}})}}{\sum \frac{\beta^2_{\mathrm{exposure}}}{\mathrm{Var}(\beta_{\mathrm{outcome}})}}$$

The IVW method is used with either random-effects or fixed-effects models (Burgess et al., 2013). Fixed-effects models are used when there are three or fewer genetic variants and random-effects models are used when there are four or more genetic variants. Random-effects models allow for heterogeneity among the causal estimates targeted by the genetic variants and for overdispersion in weighted linear regression (standard linear regression or robust regression; Burgess et al., 2016), as well as residual standard errors ≥ 1. The null hypothesis is that all genetic variants estimate the same causal parameter (heterogeneity statistic (Cochran's *Q* statistic; Zhang, 2024a) and associated *p*-value). Rejection of the null suggests that one or more variants may be pleiotropic (del Greco et al., 2015; Yavorska and Staley, 2023). In fixed-effects models, residual standard errors = 1. Weights can be penalized to reduce the contribution of genetic variants to the analysis and the influence of outlying ratio estimates (Burgess et al., 2016).

In weighted linear regression, weights can be simple. In this case, IVW estimation is equivalent to meta-analyzing the ratio estimate for each variant using inverse-variance weights based on the simplest expression for the variance of the ratio estimate (δ-expanded first-order terms - standard error of the association with the outcome divided by the association with the exposure). If δ-weights are used, the variance expression is the δ-expanded second-order terms (Burgess and Bowden, 2015). The second-order terms incorporate uncertainty in the genetic association with the exposure, and this uncertainty can be ignored using simple weighting (Burgess and Bowden, 2015).

For strict two-sample Mendelian randomization analyses (i.e., no overlap), the correlation between the genetic association with the exposure and the association with the outcome for each variant generated by sample overlap is set to 0. The correlation should be set to the observed correlation between the exposure and the outcome. For δ-weights, this correlation is used only to calculate standard errors.

Yavorska and Staley (2023) note that for multiple uncorrelated genetic variants, the estimate can be thought of as: (1) an inverse-variance weighted combination of the ratio estimates from the meta-analysis (Zhang, 2024a), (2) combining the genetic variants into a weighted score and then using that score as the ratio estimate for the instrumental variable (the same estimate is obtained with two-stage least squares using individual-level data), and (3) the coefficients of a weighted linear regression of the association with the outcome on the risk factor with the intercept fixed to zero and using inverse-variance weights. Causal estimates are obtained by regressing the association with the outcome on the association with the risk factor with the intercept set to zero and the weights being the inverse variance of the association with the outcome.

For a single genetic variant, the estimate is the ratio of the coefficients $\frac{\beta_Y}{\beta_X}$ and the standard error is the first term of the δ method approximation $\frac{\beta_{Y_{se}}}{\beta_X}$.

3.3 Penalized Inverse-Variance Weighted (pIVW) Method

The penalized inverse variance weighted (pIVW) estimator simultaneously accounts for weak instruments and balanced-level pleiotropy in a two-sample MR with summary statistics, i.e., an exposure sample (with IV exposure effects and standard errors) and an outcome sample (with IV outcome effects and standard errors) (Xu et al., 2023).

The pIVW estimator also allows for IV selection in a three-sample MR, where weak IVs are screened out using an additional sample (with IV exposure effects and standard errors) independent of the exposure and outcome samples.

3.4 Median

The weighted median or simple median method introduced by Bowden et al. (2016) is used to calculate the median of the ratio instrumental variable estimates estimated using each genetic variant individually (Yavorska and Staley, 2023).

When the weights are simple (simple median), the estimates are obtained by calculating the ratio causal estimate for each genetic variant (Bowden et al., 2016)

$$\theta = \frac{\beta_Y}{\beta_X}$$

and find the median estimate.

Computing the weighted median of the instrumental variable effect estimates provides another evaluatio of the results, which is a robust method for estimating causal effects, even when up to 50% of the instrumental variables are invalid. In this method, the effect estimates for each SNP are first ranked, and then the weighted median is used as the estimate of the causal effect.

When weighted (weighted median), the estimate is obtained as follows (Bowden et al., 2016; Yavorska and Staley, 2023):

(1) Calculate the causal estimate ratio and rank the genetic variants according to the size of the estimate, i.e.,

$$\theta_1 < \theta_2 < \cdots < \theta_j.$$

(2) Calculate the normalized inverse-variance weight for each genetic variant, $w_1 < w_2 < \cdots < w_j$,

$$w_j = \frac{\mathrm{frac}(\beta_{X_j}^2)s_{\beta_{Y_j}}{}^2}{\sum_i \mathrm{frac}(\beta_{X_i}^2)s_{\beta_{Y_i}}{}^2}$$

(3) Find *k* such that

$$s_k = \sum_{i=1}^{k} w_i < 0.5,\ s_{k+1} = \sum_{i=1}^{k+1} w_i > 0.5$$

(4) Calculate the weighted median estimate by extrapolation

$$\theta_{\mathrm{WM}} = \theta_k + (\theta_{k+1} - \theta_k) * \mathrm{frac}(0.5 - s_k)(s_{k+1} - s_k)$$

When all weights are equal, the simple median estimate is the same as the weighted median estimate. The standard errors for both the simple median and weighted median methods are calculated using bootstrapping.

Compared to the inverse-variance weighted method and the Egger method, the median-based method is more robust to individual genetic variants with strong outlier causal estimates. Formally, the simple median

method gives consistent estimates of the causal effect when at least 50% of the genetic variants are valid instrumental variables (for the weighted median method, when 50% of the weights come from valid instrumental variables) (Yavorska and Staley, 2023). When weights are penalized, the weighted method is used, but the contribution of genetic variants with outlier (heterogeneous) ratio estimates to the analysis is reduced.

3.5 Egger Method

The Egger method uses a random effects model. The Egger method provides: (1) directional pleiotropy tests (Egger intercept tests), (2) causal effect tests, and (3) causal effect estimates (Yavorska and Staley, 2023). Causal estimates are obtained by regressing the association with the outcome on the association with the risk factor, with weights being the inverse variance of the association with the outcome. The intercept term in the regression analysis of β_{outcome} on β_{exposure} (without constant term) represents the average pleiotropy of genetic variants (average direct effect on the outcome) and is used to test whether there is horizontal pleiotropy, taking into account possible heterogeneity of instrumental variables, and providing a corrected estimate of causal effect. If the intercept term is not zero, then not all genetic variants are effective instrumental variables and the standard (inverse-variance weighted) estimate is biased, specifically, there is directional pleiotropy. If the InSIDE (instrumental variable strength independent of direct effect) assumption holds, then the Egger slope parameter provides a test of causal effect and provides a consistent estimate of causal effect even if the intercept is different from zero. The method is implemented in three steps:

(1) Regression analysis (robust regression or standard linear regression) using effect estimates of all SNPs;

(2) Corrected causal effect estimate is obtained through regression;

(3) Correct and test for certain biases using the Egger regression intercept term (Bowden et al., 2015).

If genetic variants are correlated, then this correlation can be explained. A correlation matrix must be provided: the elements of this matrix are the correlations between the individual variants (diagonal elements are 1). If correlations are specified, robust regression and penalization are not allowed (Burgess et al., 2016).

The null hypothesis is that the Egger regression model describes the association with the outcome without excessive heterogeneity (using the heterogeneity statistic (Cochran's Q statistic) and the associated p-value). If the genetic variants are pleiotropic, the null hypothesis is expected to be rejected, but this does not mean that the Egger analysis or the InSIDE assumptions are invalid. I^2 is used to measure heterogeneity between genetic associations and exposures (Bowden et al., 2016; Zhang, 2024a). Low I^2 values are associated with both larger precision differences between Egger and IVW estimates and weaker instrumental variable bias (Bowden et al., 2016; Yavorska and Staley, 2023). Low p-value for the Egger intercept test indicate directional pleiotropy or failure of the InSIDE assumption and suggest that the IVW estimates are biased.

3.6 Lasso Method

Lasso extends the IVW model to include an intercept term for each genetic variant. These intercept terms represent the association between the genetic variant and the outcome bypassing the risk factor. The causal effect is estimated by weighted linear regression, where the intercept term is subject to a Lasso penalty. The Lasso method applies a Lasso penalty to the direct effect of the genetic variant on the outcome. The Lasso penalty tends to shrink the intercept term corresponding to the effective instrument to zero. The causal estimate is described as the post-Lasso estimate and is obtained by performing the IVW method using only those genetic variants identified as effective by the Lasso procedure. The Lasso method is implemented in two steps:

(1) Fit a regularized regression model and identify some genetic variants as effective instruments.

(2) Estimate the causal effect using a standard multivariate IVW that includes only effective genetic variants.

The post-Lasso method is performed whenever the number of genetic variants identified as effective instruments is greater than the number of risk factors. The default heterogeneity stopping rule will always

return more genetic variants as effective instruments than identified risk factors.

If a significant proportion of genetic variants are removed from the analysis, the MR-Lasso method may give a false impression of confidence in the causal estimate due to the homogeneity of the causal estimate for a particular variant among the remaining variants. However, it is unreasonable to claim strong evidence for causality after removing a large number of variants with heterogeneous estimates from the analysis.

The tuning parameter λ used by the Lasso procedure controls the degree of sparsity. If not specified, the tuning parameter is calculated by the heterogeneity stopping rule. For completeness, parameter estimates of the regularized regression model used to identify invalid variants can be also provided (Yavorska and Staley, 2023). The intercept is estimated from the regularized regression model in the Lasso method. An intercept estimate of zero identifies the corresponding genetic variant as a valid instrument. Genetic variants with non-zero intercept estimates are excluded from the post-Lasso estimator (Yavorska and Staley, 2023).

3.7 Maximum-likelihood Method

Burgess et al (2013) introduced the maximum likelihood method. The likelihood function is defined by assuming that the summary data for each genetic variant are normally distributed. For the association of each genetic variant with exposure and outcome, a bivariate normal distribution was assumed. The mean association with an outcome is considered as the mean association with the exposure multiplied by the causal effect parameter. Therefore, if there are k genetic variants, k+1 parameters are estimated by: one for each gene - the exposure association, plus the causal parameter. If the number of genetic variants is large, then maximizing this function may be a problem. If the maximum-likelihood estimate differs significantly from the inverse-variance weighted estimate, this may indicate that convergence has not occurred in the optimization algorithm. The variance-covariance matrix of the bivariate normal distribution is obtained from the provided standard error estimate. The correlation between genetic associations and outcomes due to exposure and sample overlap can be specified and has a default value of zero.

Two features that make this approach superior to the inverse-variance weighting approach are that uncertainty about the genetic association with the exposure is incorporated into the model and the correlation between the genetic association estimate for each variant and the exposure and outcome. The method works for both unrelated and related genetic variants. It can also be used for individual genetic variants.

The original version of the maximum-likelihood method assumes that all genetic variants have the same causal estimate; if the fixed-effects model is incorrect and there is large heterogeneity in the causal estimates of different variables, the causal estimates may be too precise. The random-effects analysis implemented is an ad hoc solution to the heterogeneity problem, but should produce reasonable confidence intervals that incorporate this heterogeneity.

This method naturally estimates fixed-effects models, assuming that each genetic variant estimates the same causal effect. However, if there is heterogeneity in the causal estimates of different variables, the confidence intervals under the fixed-effects model will be too narrow. Random-effects models add additional uncertainty by multiplying the standard error by the square root of the likelihood ratio heterogeneity statistic divided by the number of genetic variants minus one (unless the number is less than 1, in which case no modification of the criterion is made). This is similar to the residual standard errors in regression models (the Cochran's Q heterogeneity statistic is equal to the RSE squared times the number of genetic variants minus one).

If genetic variants are correlated, then this correlation can be explained. A correlation matrix must be provided, the elements of which are the correlations between variants (the diagonal elements are 1).

The null hypothesis is that all genetic variants estimate the same causal parameter; rejection of the null value indicates that one or more variants may be pleiotropic (using heterogeneity statistics (likelihood ratio

statistics) and associated *p* values).

3.8 Constrained Maximum-likelihood (cML) Method

The cML method uses constrained maximum likelihood to select null IVs with relevant and/or irrelevant tropism effects (Lin et al., 2023). Data perturbation can be used to explain selection uncertainty when many null IVs have weak pleiotropy. When performing DP (data perturbation), two goodness-of-fit (GOF) tests are developed to check whether the model-based and DP-based variance estimates converge to the same estimate. The small *p*-value of the GOF test indicates that the selection uncertainty is not negligible and that the results of DP are more reliable.

Since the constrained maximum likelihood function is non-convex, multiple random starting points can be used to find the global minimum. For some starting points, the algorithm may not converge.

3.9 Hartwig Method

Hartwig's mode-based approach obtains variant-specific ratio estimates from each genetic variant in turn and computes the mode estimate. This is done by constructing a kernel smoothed density from the ratio estimates and taking the maximum as the mode estimate. Standard errors are computed using a bootstrap procedure and confidence intervals are based on estimates with a normal distribution (Zhang, 2022a, b). If multiple (or weighted multiple) genetic variants are valid instruments, this approach should give consistent estimates as sample size increases. This means that the largest set of variants with the same causal estimate in the asymptotic limit are valid instruments.

In this approach, standard error estimates can be either

(1) simple - computed as a first-order term of a δ expansion - the standard error of the association with the outcome divided by the association with the exposure, or

(2) δ - computed as a second-order term of a δ expansion (the default option). The second-order term incorporates uncertainty about the genetic association with the exposure, which can be ignored using simple weighting.

The bandwidth of 1 in the kernel smoothed density approach represents the bandwidth value chosen by the modified Silverman bandwidth rule recommended by Hartwig et al. (2017). A bandwidth of 0.5 represents half this value.

3.10 Multivariable MR

If one want to evaluate the joint effects of multiple exposure factors on an outcome at the same time, one can use multivariable MR, such as multivariable inverse-variance weighting method, multivariable Egger method, multivariable MR-Lasso method, multivariable median method, etc. Multivariable MR is based on Mendel's law of inheritance. It randomly groups multiple variables at the same time to make the distribution of variables between groups random, so as to improve the credibility and accuracy of the results. In multivariable MR, a set of SNPs can be selected as instrumental variables for each exposure, and multivariate statistical analysis can be performed. This method allows the genetic association of multiple risk factors to be considered simultaneously in MR analysis, which helps to adjust known confounders or explore the mediating effects between different factors. For example, researchers may simultaneously consider multiple SNPs associated with exposure factors and their relationship with the occurrence of the disease. Multivariable MR requires that the number of SNPs exceeds the number of exposure factors. The null hypothesis is that all genetic variants estimate the same causal parameter; rejection of the null indicates that one or more variants may be pleiotropic (using the heterogeneity statistic (likelihood ratio statistic) and the associated *p*-value).

(1) Multivariable IVW method

Multivariable Mendelian randomization is an extension of Mendelian randomization to handle genetic variants associated with multiple risk factors. Two scenarios are envisioned for its use:

(i) Biologically relevant risk factors, such as lipid composition;
(ii) Risk factors for which there is a network of causal effects (mediation) between one risk factor and another.
In both cases, under the extended assumption of multivariable Mendelian randomization, the coefficients represent the direct causal effect of each risk factor in turn, while the other risk factors are fixed.

The method is implemented using multivariable weighted linear regression. If the variants are correlated, the method is implemented using generalized weighted linear regression. Causal estimates are obtained by regressing the association with the outcome on the association with the risk factor, with the intercept set to zero and the weights being the inverse variance of the association with the outcome.

(2) Multivariable Median method

It performs multivariable Mendelian randomization via the median method, implemented via multivariable weighted quantile regression with the quantile set to 0.5. The regression model is multivariate and weighted by the inverse variance of the specific variant estimate. Confidence intervals are calculated via a parametric bootstrap procedure to estimate the standard error of the estimate, and then use quantiles from the normal distribution or *t* distribution (Yavorska and Staley, 2023; Zhang, 2022a, b).

(3) Multivariable Egger method

Multivariate Egger is an extension of the Egger method to handle genetic variants associated with multiple risk factors. The method is implemented using multivariate weighted linear regression. If the variants are correlated, the method is implemented using generalized weighted linear regression. Causal estimates are obtained by regressing the association with the outcome on the association with the risk factor, with an intercept estimate and weights that are the inverse variance of the association with the outcome (Yavorska and Staley, 2023).

Both the univariate and multivariate versions of Egger are sensitive to the choice of parameterization of the genetic association - which allele the association is relative to (i.e., which allele is the effect allele). For univariate Egger, this problem can be solved by setting the genetic associations with the exposure to all positive. In multivariate Egger, we must choose which exposures to target the genetic associations to.

(4) Multivariable Lasso method

The multivariable Lasso method applies a Lasso penalty to the direct effect of a genetic variant on the outcome (Grant and Burgess, 2020). The causal estimate is described as a post-Lasso estimate and is obtained by performing the multivariable IVW method using only those genetic variants identified as significant by the Lasso procedure.

The multivariable Lasso extends the multivariable IVW model to include an intercept term for each genetic variant. These intercept terms represent the association between the genetic variant and the outcome bypassing the risk factor. The regularized regression model is estimated by multivariable weighted linear regression, where the intercept term is subject to a Lasso penalty. The Lasso penalty tends to shrink the intercept term corresponding to significant instruments to zero (Yavorska and Staley, 2023).

The main estimate given by this method is the post-Lasso estimate. If a significant proportion of the genetic variants are removed from the analysis, the multivariable Lasso method may give a false impression of the confidence in the causal estimate because of the homogeneity of the causal estimates for specific variants among the remaining variants. However, it is unreasonable to claim strong evidence for causality after removing a large number of variants with heterogeneous estimates from the analysis.

The Lasso penalty relies on an adjustment parameter that controls the degree of sparsity. The default is to use a heterogeneity stopping rule, but a fixed value can be specified.

As part of the analysis, the genetic variants are oriented so that all associations with one risk factor are positive (the first risk factor is used by default). Reorientation of genetic variants is performed automatically as part of this function.

(5) Multivariable Constrained Maximum-likelihood (MVcML) method

The multivariable constrained maximum-likelihood (MVcML) method is robust to both correlated and uncorrelated pleiotropy (Lin, 2023). Multivariate cML (MVcML) is an extension of cML that handles multiple exposures of interest. As its univariate version, it is robust to both correlated and uncorrelated pleiotropy.

In practice, the data perturbation (DP) version is preferred in practice because it can account for uncertainty in model selection for more robust inference (Yavorska and Staley, 2023).

(6) Multivariable Generalized Method of Moments (GMM) method

This is a robust inference of two-sample multivariate Mendelian randomization using the generalized method of moments. This method accounts for overdispersed heterogeneity in genetic variant-outcome associations (Hanson, 1982).

3.11 Simple Mode

If there is only one instrumental variable, the effect estimate of the instrumental variable is used directly to estimate the causal effect. First, a strongly correlated SNP is selected as the instrumental variable, and then the effect estimate of the SNP is used to estimate the causal effect of the exposure on the outcome

$$\beta_{\text{simple}} = \frac{\beta_{\text{outcome}}}{\beta_{\text{exposure}}}$$

3.12 *F* Statistic

The *F* statistic is a statistical indicator used to evaluate the strength of an instrumental variable. If the *F* statistic of an instrumental variable (such as a SNP) is low, it may indicate that the instrumental variable is weak, which may lead to weak instrument bias. The conditional *F* statistic is an approximation of the first-stage conditional *F* statistic based on all variants of the aggregated data. This represents the strength of the instrument for each exposure conditional on other exposures in the model. This is only reported when the sample size of the genetic association associated with the exposure is provided (Yavorska and Staley, 2023).

3.13 Two-Sample MR

Two-Sample MR is used to estimate causal relationships when the data on exposure and outcome come from different populations (when a single sample containing both exposure and outcome data is lacking). Two different data sets from similar backgrounds can be used, one for analyzing the association between *Z* and *X* (exposure factors) and the other for analyzing the association between *Z* and *Y* (disease outcomes), and better causal estimates and sensitivity analyses can be guaranteed through sample size advantages and optimized statistical analysis methods. For example, one population provides exposure data and the other provides disease occurrence data, and the two-sample MR method is used to exploit the causal relationship between exposure and disease occurrence.

In a two-sample MR, if there is an instrumental variable associated with *X*, the association between this instrumental variable and *Y* can only be observed if *X* has a causal effect on *Y*. This means that

$$\beta_{z,y} = \beta_{z,x} \times \beta_{x,y}$$

That is, instead of estimating β by regressing *X* on Y, we can simply use

$$\beta_{x,y} = \beta_{z,y} / \beta_{z,x}$$

to calculate the effect size of *X* on *Y*. This means that in contrast to the two-stage least squares method, the summary statistics of two independent GWAS can be used to calculate this ratio.

Two-sample MR can provide a richer selection of data sources and can evaluate the generality of causal relationships in different groups. Two-sample MR is also the most common type of MR designs in the current big data context, but it should be noted that since two-sample MR requires the use of external two-stage MR, thus selection bias may be introduced, such as the winner's curse caused by using GWAS results as instrumental variables, which may lead to the overestimation of the association between the instrumental variable and *X* and the underestimation of the causal association.

Two Stage Mendelian randomization (TSMR) is a variant of two-sample MR (Spiller et al., 2019). It can be used to evaluate whether the mediator mediates the effect of exposure *X* on outcome *Y*, not just the association between a single factor and a certain outcome. It is suitable for finding complex relationships between multiple factors and inferring the mechanism of exposure *X* to outcome *Y* through unpacking methods. Two-stage MR supports the simultaneous evaluation of multiple causal relationships and can discover complex causal networks, but the effectiveness of instrumental variables affected by the intensity and frequency of genetic variations requires more data and statistical analysis, and the interpretation of the results is more challenging (Wang, 2023).

3.14 Multi-Sample MR

Multi-Sample MR uses multiple samples from different populations to enhance the statistical power and generalization ability of the analysis. The method is to integrate genetic and phenotypic data from different populations for comprehensive analysis. Multi-sample MR can be used when the exposure or outcome involved in the study may be heterogeneous in different populations. For example, analyze how genetic susceptibility affects the risk of hypertension in different ethnic populations.

3.15 Bidirectional MR

Bidirectional MR studies whether there is a bidirectional causal relationship between two variables. It consists of two two-sample MR analyses and is a variant of two-sample MR. In essence, it is to evaluate whether there is a reverse causal relationship between exposure and outcome, that is, whether the outcome can cause the exposure. When it is suspected that two variables may affect each other, the associated SNPs can be used to evaluate the causal effect of the two variables as exposure and outcome. For example, study the relationship between depression and sleep disorders, evaluate whether depression causes sleep disorders, and whether sleep disorders increase the risk of depression.

Bidirectional MR can avoid the confusion caused by reverse causality and have a more comprehensive understanding of causality. However, the instrumental variable assumption in both directions needs to be met at the same time.

3.16 Mediation MR

When it is necessary to clarify how a certain lifestyle habit affects the development of a disease through physiological mechanisms, mediation MR can be used to explore how exposure (such as smoking) affects the outcome (such as the risk of respiratory infection) through one or more mediating variables (such as lung function). Mediation MR solves the confounding problem in causal inference through the natural random assignment of genetic variants to determine whether there is a mediating mechanism that can explain the effect of exposure on the outcome. In mediation MR, the effect of the main exposure on the mediating variable is first evaluated, and then how the mediating variable affects the outcome and whether the main exposure has other direct ways to affect the outcome. For example, researchers evaluate the effect of smoking on lung function and how lung function impairment increases the risk of respiratory infection, and then evaluate whether smoking has other direct ways to affect the risk of respiratory infection.

3.17 Nonlinear MR

Non-linear MR is a method used to evaluate the nonlinear causal relationship between exposure and outcome, especially whether the effect of exposure on the outcome is different at different exposure levels.

3.18 Time-Series MR

Time-Series MR studies how exposure factors that change over time affect outcomes. First, data on exposure and outcomes of individuals at different time points are collected, and then time series analysis techniques are used to evaluate the dynamic causal relationship between exposure and outcome. In long-term cohort studies, time series MR can be used to understand how a certain exposure affects health outcomes over time. For example, the dynamic process of how long-term alcohol intake affects liver health can be evaluated.

4 Common Used Graphs in Mendelian randomization

4.1 Scatter Plot

Scatter plots can reveal the association pattern of a single SNP with exposure and outcome. In the MR scatter plot, each point represents a specific genetic variant (SNP). The horizontal axis usually represents the strength of association of each SNP with the exposure factor, while the vertical axis represents the strength of association of each SNP with the research outcome. The distribution of these points can intuitively show the association of genetic tools with exposure and outcome. The key to this graph is to find the slope, which represents the strength of the causal relationship between exposure and outcome. If all the points are roughly arranged along a straight line, this may indicate the existence of a consistent causal relationship.

4.2 Forest Plot

Forest plots (Zhang 2024a; Zhang and Liu, 2024) can show the effect size and confidence interval of each SNP, which helps to understand the contribution of each SNP to the overall causal effect estimate. The standard forest plot shows the effect size and 95% confidence interval of each SNP on the exposure factor. The horizontal axis is usually the effect size (Zhang, 2022c), such as the risk ratio or regression coefficient, and the vertical axis lists all SNPs. Each point represents the effect estimate of a SNP, and the horizontal line represents the 95% confidence interval. If the confidence interval (Zhang, 2022a) contains the zero point (usually the vertical line with an effect size of 1), it means that the effect of the SNP is not significant. Leave-One-Out Forest Plot: remove one SNP each time and recalculate the MR estimate to evaluate its impact on the overall result. If the effect estimate changes significantly after removing a certain SNP, it may indicate that the SNP has too large individual effect on the outcome.

4.3 Funnel Plot

Funnel plots were used to examine the heterogeneity of estimated values and judge the reliability of MR estimates. In the ideal case without heterogeneity, the graph should be symmetrical, meaning that all genetic instrument variants (SNPs) are symmetrically distributed around the true causal effect estimate. The horizontal axis usually represents the effect estimate (such as the impact of each SNP on the outcome), and the vertical axis represents the precision (such as the standard error or inverse variance of the effect estimate). The symmetry of the funnel plot suggests that the results are unlikely to be affected by some unobserved confounder. If the funnel plot appears asymmetrical, it may indicate the presence of heterogeneity, indicating that some SNPs may affect the outcome through other pathways besides exposure factors, or may be the result of measurement error and chance.

5 Cases of Mendelian Randomization

Case 1

Suppose a researcher wants to study the effect of alcohol consumption (an exposure factor) on myocardial

infarction (a disease outcome).

① Find SNPs: Use GWAS data to find several SNPs that are significantly related to smoking behavior as instrumental variables.

② Exclude the influence of LD: Check the LD relationship between these SNPs, and use SNP-Clumping technology to select some representative SNPs to reduce the bias caused by LD.

③ Sensitivity analysis: During the analysis process, sensitivity analysis, such as Leave-One-Out analysis, is performed to ensure that a single SNP will not have an excessive impact on the overall estimate.

④ Pleiotropic effects: At the same time, considering the pleiotropic problem, the analysis pays special attention to those SNPs that may affect myocardial infarction through different biological pathways to ensure that the core assumptions of MR analysis are met. This method can more reliably evaluate the impact of drinking on the risk of myocardial infarction.

Case 2

In order to estimate whether weight gain leads to an increased risk of diabetes, the 2SLS method can be used first to estimate its impact on the risk of diabetes through the predicted exposure (based on the BMI value of the SNP). Then, the IVW method can be used to combine the results of multiple SNPs to obtain an overall causal effect estimate.

① Inverse-Variance Weighted (IVW). It is used for evaluating the causal relationship between independent variables and the risk of disease. The existence of the intercept term is not considered in regression, and the inverse variance is used as the weight for fitting. In the IVW hypothesis, these SNPs (as instrumental variables) are considered to have no pleiotropic effects. At the same time, considering that the results of GWAS are mostly obtained after phenotype standardization, it is considered that there is a positive proportional relationship between outcome and exposure. In general, the IVW method is used to determine whether it is a positive result.

② For example, in IVW analysis, the effect of each SNP on stomatitis and the effect of each SNP on COVID-19 (i.e., how each SNP affects COVID-19) are calculated. Then, the effect of each SNP on COVID-19 is divided by the effect of the SNP on stomatitis to obtain the causal ratio of each SNP. These causal ratios are integrated using the IVW to obtain the overall causal effect of stomatitis on COVID-19. A larger value of the overall causal effect means that stomatitis has a greater causal effect on COVID-19, and p must be less than 0.05 to be significant.

6 Disadvantages of MR Studies

Some disadvantages of MR studies are as follows:

(1) Limited applicability: MR studies are only applicable to exposure factors with suitable genetic instrumental variables, which limits their scope of application.

(2) Statistical power issues: The effect of a single genetic variant on most exposure factors may be very small, which leads to insufficient statistical power of MR analysis and increases the risk of false and negative results.

(3) Low variation interpretation: Known genetic variants can usually only explain a small part of the variation in complex phenotypes (such as BMI), and a very large sample size is required to detect weak to moderate effects.

(4) Sample size requirements: Especially in MR studies of complex phenotypes, since the variation explained by known genetic variants is usually low, a large sample size is required to achieve sufficient statistical power (Zhang, 2022b, 2024c).

(5) Influence of environmental factors: For those risk factors that are greatly affected by environmental factors, MR studies may not be sufficient to accurately estimate the contribution of genetic factors to the variation in

exposure factors.

In general, MR studies provide a powerful method to evaluate the potential causal relationship between exposure factors and diseases, especially when randomized controlled trials are not feasible or economical. However, it also has a series of limitations, especially when high genetic variation interpretation and large sample size are required. Researchers need to consider these limitations when designing MR studies and use as many methods as possible to improve the statistical power and accuracy of the studies (Zhang, 2024a-c).

Acknowledgment

I am thankful to the support of Research on New Technologies for Tannery Wastewater Treatment (2020.9-2024.9), from Zhongmeng Environmental Construction Co., Ltd., China, and Discovery and Crucial Node Analysis of Important Biological and Social Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, China.

References

All of Us Research Program Genomics Investigators. 2024. Genomic Data in the All of Us Research Program. Nature. doi: https://doi.org/10.1038/s41586-023-06957-x

Bowden J. 2016. Evaluating the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: The role of the I^2 statistic. International Journal of Epidemiology, 45(6): 1961-1974. doi: https://doi.org/10.1093/ije/dyw220

Bowden J, Smith GD, Burgess S. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology, 44: 512-525. doi: https://doi.org/10.1093/ije/dyv080

Bowden J, Smith GD, Haycock PC, Burgess S. 2016. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic Epidemiology, 40(4): 304-314. doi: https://doi.org/10.1002/gepi.21965

Burgess S, Bowden J. 2015. Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods. arXiv, 1512.04486

Burgess S, Bowden J, Dudbridge F, Thompson SG. 2016. Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. arXiv, 1606.03729

Burgess S, Butterworth AS, Thompson SG. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic Epidemiology, 37: 658-665. doi: https://doi.org/10.1002/gepi.21758

del Greco F, Minelli C, Sheehan NA, Thompson JR. 2015. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. Stat Med, 34(21): 2926-2940. doi: https://doi.org/10.1002/sim.6522

Deng MG, Liu F, Liang YH, et al. 2023. Association between frailty and depression: A bidirectional Mendelian randomization study. Science Advances, 9(38): eadi3902. doi: https://doi.org/10.1126/sciadv.adi3902

DePaolo J, Levin MG, Tcheandjieu T, et al. 2023. Relationship between ascending thoracic aortic diameter and blood pressure: A Mendelian randomization study. Arteriosclerosis, Thrombosis, and Vascular Biology, 43: 359-366. doi: https://doi.org/10.1161/ATVBAHA.122.318149

Fu Y, Xu F, Jiang L, Miao Z, Liang X, Yang J, et al. 2021. Circulating vitamin C concentration and risk of

cancers: a Mendelian randomization study. BMC Medicine, 19(1): 171. doi: https://doi.org/10.1186/s12916-021-02041-1

Grant AJ, Burgess S. 2020. Pleiotropy robust methods for multivariable Mendelian randomization. arXiv, 2008.11997

GWASLab. 2024. Mendelian Randomization Series No. 1: Basic Concepts Mendelian randomization. https://gwaslab.org/2021/06/24/mr/

Hartwig FP, Smith GD, Bowden J. 2017. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. International Journal of Epidemiology, 46(6): 1985-1998. doi: https://doi.org/10.1093/ije/dyx102

Hemani G, Tilling K, Davey Smith G. 2017. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLOS Genetics, 13(11): e1007081. doi: https://doi.org/10.1371/journal.pgen.1007081

Hernán MA, Robins JM. 2006. Instruments for causal inference an epidemiologist's dream? Epidemiology, 17(4): 360-372. doi: https://doi.org/10.1097/01.ede.0000222409.00878.37

Henry A, Gordillo-Maraón M, Finan C. 2022. Therapeutic targets for heart failure identified using proteomics and Mendelian randomization. Circulation, 145(16): 1205-1217. doi: https://doi.org/10.1161/CIRCULATIONAHA.121.056663

Kim EJ, Hoffmann TJ, Nah G, et al. 2021. Coffee consumption and incident tachyarrhythmias reported behavior, Mendelian randomization, and their interactions. JAMA Internal Medicine, 181(9): 1185-1193. doi: https://doi.org/10.1001/jamainternmed.2022.6962

Kim JY, Song M, Kim MS, et al. 2023. An atlas of associations between 14 micronutrients and 22 cancer outcomes: Mendelian randomization analyses. BMC Medicine, 21(1): 316. doi: https://doi.org/10.1186/s12916-023-03018-y

Lin Z, Xue H, Pan W. 2023. Robust multivariable Mendelian randomization based on constrained maximum likelihood. The American Journal of Human Genetics, 110(4): 592-605. doi: https://doi.org/10.1016/j.ajhg.2023.02.014

Papadimitriou N, Dimou N, Gill D, Tzoulaki I, Murphy N, Riboli E, et al. 2021. Genetically predicted circulating concentrations of micronutrients and risk of breast cancer: a Mendelian randomization study. International Journal of Cancer, 148(3): 646-653. doi: https://doi.org/10.1002/ijc.33246

Rasooly D, Peloso GM, Pereira AC, et al. 2023. Genome-wide association analysis and Mendelian randomization proteomics identify drug targets for heart failure. Nature Communications, 14: 3826. https://www.nature.com/articles/s41467-023-39253-3

Wang XT. 2023. Briefly Describe The Common Designs of Mendelian Randomization Analysis. https://mp.weixin.qq.com/s/iY4LXS4Rg_D7Y3tVd5xNTg

XM. 2024. Clinical Research and Medical Statistics. Mendelian Series in R language: Understanding Mendelian randomization in one article. https://mp.weixin.qq.com/s/tsvkrkPom1Gz9n4bn94swg

Xu S, Wang P, Fung WK, Liu Z. 2023. A novel penalized inverse-variance weighted estimator for Mendelian Randomization with applications to COVID-19 outcomes. Biometrics, 79(3): 2184-2195. doi: https://doi.org/10.1111/biom.13732

Yavorska O, Staley J. 2023. MendelianRandomization. https://cran.r-project.org/web/packages/MendelianRandomization/index.html

Zheng J, Baird D, Borges MC, et al. 2017. Recent developments in Mendelian Randomization studies. Current Epidemiology Reports, 4(4): 330-345. doi: https://doi.org/10.1007/s40471-017-0128-6

Zhu S, Kong XJ, Han FL, et al. 2024. Association between social isolation and depression: Evidence from

longitudinal and Mendelian randomization analyses. Journal of Affective Disorders, 350: 182-187. doi: https://doi.org/10.1016/j.jad.2024.01.106

Spiller W, Davies NM, Palmer TM. 2019. Software Application Profile: mrrobust — A tool for performing two-sample summary Mendelian randomization analyses. International Journal of Epidemiology, 48(3): 684-690. https://doi.org/10.1093/ije/dyy195. Mrrobust. https://github.com/remlapmot/mrrobust

Palmer T, Spiller W, Sanderson E. 2024. OneSampleMR: Useful functions for one-sample Mendelian randomization and instrumental variable analyses. https://remlapmot.github.io/OneSampleMR/

Zhang WJ. 2021a. A statistical simulation method for causality inference of Boolean variables. Network Biology, 11(4): 263-273.
http://www.iaees.org/publications/journals/nb/articles/2021-11(4)/a-method-for-causality-inference-of-Boolean-variables.pdf

Zhang WJ. 2021b. Causality inference of linearly correlated variables: The statistical simulation and regression method. Computational Ecology and Software, 11(4): 154-161.
http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/causality-inference-of-linearly-correlated-variables.pdf

Zhang WJ. 2021c. Causality inference of nominal variables: A statistical simulation method. Computational Ecology and Software, 11(4): 142-153.
http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/causality-inference-of-nominal-variables-with-statistical-simulation-method.pdf

Zhang WJ. 2022a. Confidence intervals Concepts, fallacies, criticisms, solutions and beyond. Network Biology, 12(3): 97-115.
http://www.iaees.org/publications/journals/nb/articles/2022-12(3)/confidence-intervals-fallacies-criticisms-solutions.pdf

Zhang WJ. 2022b. Dilemma of *t*-tests: Retaining or discarding choice and solutions. Computational Ecology and Software, 12(4): 181-194.
http://www.iaees.org/publications/journals/ces/articles/2022-12(4)/dilemma-of-t-tests.pdf

Zhang WJ. 2022c. *p*-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond. Computational Ecology and Software, 12(3): 80-122.
http://www.iaees.org/publications/journals/ces/articles/2022-12(3)/p-value-based-statistical-significance-tests.pdf

Zhang WJ. 2024a. MetaAnaly: The platform-independent computational tool for meta-analysis in the paradigm of new statistics. Network Biology, 14(2): 187-214.
http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/MetaAnaly.htm

Zhang WJ. 2024b. SampSizeCal: The platform-independent computational tool for sample sizes in the paradigm of new statistics. Network Biology, 14(2): 100-155.
http://www.iaees.org/publications/journals/nb/articles/2024-14(2)/5-Zhang-Abstract.asp

Zhang WJ, Liu GH. 2024. Dynamically insert the forest plot into a web page: The full Javascript codes. Computational Ecology and Software, 14(3): 168-173.
http://www.iaees.org/publications/journals/ces/articles/2024-14(3)/1-Zhang-Abstract.asp