

Mendelian Randomization Methods for Causal Inference: Estimands, Identification and Inference

Minhao Yao¹, Anqi Wang², Xihao Li^{3,4}, Zhonghua Liu^{5*}

¹ Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

² Department of Neurology, Columbia University Irving Medical Center, New York, NY, USA

³ Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

⁴ Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

⁵ Department of Biostatistics, Columbia University, New York, NY, USA

* Correspondence to: Zhonghua Liu (zl2509@cumc.columbia.edu)

Abstract

Mendelian randomization (MR) has become an essential tool for causal inference in biomedical and public health research. By using genetic variants as instrumental variables, MR helps address unmeasured confounding and reverse causation, offering a quasi-experimental framework to evaluate causal effects of modifiable exposures on health outcomes. Despite its promise, MR faces substantial methodological challenges, including invalid instruments, weak instrument bias, and design complexities across different data structures. In this tutorial review, we provide a comprehensive overview of MR methods for causal inference, emphasizing clarity of causal interpretation, study design comparisons, availability of software tools, and practical guidance for applied scientists. We organize the review around causal estimands, ensuring that analyses are anchored to well-defined causal questions. We discuss the problems of invalid and weak instruments, comparing available strategies for their detection and correction. We integrate discussions of population-based versus family-based MR designs, analyses based on individual-level versus summary-level data, and one-sample versus two-sample MR designs, highlighting their relative advantages and limitations. We also summarize recent methodological advances and software developments that extend MR to settings with many weak or invalid instruments and to modern high-dimensional omics data. Real-data applications, including UK Biobank

and Alzheimer’s disease proteomics studies, illustrate the use of these methods in practice. This review aims to serve as a tutorial-style reference for both methodologists and applied scientists.

1 Introduction

1.1 Motivation for causal inference in observational studies

A central goal of causal inference is to assess the causal relationships between variables (Holland, 1986; Pearl, 2009; Hernán and Robins, 2020; Imbens, 2024). This involves determining whether changes in one variable (the treatment or exposure) directly influence changes in another variable (the outcome). Randomized experiments are generally regarded as the gold standard study design in statistical research and practice due to their ability to facilitate causal inference (Neyman, 1923; Fisher, 1935). In essence, these experiments employ random assignment to allocate participants to treatment and control groups, which ensures that the comparison groups are balanced regarding all (measured and unmeasured) covariates except for the treatment assignment itself (Neyman, 1923; Fisher, 1935; Rubin, 1977; Imbens and Rubin, 2015). This randomization minimizes bias and enhances the internal validity of the findings. In contrast, observational (non-randomized) studies are often employed when randomization is unfeasible or unethical (Rubin, 2007; Rosenbaum et al., 2010). Observational studies aim to draw causal conclusions from real world data; however, such studies can be affected by confounding variables—factors that may influence both the treatment assignment and outcome variables, complicating establishing treatment-outcome causal relationships (Hernán and Robins, 2020).

1.2 Mendelian randomization as a natural experiment

An instrumental variable (IV) serves as a powerful tool in causal inference by leveraging a natural experiment, allowing researchers to uncover causal relationships even in the presence of unobserved confounding (Wright, 1928; Angrist et al., 1996; Angrist and Pischke, 2009; Baiocchi et al., 2014; Wooldridge, 2016). By leveraging an exogenous source of variation, such as genotype (Katan, 1986; Davey Smith and Ebrahim, 2003; VanderWeele et al., 2014) or draft lottery (Angrist, 1990), the IV approach isolates the variation in the treatment variable that is as good as randomly assigned, much like a randomized controlled trial. This natural experiment framework helps address confounding

concerns, providing more credible estimates of causal effects (Dunning, 2012). Embracing IV as a natural experiment not only strengthens empirical research but also brings us closer to the gold standard of causal inference using randomized experiments.

Mendelian randomization (MR) is a causal inference method that applies Mendel’s laws of inheritance, using genetic variants as instrumental variables to assess causal relationships between modifiable risk factors and health outcomes. By leveraging Gregor Mendel’s principles of random segregation and independent assortment of alleles, MR mimics a randomized experiment, reducing confounding biases inherent in observational studies (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008; Davey Smith and Hemani, 2014; Sanderson et al., 2022a). For illustrative purposes, we compare the designs of a randomized experiment and Mendelian randomization in Figure 1. Since genetic variants are randomly assigned at conception, much like the randomization in a clinical trial, they serve as ideal instruments to assess causal relationships between modifiable exposures (e.g., cholesterol levels) and health outcomes (e.g., heart disease) (Thanassoulis and O’Donnell, 2009; Palmer et al., 2012; Emdin et al., 2017). By leveraging the unconfounded nature of genetic inheritance, MR minimizes biases from reverse causation and unmeasured confounding, offering a robust framework for causal inference in biomedical research (Lawlor et al., 2008; Burgess et al., 2021; Sanderson et al., 2022a). This approach has been transformative in public health and medicine, helping to validate drug targets, debunk spurious associations, and guide public health policies (Haycock et al., 2016; Smith et al., 2017; Yao et al., 2024). By employing genetic variants as IVs, MR leverages the natural randomization of alleles conferred by Mendelian inheritance, transforming observational data into a quasi-experimental framework that robustly infers causal relationships.

1.3 From causal estimands to statistical inference

In this paper, we adopt the *estimand framework* to elucidate key concepts, study designs, statistical inference methods, and causal interpretations in causal inference, aiming to clarify common misconceptions and provide practical guidance for MR analysis (Lundberg et al., 2021; Kahan et al., 2024). By formally defining causal estimands, such as the average treatment effect or local average treatment effect, we align MR with the underlying causal questions of interest (Imbens and Angrist, 1994; Imbens, 2004; Lundberg et al., 2021). We then discuss how MR designs and analytical approaches target these causal estimands under certain assumptions (Haycock et al., 2016; Ference

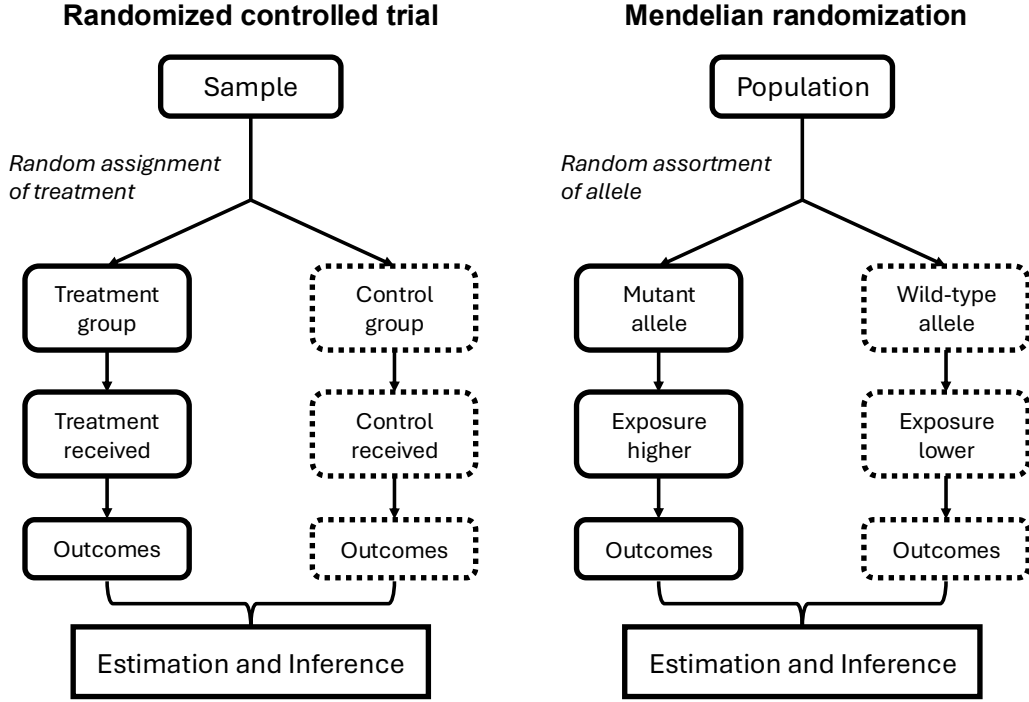


Figure 1: Comparisons of randomized controlled trial and Mendelian randomization.

et al., 2021). This estimand framework not only enhances the interpretability of MR results but also helps researchers navigate methodological challenges, such as pleiotropy and weak instrument bias, ensuring more reliable causal inference in practice (Lewis, 1999; VanderWeele, 2016; Little and Lewis, 2021; Han and Zhou, 2023; Keene et al., 2023; Kahan et al., 2024).

To formulate a coherent causal inference framework, it is essential to distinguish the following three key concepts: causal population, observed population, and sample, as illustrated in Figure 2. The *causal population* consists of all subjects in the study domain, where each subject is associated with multiple potential outcomes, one corresponding to each level of the treatment or exposure (Neyman, 1923; Rubin, 1974; Imbens and Angrist, 1994; Angrist et al., 1996; Rubin, 2005). Causal estimands are precisely defined target quantities in causal population specifying the causal effects that we are interested in. The *observed population* consists of subjects for whom only one potential outcome is realized due to the actual treatment assignment. Statistical estimands are quantities that are defined in the observed population. Causal assumptions (e.g., consistency, unconfounded-

ness) are required to establish causal identification, linking a causal estimand to its corresponding statistical estimand defined in the observed population (Neyman, 1923; Rubin, 1974; Han and Zhou, 2023). A *sample* consists of a subset drawn from the target observed population through either random or non-random selection procedures during data collection. Estimation and inference within this sample necessitates accounting for the sampling design to draw valid conclusions about the target observed population.

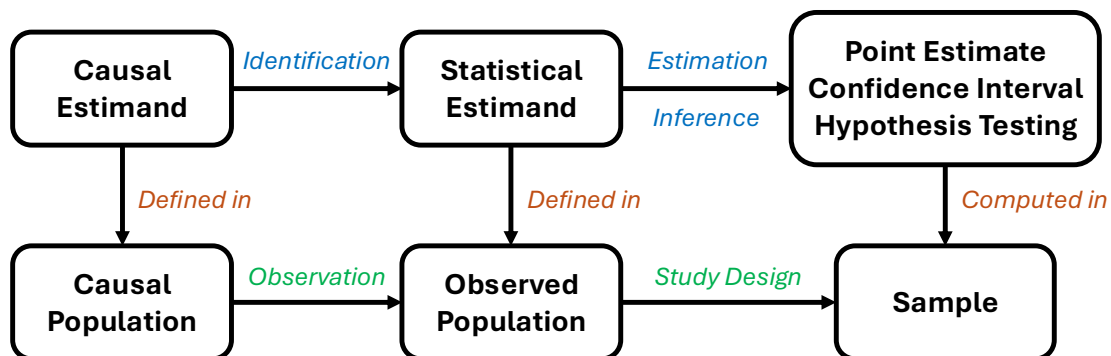


Figure 2: Conceptual flowchart bridging causal population, observed population, and sample in causal inference.

1.4 Outline of the paper

The remainder of the paper is organized as follows. Section 2 introduces how natural experiments based on genetic variation underpin MR analysis. Section 3 defines key causal estimands and the identification assumptions under the potential outcomes framework. Section 4 focuses on causal estimand under the Additive Linear Constant Effect (ALICE) model framework. Section 5 discusses identification and inference when some genetic instruments are invalid. Section 6 introduces weak instrument bias and reviews recent methods aiming at mitigating this bias. Section 7 compares MR analyses based on population-based versus family-based study designs. Section 8 discusses the use of individual-level and summary-level data in MR, highlighting their respective advantages and limitations. Section 9 compares one-sample and two-sample MR designs, focusing on the assumptions and the behavior of weak instrument bias. Section 10 outlines the procedure and

key considerations for selecting genetic instruments in MR analyses. Section 11 illustrates method comparisons using two real-data applications. Finally, Section 12 outlines future directions for MR, including binary and survival outcomes, longitudinal designs, and multivariate MR.

2 Using Genetic Variants as Instruments: Natural Experiments in Health Research

2.1 From randomized trials to natural experiments

Randomized controlled trials (RCTs) serve as the gold standard to establish the causal relationship between an exposure and an outcome (Fisher, 1935; Stolberg et al., 2004). However, some exposures are unethical or even infeasible to be randomized (Hellman and Hellman, 2017; Goldstein et al., 2018). As an alternative to RCTs, natural experiments are observational (non-randomized) studies where subjects are assigned to the treatment or control groups based on events determined by other factors beyond the control of researchers (DiNardo, 2010; Dunning, 2012; Craig et al., 2017). Natural experiments are common and have been used extensively in many fields, especially when the exposure in view cannot be ethically or practically manipulated in experimental settings (Sanson-Fisher et al., 2014; Craig et al., 2017; Leatherdale, 2019).

2.2 Genetic inheritance as a natural experiment

Mendelian randomization (MR), named after Gregor Mendel (1822–1884) who established the laws of Mendelian inheritance (Castle, 1903; Biffen, 1905; Bateson and Mendel, 2013), leverages the random assortment of genetic information during meiosis as a natural experiment to assess the causality between a modifiable exposure and an outcome of interest from observational studies (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008; Davey Smith and Hemani, 2014). In biallelic single-nucleotide polymorphisms (SNPs) where two possible alleles exist at a specific locus, the predominant allele in the population is referred to as the wild-type or major allele, while the less common allele is referred to as the variant or minor allele (International HapMap Consortium, 2005; Chari and Dworkin, 2013). During meiosis, alleles for unlinked genes are inherited independently, which is a process governed by Mendel’s Law of Independent Assortment (Castle, 1903; Biffen,

1905; Kleckner, 1996). This process forms the basis for the natural experiment underpinning MR framework (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008; Davey Smith et al., 2020).

2.3 Instrumental variable assumptions and potential violations in MR

Just as Archimedes famously claimed, “Give me a place to stand, and I will move the Earth” (Dijksterhuis, 2014), IV methods echo: “Give me a valid instrument, and I will eliminate confounding.” For reliable causal findings, genetic instruments included in the conventional MR analysis are required to be valid IVs, that is, they should satisfy the following three core IV assumptions (Lawlor et al., 2008; Didelez and Sheehan, 2007):

Assumption A1 (IV relevance). *The genetic variant is associated with the exposure.*

Assumption A2 (IV independence). *The genetic variant is not associated with unmeasured confounder of the exposure-outcome relationship.*

Assumption A3 (Exclusion restriction). *The genetic variant affects the outcome only through the exposure.*



Figure 3: Directed acyclic graphs (DAGs) that show the relationship among an instrumental variable Z , a treatment/exposure D , an outcome Y , and the unmeasured confounding U . In the right DAG, the dashed red, solid green and blue lines represent violations of the IV relevance (A1), IV independence (A2) and exclusion restriction (A3) assumptions, respectively.

However, all of the above three core IV assumptions might be violated in large-scale genetics data, as shown in Figure 3(b). Among the three core IV assumptions A1-A3, only IV relevance assumption A1 is empirically testable by selecting genetic variants associated with the exposure, while IV independence assumption A2 and exclusion restriction assumption A3 cannot be empirically verified in general (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008; Sanderson et al.,

2022a). The near violation of the testable assumption A1 may happen when genetic variants exhibit weak associations with the exposure, leading to the potential weak IV bias (Staiger and Stock, 1997; Stock et al., 2002; Burgess et al., 2011; Andrews et al., 2019). The violation of assumption A2 may arise due to the presence of population stratification, assortative mating and dynastic effect (Lawlor et al., 2008; Brumpton et al., 2020; Sanderson et al., 2021). The violation of assumption A3 may occur due to the widespread horizontal pleiotropy, where the genetic variant influences the outcome through other biological pathways that do not involve the exposure in view (Lawlor et al., 2008; Sivakumaran et al., 2011; Solovieff et al., 2013; Hemani et al., 2018a). Recently, a number of MR methods have been proposed for the identification, estimation and inference of the causal effect of interest when one or more of the three core IV assumptions are potentially violated (Bowden et al., 2015, 2017; Verbanck et al., 2018; Guo et al., 2018; Zhao et al., 2020; Sun et al., 2023a; Liu et al., 2023; Yao et al., 2024; Zhang et al., 2025). For a more comprehensive review of identification and inference with invalid IVs, see Kang et al. (2024).

3 Causal Estimands in the Potential Outcomes Framework

3.1 Definition of individual treatment effect and average treatment effect

Under the potential outcomes framework (Neyman, 1923; Rubin, 1974, 2005), let $D_i \in \{0, 1\}$ denote the binary treatment status (also referred to as the exposure) for subject i , and $Y_i(d)$ denote the potential outcome for subject i if we set $D_i = d \in \{0, 1\}$. The *individual treatment effect* (ITE) (Neyman, 1923; Rubin, 1974) for subject i is defined as

$$\text{ITE}_i = Y_i(1) - Y_i(0),$$

which quantifies the difference in the outcome for subject i under treatment versus no treatment. The ITE is generally not identifiable since we cannot observe both $Y_i(1)$ and $Y_i(0)$ for the same subject i at one time (Holland, 1986; Hernán and Robins, 2020), a concept known as the fundamental problem of causal inference.

The *average treatment effect* (ATE) (Rubin, 1974; Imbens and Angrist, 1994) is defined as

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)],$$

which measures the difference in mean outcomes had everyone been treated versus had everyone

been untreated. Likewise, for a continuous treatment, we can also define similar treatment effect for any two distinct levels: d, d' . Let Y_i be the observed outcome for subject i . Under the following assumptions: (1) the consistency assumption, i.e., $Y_i = Y_i(d)$ for $d \in \{0, 1\}$, and (2) the random assignment of treatment status, i.e., $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i$, the ATE can be identified as follows (Imbens and Angrist, 1994; Hirano et al., 2003):

$$\text{ATE} = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] = \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0],$$

where the second equation holds because of the random assignment of treatment status, and the third equation holds because of the consistency assumption. However, when the treatment status is not randomized, the ATE cannot be identified using the above formula, because $\mathbb{E}[Y_i(d)] \neq \mathbb{E}[Y_i(d)|D_i = d']$ for $d, d' \in \{0, 1\}$.

3.2 Constant treatment effect

Consider a binary instrument variable (IV) $Z_i \in \{0, 1\}$. Let $D_i(z)$ denote the binary treatment status for subject i when the IV is set to $Z_i = z \in \{0, 1\}$, and $Y_i(z, d)$ denote the potential outcome for subject i if we set $Z_i = z \in \{0, 1\}$ and $D_i = d \in \{0, 1\}$. Then, the core IV assumptions A1-A3 can be stated as:

Assumption A1' (IV relevance). $D_i \not\perp\!\!\!\perp Z_i$.

Assumption A2' (IV independence). $\{Y_i(0, D_i(0)), Y_i(1, D_i(1)), D_i(0), D_i(1)\} \perp\!\!\!\perp Z_i$.

Assumption A3' (Exclusion restriction). $Y_i(0, d) = Y_i(1, d) = Y_i(d)$ for $d \in \{0, 1\}$.

However, the above assumptions A1'-A3' alone are insufficient for the point identification of the causal effect, and hence a fourth assumption is required (Hernán and Robins, 2020). One such assumption is the following constant treatment effect (CTE) assumption (Haavelmo, 1944; Christ, 1966; Goldberger, 1972; Hernán and Robins, 2006, 2020):

Assumption A4.1 (Constant treatment effect). $Y_i(1) - Y_i(0) = \beta$ for all subjects i .

Let $Y_i(0) = y_0 + \varepsilon_i$, where $y_0 = \mathbb{E}[Y_i(0)]$, then the observed outcome Y_i can be written as the following model (Haavelmo, 1944; Goldberger, 1972; Wooldridge, 2010):

$$Y_i = y_0 + \beta D_i + \varepsilon_i.$$

Since D_i might be correlated with ε_i , regressing the outcome Y_i on the treatment D_i does not consistently estimate β . However, under the IV independence assumption [A2'](#), ε_i should be independent of the instrument Z_i , implying $\mathbb{E}[\varepsilon_i|Z_i = 0] = \mathbb{E}[\varepsilon_i|Z_i = 1]$. Substituting $\varepsilon_i = Y_i - y_0 - \beta D_i$ and solving for β , it can be shown that the CTE β equals the following usual IV estimand β_{IV} ([Wald, 1940](#); [Hernán and Robins, 2020](#)):

$$\beta_{IV} = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]}. \quad (1)$$

As illustrated in [Figure 4](#), the usual IV estimand β_{IV} is the slope of the line that captures the relationship between the expected outcome $\mathbb{E}[Y_i|Z_i]$ and the expected treatment $\mathbb{E}[D_i|Z_i]$ conditional on two levels of the IV $Z_i = z \in \{0, 1\}$.

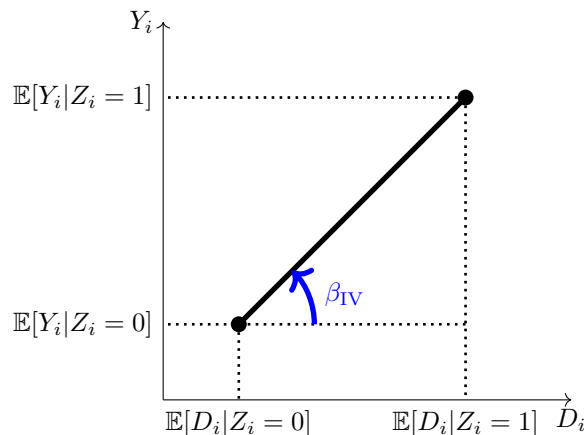


Figure 4: Graphical illustration of the usual IV estimand β_{IV} , represented by the slope of the solid line.

Conclusion 1. *Under assumptions [A1'-A3'](#) and [A4.1](#), the usual IV estimand (1) identifies the constant treatment effect.*

3.3 Average treatment effect on the treated

In this section, we consider the following additive homogeneity assumption ([Robins, 1994](#); [Hernán and Robins, 2006, 2020](#)), which is weaker than assumption [A4.1](#) and only requires that the average treatment effect is the same across different levels of Z_i for both treated and untreated groups, i.e.,

Assumption A4.2 (Additive homogeneity). $\mathbb{E}[Y_i(1) - Y_i(0) | D_i = d, Z_i = 1] = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = d, Z_i = 0]$ for $d \in \{0, 1\}$.

For binary treatment D_i and binary IV Z_i , we can express the average treatment effect among the treated across different levels of Z_i using the following saturated additive structural mean model (Robins, 1994; Hernán and Robins, 2006, 2020):

$$\mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1, Z_i] = \beta_0 + \beta_1 Z_i.$$

With the consistency assumption, the above model can be re-written as $\mathbb{E}[Y_i - Y_i(0) | D_i, Z_i] = D_i(\beta_0 + \beta_1 Z_i)$ (Hernán and Robins, 2020). Here, β_0 represents the average treatment effect among the treated individuals with $Z_i = 0$, and $\beta_0 + \beta_1$ represents the average treatment effect among the treated individuals with $Z_i = 1$. The additive homogeneity assumption A4.2 implies $\beta_1 = 0$, and then the parameter β_0 corresponds to the *average treatment effect on the treated* (ATT) (Heckman and Robb Jr, 1985; Robins, 1994; Hernán and Robins, 2006):

$$\text{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) | D_i = 1].$$

Under assumption A4.2, $\mathbb{E}[Y_i(0) | Z_i = z] = \mathbb{E}[Y_i - D_i \beta_0 | Z_i = z]$. Under the IV independence assumption A2', $\mathbb{E}[Y_i(0) | Z_i = 0] = \mathbb{E}[Y_i(0) | Z_i = 1]$. By solving the equation $\mathbb{E}[Y_i - D_i \beta_0 | Z_i = 0] = \mathbb{E}[Y_i - D_i \beta_0 | Z_i = 1]$, the parameter β_0 equals the usual IV estimand β_{IV} defined in equation (1). More recently, Liu et al. (2023) and Liu et al. (2025) have further investigated the identification of the ATT under potential violations of core IV assumptions.

Conclusion 2. *Under assumptions A1'-A3' and A4.2, the usual IV estimand (1) identifies the average treatment effect on the treated.*

3.4 Local average treatment effect

An alternative fourth identification assumption is the following monotonicity assumption (Imbens and Angrist, 1994; Angrist et al., 1996):

Assumption A4.3 (Monotonicity). $D_i(1) \geq D_i(0)$ for all subjects i .

Then, under assumptions A1'-A3' and A4.3, the usual IV estimand β_{IV} in equation (1) identifies the *local average treatment effect* (LATE) in the subgroup of compliers (i.e., subjects with $D_i(0) = 0$

and $D_i(1) = 1$) (Imbens and Angrist, 1994; Angrist et al., 1996), which is defined as:

$$\text{LATE} = \mathbb{E}[Y_i(1) - Y_i(0) | \text{Compliers}].$$

For binary IV Z_i and binary treatment status D_i , the entire population is divided into four latent subgroups, known as compliance types (Imbens and Angrist, 1994; Angrist et al., 1996), as shown in the following table:

Table 1: Four compliance types based on the values of $D_i(z)$ for $z \in \{0, 1\}$.

	$Z_i = 0$	$Z_i = 1$
Complier	$D_i(0) = 0$	$D_i(1) = 1$
Always-taker	$D_i(0) = 1$	$D_i(1) = 1$
Never-taker	$D_i(0) = 0$	$D_i(1) = 0$
Defier	$D_i(0) = 1$	$D_i(1) = 0$

The compliance type of subject i is generally latent since we cannot observe both $D_i(0)$ and $D_i(1)$ at one time. Under the monotonicity assumption A4.3 there are no defiers in the population. Additionally, the IV Z_i has no effect on Y_i in the subgroups of always-takers or never-takers, as the treatment status D_i is fixed across different levels of Z_i in these two subgroups. Therefore, the IV Z_i can only affect the outcome Y_i in the subgroup of compliers, meaning that the usual IV estimand identifies the LATE in compliers.

Conclusion 3. *Under assumptions A1'-A3' and A4.3, the usual IV estimand (1) identifies the local average treatment effect in compliers.*

3.5 Identification of average treatment effect

Let U_i denote the unmeasured confounders. Wang and Tchetgen Tchetgen (2018) proposes the following two no-interaction assumptions for the identification of ATE:

Assumption A4.4 (No additive $U_i - Z_i$ interaction). *There is no additive $U_i - Z_i$ interaction in $\mathbb{E}[D_i | Z_i, U_i]$, that is, $\mathbb{E}[D_i | Z_i = 1, U_i] - \mathbb{E}[D_i | Z_i = 0, U_i] = \mathbb{E}[D_i | Z_i = 1] - \mathbb{E}[D_i | Z_i = 0]$.*

Assumption A4.5 (No additive $U_i - d$ interaction). *There is no additive $U_i - d$ interaction in $\mathbb{E}[Y_i(d) | U_i]$, that is, $\mathbb{E}[Y_i(1) - Y_i(0) | U_i] = \mathbb{E}[Y_i(1) - Y_i(0)]$.*

Intuitively, assumption A4.4 rules out modification of instrument-treatment association by U_i on the additive scale, whereas assumption A4.5 rules out modification of the effect of the treatment on the outcome by U_i on the additive scale. In addition, Wang and Tchetgen Tchetgen (2018) also imposes the following assumption for confounding control:

Assumption A5 (Sufficiency of U_i for confounding control). $Y_i(d) \perp\!\!\!\perp (D_i, Z_i) \mid U_i$.

Assumption A5 requires that, conditional on the unmeasured confounders U_i , the potential outcomes are independent of the treatment status and the instrument. This assumption is originally formulated by Richardson and Robins (2014).

Conclusion 4. *Under assumptions A1'-A3' and A5, together with either assumption A4.4 or A4.5, the usual IV estimand (1) identifies the average treatment effect.*

3.6 Comparisons of the causal estimands

We compare the four causal estimands (CTE, ATT, LATE and ATE) in Table 2. The identification of all four causal estimands requires assumptions A1'-A3', which are therefore referred to as the *core IV assumptions* (Angrist et al., 1996; Baiocchi et al., 2014). However, these core IV assumptions A1'-A3' are insufficient for point identification, and the four causal estimands differ in their additional identification assumption (Hernán and Robins, 2020). The CTE relies on the strong constant treatment effect assumption A4.1, which posits that the treatment effect is the same across all subjects. By contrast, the ATT relies on the additive homogeneity assumption A4.2, a weaker identification assumption than A4.1, and corresponds to the average treatment effect among those who actually received the treatment. The LATE is identified by imposing the monotonicity assumption A4.3 as the additional identification assumption, and is interpreted as the average treatment effect in the subgroup of compliers, i.e., subjects who would receive the treatment if assigned to it and not receive it otherwise. Finally, the ATE, which captures the average treatment effect for the entire population, is identified under either the no-interaction assumption A4.4 or A4.5, together with the confounding control assumption A5.

Table 2: Comparison of identification assumptions and interpretations for CTE, ATE, ATT, and LATE.

Causal estimand	Identification assumptions	Causal interpretation
CTE	IV relevance A1' ;	Constant treatment effect of the treatment versus control across all subjects.
	IV independence A2' ;	
	Exclusion restriction A3' ;	
	Constant treatment effect A4.1 .	
ATT	IV relevance A1' ;	Average treatment effect of the treatment versus control specifically for subjects that actually received the treatment.
	IV independence A2' ;	
	Exclusion restriction A3' ;	
	Additive homogeneity A4.2 .	
LATE	IV relevance A1' ;	Average treatment effect of the treatment versus control specifically for the compliers.
	IV independence A2' ;	
	Exclusion restriction A3' ;	
	Monotonicity A4.3 .	
ATE	IV relevance A1' ;	Average treatment effect of the treatment versus control for the entire population.
	IV independence A2' ;	
	Exclusion restriction A3' ;	
	No-interaction A4.4 or A4.5 ; Confounding control A5 .	

4 Causal Estimand Defined in the ALICE Model

4.1 Definition of the ALICE model

Having introduced the key causal estimands (CTE, ATT, LATE, and ATE) and their identification under IV assumptions, we now turn to a specific causal model that has become the workhorse of Mendelian randomization studies. The Additive Linear Constant Effects (ALICE) model ([Holland, 1988](#)) formalizes the constant treatment effect assumption [A4.1](#) introduced in Section 3, providing a simple yet widely used framework for characterizing causal effects in MR. Let $D_i \in \mathbb{R}$ denote the exposure of subject i , and $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})^\top \in \mathbb{R}^p$ denote the vector of p genetic instru-

ments of subject i . Let $Y_i(\mathbf{z}, d) \in \mathbb{R}$ denote the continuous potential outcome if subject i had $\mathbf{Z}_i = \mathbf{z} = (z_1, \dots, z_p)^\top$ and $D_i = d$. Then, for two possible values of instruments \mathbf{z}, \mathbf{z}' and the exposure d, d' , we assume the following model (Holland, 1986; Kang et al., 2016; Guo et al., 2018):

$$\begin{aligned} Y_i(\mathbf{z}', d') - Y_i(\mathbf{z}, d) &= \beta(d' - d) + \sum_{j=1}^p \psi_j(z'_j - z_j), \\ \mathbb{E}[Y_i(\mathbf{0}, 0) | \mathbf{Z}_i] &= \sum_{j=1}^p \phi_j Z_{ij}, \end{aligned} \tag{2}$$

where $\beta \in \mathbb{R}$ is the primary causal parameter of interest, representing the constant effect of a one-unit change in the exposure on the outcome across all subjects in the whole population. The parameter $\psi_j \in \mathbb{R}$ quantifies the degree of violation of the exclusion restriction assumption A3 for j th instrument, capturing the direct effects of the instrument on the potential outcome. The parameter $\phi_j \in \mathbb{R}$ quantifies the degree of violation of the IV independence assumption A2 for j th genetic instrument. Under the IV independence assumption A2, the instruments \mathbf{Z}_i should be independent of the baseline potential outcome $Y_i(\mathbf{0}, 0)$ in the absence of confounding. However, in model (2), the relationship between \mathbf{Z}_i and $Y_i(\mathbf{0}, 0)$ is modeled through ϕ_1, \dots, ϕ_p , allowing for potential violations of assumption A2 (Angrist et al., 1996; Kang et al., 2016; Guo et al., 2018). Let $\pi_j = \psi_j + \phi_j$ for $j = 1, \dots, p$, and $\varepsilon_i = Y_i(0, \mathbf{0}) - \mathbb{E}[Y_i(0, \mathbf{0}) | \mathbf{Z}_i]$, then under the consistency assumption, we have the following observed outcome model (Small, 2007; Kang et al., 2016; Guo et al., 2018):

$$Y_i = \beta D_i + \sum_{j=1}^p \pi_j Z_{ij} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i | \mathbf{Z}_i) = 0. \tag{3}$$

The causal effect β in model (3) cannot be estimated by directly fitting a usual linear regression because the exposure D_i might be correlated with the error term ε_i . Moreover, in model (3), the parameter $\pi_j \in \mathbb{R}$ encodes the degrees of violation of assumptions A2 and A3 for j th genetic instrument. Specifically, if the j th genetic instrument satisfies both the exclusion restriction assumption and IV independence assumption, then $\pi_j = 0$; otherwise, if $\pi_j \neq 0$, the j th genetic instrument violates at least one of the exclusion restriction assumption or IV independence assumption (Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2021; Guo, 2023; Sun et al., 2023a; Kang et al., 2024; Zhang et al., 2025). Therefore, we say the j th genetic instrument is a valid IV if $\pi_j = 0$, and an invalid IV if $\pi_j \neq 0$. In Section 5, we will discuss the identification and inference in the presence of invalid IVs under the ALICE model framework.

Remark 1. *Kang et al. (2016)* extends model (2) to incorporate heterogeneous causal effect as follows:

$$Y_i(\mathbf{z}', d') - Y_i(\mathbf{z}, d) = \beta_i(d' - d) + \sum_{j=1}^p \psi_j(z'_j - z_j),$$

where β_i is the individual causal effect of subject i . Let $\beta = \mathbb{E}[\beta_i]$ be the average causal effect, the observed outcome model (3) becomes

$$Y_i = \beta D_i + \sum_{j=1}^p \pi_j Z_{ij} + (\beta_i - \beta) D_i + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i | \mathbf{Z}_i) = 0.$$

This model reduces to the constant causal effect model (3) if $(\beta_i - \beta)$ is independent of D_i given \mathbf{Z}_i (Kang et al., 2016).

4.2 ALICE model is widely used in MR studies

To model the relationship between a continuous exposure and genetic instruments, we further consider a linear model between the exposure D_i and the genetic instruments \mathbf{Z}_i (Angrist et al., 1996; Small, 2007; Guo et al., 2018):

$$D_i = \sum_{j=1}^p \gamma_j Z_{ij} + \delta_i, \quad \mathbb{E}(\delta_i | \mathbf{Z}) = 0, \quad (4)$$

where γ_j represents the IV strength of j th genetic instrument. Note that the error term δ_i in the exposure model (4) might be correlated with the error term ε_i in the outcome model (3) due to unmeasured confounders. By plugging in the exposure model (4) into the outcome model (3), we can obtain the reduced-form model for the outcome (Small, 2007; Guo et al., 2018):

$$Y_i = \sum_{j=1}^p \Gamma_j Z_{ij} + e_i, \quad \mathbb{E}(e_i | \mathbf{Z}_i) = 0, \quad (5)$$

where $\Gamma_j = \beta\gamma_j + \pi_j$, and $e_i = \beta\delta_i + \varepsilon_i$.

Most summary-level MR methods for continuous outcomes build upon the ALICE model. For a single genetic instrument j , according to the equation $\Gamma_j = \beta\gamma_j + \pi_j$, the ratio estimand is defined as follows (Burgess et al., 2013; Slob and Burgess, 2020):

$$\beta_j = \frac{\Gamma_j}{\gamma_j} = \beta + \frac{\pi_j}{\gamma_j}. \quad (6)$$

When j th genetic instrument is a valid IV (i.e., $\pi_j = 0$), the ratio estimand β_j equals the causal effect β in the ALICE model. In summary-level MR analysis, the ratio estimate of j th SNP is

defined as $\hat{\beta}_j = \hat{\Gamma}_j / \hat{\gamma}_j$, where $\hat{\Gamma}_j$ and $\hat{\gamma}_j$ are marginal estimates of Γ_j and γ_j in genome-wide association studies (GWAS) summary statistics.

4.3 Practical limitations and interpretational caveats of the ALICE model

Although the ALICE model provides a useful framework for causal inference with potentially invalid instrumental variables, it is subject to several important limitations and interpretational caveats. First, the causal effect defined in the ALICE model should be interpreted as a constant treatment effect of a one-unit increase in the exposure on the outcome (Holland, 1986; Small, 2007; Kang et al., 2016). However, the constant treatment effect assumption may not hold in real-world settings where treatment effects may vary across subjects (Angrist, 2004; Powers et al., 2018; Künzel et al., 2019). Second, the ALICE model assumes linearity not only in the causal effect but also in the violation of core IV assumptions. This linearity assumption might be violated when the underlying relationships are nonlinear, such as in complex genetic architectures (Veitia et al., 2013; Guindo-Martínez et al., 2021; Sun et al., 2023a). Therefore, when applying the ALICE model, it is essential to carefully assess the plausibility of its assumptions within the context of the study and to interpret the resulting estimates with appropriate caution.

5 Identification and Inference in the Presence of Invalid IVs

5.1 Additional identification assumption under the ALICE model

When j th genetic IV is a valid instrument (i.e., $\pi_j = 0$), the ALICE model in Section 4 enables causal identification through the ratio Γ_j / γ_j . However, when invalid instruments are present without prior knowledge of IV validity status, the causal effect β in model (3) becomes non-identifiable. This is because the parameters in models (4) and (5) should satisfy the following equation system:

$$\Gamma_j = \beta \gamma_j + \pi_j, \quad j = 1, \dots, p, \quad (7)$$

where the IV-exposure associations $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^\top$ and IV-outcome associations $\boldsymbol{\Gamma} = (\Gamma_1, \dots, \Gamma_p)^\top$ can be identified using population ordinary least squares (OLS) through $\boldsymbol{\gamma} = \mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top)^{-1} \mathbb{E}(\mathbf{Z}_i D_i)$ and $\boldsymbol{\Gamma} = \mathbb{E}(\mathbf{Z}_i \mathbf{Z}_i^\top)^{-1} \mathbb{E}(\mathbf{Z}_i Y_i)$. Given $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$, there are p equations with $p+1$ unknown parameters $\{\beta, \pi_1, \dots, \pi_p\}$, resulting in an underdetermined equation system that precludes unique identification of $\{\beta, \pi_1, \dots, \pi_p\}$, and renders models (3) and (4) under-identified. Consequently, additional

assumptions regarding $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)^\top$ are required to address the identifiability issue. Below, we list three commonly adopted additional identification assumptions in the ALICE model.

Assumption A6 (Instrument strength independent of the direct effect (InSIDE)). *The IV-exposure association γ_j is asymptotically independent of the degree of IV invalidity π_j as the number of genetic IVs p goes to infinity.*

From the equation system (7), we have

$$\frac{\text{Cov}(\mathbf{\Gamma}, \boldsymbol{\gamma})}{\text{Var}(\boldsymbol{\gamma})} = \beta + \frac{\text{Cov}(\boldsymbol{\pi}, \boldsymbol{\gamma})}{\text{Var}(\boldsymbol{\gamma})}.$$

Under the InSIDE assumption, $\text{Cov}(\boldsymbol{\pi}, \boldsymbol{\gamma}) \rightarrow 0$ as $p \rightarrow \infty$, yielding the identification of β (Kolesár et al., 2015; Bowden et al., 2015). The InSIDE assumption has been adopted in some summary-level MR methods, for example, MR-Egger (Bowden et al., 2015), random-effects inverse-variance weighed (IVW) method (Bowden et al., 2017), and MR using the Robust Adjusted Profile Score (MR-RAPS) (Zhao et al., 2020).

Assumption A7 (Majority rule). *The number of valid genetic IVs is more than half of the relevant genetic IVs.*

The majority rule assumption is a sufficient condition for the identification of β under the ALICE model framework (Han, 2008; Kang et al., 2016). Formally, let $\mathcal{S} = \{j : \gamma_j \neq 0\}$ denote the set of all relevant genetic IVs with non-zero IV-exposure associations, and $\mathcal{V} = \{j \in \mathcal{S} : \gamma_j \neq 0 \text{ and } \pi_j = 0\}$ denote the set of all valid genetic IVs, then the majority rule assumption can be expressed as $|\mathcal{V}| > \frac{1}{2}|\mathcal{S}|$. Under the majority rule assumption, more than half of the ratio estimand β_j defined in equation (6) equal the true causal effect β , since they arise from valid instruments with $\pi_j = 0$ (Bowden et al., 2016). A natural identification strategy is therefore to find the median of $\{\beta_j\}_{j \in \mathcal{S}}$ (Bowden et al., 2016). Some MR methods based on the majority rule assumption include Some Invalid Some Valid IV Estimator (sisVIVE) (Kang et al., 2016), weighted median method (Bowden et al., 2016), and MR Pleiotropy RESidual Sum and Outlier test (MR-PRESSO) (Verbanck et al., 2018).

Assumption A8 (Plurality rule). *Valid genetic IVs form the largest group among relevant genetic IVs based on the ratio of IV-outcome association to IV-exposure association.*

As shown in [Guo et al. \(2018\)](#), the plurality rule assumption is weaker than the majority rule assumption, and is a sufficient condition for the identification of causal effect β under the ALICE model framework. Formally, the plurality rule assumption can be expressed as $|\mathcal{V}| > \max_{c \neq 0} |\{j \in \mathcal{S} : \frac{\pi_j}{\gamma_j} = c\}|$. Under this assumption, the true causal effect β corresponds to the mode of the distribution of $\{\beta_j\}_{j \in \mathcal{S}}$ ([Hartwig et al., 2017](#)). Thus, the identification of the causal effect β can be achieved by detecting the largest group of ratio estimands, either by direct mode estimation ([Hartwig et al., 2017](#)) or via voting-based procedures ([Guo et al., 2018](#); [Yao et al., 2024](#)). The plurality rule assumption is also termed as the ZERo Modal Pleiotropy Assumption (ZEMPA) ([Hartwig et al., 2017](#)), and is adopted in MR methods including the mode-based estimation ([Hartwig et al., 2017](#)), Two-Stage Hard Thresholding (TSHT) ([Guo et al., 2018](#)), MRMix ([Qi and Chatterjee, 2019](#)), the contamination mixture method ([Burgess et al., 2020](#)), Confidence Interval method for Instrumental Variable (CIIV) ([Windmeijer et al., 2021](#)), and MR with valid IV Selection and Post-selection Inference (MR-SPI) ([Yao et al., 2024](#)).

Remark 2. Equation (7) demonstrates that, in the presence of unknown instrument invalidity, the causal effect β in the ALICE model is generally not identifiable. Assumptions A6-A8 restore identification by imposing different constraints on $\boldsymbol{\pi}$, which encodes the degree of violation of assumptions A2 and A3. Because assumptions A6-A8 cannot be empirically tested with data, a common practice is to employ multiple MR methods relying on different assumptions as sensitivity analyses to evaluate the robustness of MR findings.

5.2 Alternative identification strategies beyond the ALICE model framework

[Sun et al. \(2023a\)](#) considers the following model under the potential outcomes framework:

$$Y_i(\mathbf{z}, d') - Y_i(\mathbf{0}, d) = \beta(d' - d) + \psi(\mathbf{z}),$$

where $\psi(\cdot)$ is an unknown function that satisfies $\psi(\mathbf{0}) = 0$, which allows for arbitrary interactions among the direct effects of the instruments on the outcome. The ALICE model is a special case of the above model by specifying $\psi(\mathbf{z}) = \sum_{j=1}^p \psi_j z_j$ and $\mathbb{E}[Y_i(\mathbf{0}, 0) | \mathbf{Z}_i] = \sum_{j=1}^p \phi_j Z_{ij}$ ([Sun et al., 2023a](#)). Under this model, the set of valid instruments is defined as the index set $\mathcal{V} \subseteq \{1, \dots, p\}$ such that $\psi(\mathbf{Z}_i) = \psi(\mathbf{Z}_{i,-\mathcal{V}})$ and $\mathbb{E}[Y_i(\mathbf{0}, 0) | \mathbf{Z}_i] = \mathbb{E}[Y_i(\mathbf{0}, 0) | \mathbf{Z}_{i,-\mathcal{V}}]$ holds almost surely, where

$\mathbf{Z}_{i,-\mathcal{V}} = (Z_{ij} : j \notin \mathcal{V})$ (Sun et al., 2023a). When all p instruments are mutually independent and there are at least v valid instruments, Sun et al. (2023a) shows that the causal effect β in the above model is the unique solution to the following equation:

$$\mathbb{E}[\mathbf{h}^{[v]}(\mathbf{Z}_i)(Y_i - \beta D_i)] = \mathbf{0},$$

where the function $\mathbf{h}^{[v]}(\mathbf{Z}_i) \in \mathbb{R}^m$ with $m = \sum_{j=0}^{v-1} \binom{p}{j}$ represents all demeaned interactions involving at least $p - v + 1$ instruments. For example, when there are $p = 2$ instruments and there is at least $v = 1$ valid instrument, then there is only one demeaned interaction $(Z_{i1} - \mu_1)(Z_{i2} - \mu_2)$ involving at least 2 instruments, where μ_1 and μ_2 are the expectations of Z_{i1} and Z_{i2} , respectively. When $(Z_{i1} - \mu_1)(Z_{i2} - \mu_2)$ is associated with the exposure D_i , then β is the unique solution to

$$\mathbb{E}[(Z_{i1} - \mu_1)(Z_{i2} - \mu_2)(Y_i - \beta D_i)] = 0.$$

Remark 3. As discussed in Kang et al. (2024), Sun et al. (2023a) uses higher-order interactions to create “new” instruments from the p instruments, which can capture possible nonlinear effects of instruments on the exposure. In contrast, Guo et al. (2022) applies machine learning algorithms to explore nonlinear effects of instruments on the exposure.

Tchetgen Tchetgen et al. (2021) proposes the MR G-Estimation under No Interaction with Unmeasured Selection (MR-GENIUS) approach that leverages heteroscedasticity in the exposure to identify the causal effect. Specifically, Tchetgen Tchetgen et al. (2021) considers the following model:

$$\mathbb{E}[Y_i | D_i, \mathbf{Z}_i, U_i] = \beta_y(U_i)D_i + \alpha_y(U_i, \mathbf{Z}_i) + \eta_y(U_i),$$

$$\mathbb{E}[D_i | \mathbf{Z}_i, U_i] = \alpha_d(U_i, \mathbf{Z}_i) + \eta_d(U_i),$$

where $\beta_y(\cdot)$ is an unspecified function of the unmeasured confounder U_i , which affects both the exposure D_i and the outcome Y_i , and is independent of the instruments \mathbf{Z}_i . The terms $\eta_y(\cdot)$ and $\eta_d(\cdot)$ are two unspecified functions of U_i , and $\alpha_y(\cdot)$ and $\alpha_d(\cdot)$ are two unspecified functions of (U_i, \mathbf{Z}_i) satisfying $\alpha_y(U, \mathbf{0}) = \alpha_d(U, \mathbf{0}) = 0$. When the exposure D_i is heteroscedastic, i.e., $\text{Var}(D_i | \mathbf{Z}_i)$ varies with the instruments \mathbf{Z}_i , Tchetgen Tchetgen et al. (2021) shows that the average causal effect $\beta = \mathbb{E}[\beta_y(U_i)]$ in the above model is the unique solution to the following equation:

$$\mathbb{E}[(\mathbf{Z}_i - \mathbb{E}(\mathbf{Z}_i))(D_i - \mathbb{E}(D_i | \mathbf{Z}_i))(Y_i - \beta D_i)] = \mathbf{0}.$$

Remark 4. As discussed in [Tchetgen Tchetgen et al. \(2021\)](#), MR-GENIUS might not perform well when $\text{Var}(D_i|Z_i)$ is only weakly dependent on the instruments Z_i . [Ye et al. \(2024\)](#) extends MR-GENIUS to allow for many weak invalid instruments.

By leveraging heteroscedasticity in the outcome, [Liu et al. \(2023\)](#) proposes the Mendelian Randomization Mixed-Scale Treatment Effect Robust Identification (MR-MiSTERI) approach for the average treatment effect on the treated (ATT). In this section, we focus on the case where both the treatment and the possibly invalid genetic instrument are binary; see [Liu et al. \(2023\)](#) for extensions. Specifically, MR-MiSTERI relies on the following three identification assumptions:

Assumption B1 (Homogeneous ATT). *The ATT does not vary with the possibly invalid IV on the additive scale, i.e., $\mathbb{E}[Y_i(z, d = 1) - Y_i(z, d = 0) \mid D_i = 1, Z_i = z] = \beta$.*

Assumption B2 (Homogeneous confounding bias on the odds ratio scale). *$OR(Y_i(0) = y_0, D_i = d \mid Z_i = z) = \exp(\xi dy_0)$, where ξ quantifies the magnitude of confounding bias.*

Assumption B3 (Outcome heteroscedasticity). *Define $\varepsilon_i = Y_i - \mathbb{E}(Y_i \mid D_i, Z_i)$ and suppose that $\varepsilon_i \mid D_i, Z_i \sim N(0, \sigma^2(Z_i))$, then $\sigma^2(Z_i)$ must vary with the genetic instrument Z_i .*

Remark 5. Assumption B3 can be empirically testable through genome-wide variance quantitative trait loci (vQTL) analyses ([Paré et al., 2010](#); [Wang et al., 2019](#)). As shown in [Paré et al. \(2010\)](#), gene-gene (GxG) and/or gene-environment (GxE) interactions can result in genotype-dependent changes in trait variance, providing direct evidence for heteroscedasticity in quantitative traits.

Under assumptions B1-B3, the confounding bias parameter ξ and the causal effect β are uniquely identified by

$$\begin{aligned}\xi &= \frac{D_i(Z_i = 1) - D_i(Z_i = 0)}{\sigma^2(Z_i = 1) - \sigma^2(Z_i = 0)}, \\ \beta &= D_i(Z_i) - \frac{D_i(Z_i = 1) - D_i(Z_i = 0)}{\sigma^2(Z_i = 1) - \sigma^2(Z_i = 0)} \sigma^2(Z_i), \quad Z_i = 0, 1,\end{aligned}$$

and the estimates for ξ and β can be obtained by replacing the unknown quantities with the sample counterparts in observed data.

5.3 Inference for the causal effect

Following [Kang et al. \(2024\)](#), inference for the causal effect in MR analysis can be broadly classified into two methodological paradigms: *pointwise inference* and *uniformly valid inference*.

Pointwise inference constructs the confidence interval for the causal effect either by: (1) calculating the standard error of the causal effect estimate directly from asymptotic distribution or resampling techniques (e.g., bootstrap) (Bowden et al., 2015, 2016, 2017; Hartwig et al., 2017; Qi and Chatterjee, 2019; Zhao et al., 2020; Burgess et al., 2020); or (2) selecting valid instruments from candidate genetic variants (e.g., using voting procedure or outlier detection test) and subsequently constructing confidence intervals using the selected subset (Guo et al., 2018; Verbanck et al., 2018; Windmeijer et al., 2021; Yao et al., 2024). The latter approach relies on the correct selection of valid IVs; when IV selection error occurs in finite samples, it might lead to poor coverage performance (Guo, 2023).

The second paradigm, *uniformly valid inference*, constructs CIs that remain robust to finite-sample instrument selection errors (Kang et al., 2022; Guo, 2023; Yao et al., 2024; Kang et al., 2024). Specifically, Kang et al. (2022) proposes taking the union of confidence intervals constructed from subsets of instruments passing the J test (Hansen, 1982); however, this procedure is computationally costly when the number of candidate genetic IVs is large (Kang et al., 2024). Alternatively, Guo (2023) and Yao et al. (2024) first construct “pseudo CIs” through grid-search using resampled IV-exposure and IV-outcome associations, and then construct the final robust CI by taking the union of these pseudo CIs across resamples. Uniformly valid inference generally constructs wider CIs than pointwise inference, which is a trade-off for the guaranteed finite-sample coverage level (Kang et al., 2024).

6 Weak Identification in the ALICE Model Framework

6.1 The presence of weak identification bias

In this section, we examine the bias introduced by weak instruments, i.e., instruments that are only weakly associated with the exposure, in the ALICE model framework. We begin by assuming that all instruments satisfy assumptions A2 and A3, i.e., $\pi_j = 0$ for all $j \in \{1, \dots, p\}$ in model (3). In this case, two-stage least squares (2SLS) is commonly employed to estimate the causal effect β (Angrist and Pischke, 2009; Wooldridge, 2016). Specifically, in the first stage, we fit an OLS regression of the exposure \mathbf{D} on the genetic instruments \mathbf{Z} to obtain the following fitted exposure

values $\hat{\mathbf{D}}$:

$$\hat{\mathbf{D}} = \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{D}.$$

In the second stage, the outcome \mathbf{Y} is regressed on these fitted exposures $\hat{\mathbf{D}}$ to obtain the following 2SLS estimator of the causal effect:

$$\hat{\beta}_{2\text{SLS}} = \left(\hat{\mathbf{D}}^\top \hat{\mathbf{D}} \right)^{-1} \hat{\mathbf{D}}^\top \mathbf{Y}. \quad (8)$$

Assume that the error terms satisfy $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $\delta_i \sim N(0, \sigma_\delta^2)$, and denote $\text{Cov}(\delta_i, \varepsilon_i) = \sigma_{\delta, \varepsilon}$, [Rothenberg \(1984\)](#) provides the following analytic expression for the bias of 2SLS estimator

$$\mu \left(\hat{\beta}_{2\text{SLS}} - \beta \right) = \frac{\sigma_\varepsilon}{\sigma_\delta} \frac{\eta_\varepsilon + \xi_{\varepsilon, \delta} / \mu}{1 + 2\eta_\delta / \mu + \xi_{\delta, \delta} / \mu^2}, \quad (9)$$

where $\mu^2 = \gamma^\top \mathbf{Z}^\top \mathbf{Z} \gamma / \sigma_\delta^2$ is the *concentration parameter* that measures the instrument strength ([Stock et al., 2002](#); [Stock and Yogo, 2002](#)). Here, $\eta_\varepsilon = (\sigma_\varepsilon \sqrt{\gamma^\top \mathbf{Z}^\top \mathbf{Z} \gamma})^{-1} \gamma^\top \mathbf{Z}^\top \varepsilon$ and $\eta_\delta = (\sigma_\delta \sqrt{\gamma^\top \mathbf{Z}^\top \mathbf{Z} \gamma})^{-1} \gamma^\top \mathbf{Z}^\top \delta$ are two standard normal random variables with correlation $\sigma_{\delta, \varepsilon} / (\sigma_\varepsilon \sigma_\delta)$, $\xi_{\varepsilon, \delta} = \delta^\top \mathbf{Z} (\sigma_\varepsilon \sigma_\delta \mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \varepsilon$ and $\xi_{\delta, \delta} = \delta^\top \mathbf{Z} (\sigma_\delta^2 \mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \delta$ are two quadratic forms of normal random variables that do not depend on the sample size n .

Remark 6. From equation (9), as the concentration parameter μ^2 goes to infinity, $\mu \left(\hat{\beta}_{2\text{SLS}} - \beta \right)$ has an asymptotic distribution of $N(0, \sigma_\varepsilon^2 / \sigma_\delta^2)$ ([Rothenberg, 1984](#)). Therefore, the concentration parameter μ^2 can be thought of as an effective sample size ([Stock et al., 2002](#); [Stock and Yogo, 2002](#)). When instruments are strong, the concentration parameter μ^2 increases proportionally to the sample size n ([Andrews and Stock, 2005](#)).

6.2 Measurement of weak identification

[Staiger and Stock \(1997\)](#) proposes to assess the instrument strength using the following F -statistic:

$$\hat{F} = \frac{\hat{\gamma}^\top \mathbf{Z}^\top \mathbf{Z} \hat{\gamma}}{p \hat{\sigma}_\delta^2},$$

where $\hat{\gamma}$ denotes the coefficient vector by fitting an OLS regression of the exposure on the instruments, and $\hat{\sigma}_\delta^2$ is the corresponding residual variance. This statistic provides a test of the joint null hypothesis $\gamma = \mathbf{0}$ in the first-stage regression of 2SLS, and is therefore commonly referred to as the “first-stage F -statistic” ([Staiger and Stock, 1997](#); [Stock and Yogo, 2002](#)). Under the null hypothesis $\gamma = \mathbf{0}$ and within the weak instrument asymptotics framework (i.e., IV strengths $\{\gamma_j\}_{j=1}^p$

shrink at a $1/\sqrt{n}$ rate (Staiger and Stock, 1997), $p\hat{F}$ converges in distribution to a noncentral chi-squared random variable with p degrees of freedom and noncentrality parameter μ^2 (Stock and Yogo, 2002). As suggested by Staiger and Stock (1997), $\hat{F} < 10$ is the rule-of-thumb threshold for weak instruments.

6.3 Addressing weak identification bias in MR Studies

Recent developments in IV and MR literature have advanced methodologies to address weak instrument bias. For example, Ye et al. (2021) proposes dIVW, a debiased version of the inverse-variance weighted (IVW) estimator, which is robust to many weak IVs using two-sample summary-level data. Xu et al. (2023) further develops the penalized IVW (pIVW) estimator by using a penalization approach to prevent the denominator of dIVW estimator to be too close to zero. Mikusheva and Sun (2022) defines weak identification in the context of many instruments, where the number of instruments p grows with the sample size n , and introduces a jackknifed version of the Anderson-Rubin test statistic (Anderson and Rubin, 1949) that is robust to weak identification with many instruments and heteroscedasticity in both the exposure and the outcome. Ye et al. (2024) proposes GENIUS-MAWII (G-Estimation under No Interaction with Unmeasured Selection leveraging MAny Weak Invalid IVs), which simultaneously addresses the challenges of many weak instruments and widespread horizontal pleiotropy in MR studies.

7 Population-based versus Family-based Design

7.1 Population-based MR design

Population-based designs (e.g., cohort studies and case-control studies) include unrelated subjects from the target population (Szklo, 1998; Nkomo et al., 2006; Rothman et al., 2008). A cohort study is an observational research method where a group of people with a shared characteristic, called a cohort, is followed over time to observe health outcomes or the development of a disease after a specific exposure (Szklo, 1998; Rothman et al., 2008). These studies identify groups based on factors like exposure to a risk factor and then compare the outcomes in exposed versus unexposed individuals to determine associations. There are two main types of cohort studies: prospective cohort studies (Sedgwick, 2013), which follow the group into the future, and retrospective cohort

studies (Sedgwick, 2014), which look back at historical data. For example, UK Biobank (UKB) is a large-scale, prospective cohort study that includes over 500,000 participants aged 40-69 at recruitment across the United Kingdom (Sudlow et al., 2015; Bycroft et al., 2018). In contrast, a case-control study retrospectively compares subjects with a specific outcome (cases) to those without (controls) (Schlesselman, 1982; Breslow, 1996; Rothman et al., 2008). Since outcomes are often rare, cases are oversampled and controls are undersampled in case-control studies, resulting in a sample that may not reflect the target population (Wan et al., 2021). Nevertheless, logistic regression can still provides valid association estimates on the odds ratio scale in case-control studies (Prentice and Pyke, 1979). MR studies leveraging population-based designs benefit from large sample sizes and wide coverage. For example, UKB has genotyped over 500,000 individuals, providing high statistical power to detect modest associations between exposures and outcomes (Sudlow et al., 2015; Bycroft et al., 2018). However, population-based designs are susceptible to confounding by population stratification, assortative mating, dynastic effects, and selection bias (Lawlor et al., 2008; Brumpton et al., 2020; Sanderson et al., 2022a). To mitigate these biases, researchers often apply methods such as adjusting for principal components, matching, or the use of negative controls (Price et al., 2006; Stuart, 2010; Lipsitch et al., 2010; Sanderson et al., 2021).

7.2 Family-based MR design

Family-based designs in MR use data from related individuals, typically sibling pairs or parent-offspring trios, to draw causal conclusions within families (Davies et al., 2019; Brumpton et al., 2020; Howe et al., 2022; LaPierre et al., 2023; Davies et al., 2024). By comparing genetically and demographically similar relatives, these designs inherently control for many confounding factors (Kong et al., 2018; Howe et al., 2022; Davies et al., 2024). For example, in a sibling-based MR, one sibling can serve as a control for shared family background (Howe et al., 2022). This within-family comparison helps to eliminate bias due to population stratification, dynastic effects, and assortative mating, which may otherwise confound population-based MR findings (Brumpton et al., 2020; Howe et al., 2022). However, family-based study designs typically have smaller sample sizes, limiting statistical power to detect causal relationships (Chen and Abecasis, 2007; Brumpton et al., 2020; Howe et al., 2022). Moreover, these designs require more complex modeling to properly account for family structures and relatedness among subjects (Brumpton et al., 2020; Hwang et al., 2021; LaPierre et al., 2023). Despite these challenges, family-based MR has proven valuable; for

example, within-family analyses have shown that effects of height and BMI on educational attainment, observed in population-based MR, are substantially attenuated when shared familial factors are controlled (Brumpton et al., 2020).

7.3 Choosing between population and family-based MR designs

In summary, population-based and family-based MR designs offer complementary advantages and trade-offs, as summarized in Table 3. Population-based MR relies on broad, large-scale samples, offering greater statistical power and generalizability but is more susceptible to bias from population stratification, assortative mating, dynastic effects, and selection bias. Family-based MR inherently accounts for many shared genetic and environmental factors, enhancing the reliability of causal conclusions, but typically involves smaller samples and more complex modeling. In practice, combining insights from both study designs can provide more robust causal findings.

Table 3: Comparison of population-based and family-based MR designs

Feature	Population-based design	Family-based design
Definition	Includes subjects sampled from the target population (e.g., UK Biobank).	Includes genetically related subjects (e.g., siblings or parent-offspring trios).
Study types	Includes designs such as cohort studies (prospective or retrospective) and case-control studies.	Include designs such as sibling designs and parent-offspring trio designs.
Key strengths	Generally larger sample size; high statistical power and broader generalizability.	Inherent control for population stratification, dynastic effects, and shared environment.
Limitations	Susceptible to population stratification, assortative mating, dynastic bias, and selection bias.	Smaller sample sizes; requires more complex modeling to account for family structures and relatedness.

Continued on next page

Table 3 – continued from previous page

Feature	Population-based design	Family-based design
IV assumption violation risks	More susceptible to violations of the IV independence and exclusion restriction assumptions.	Less susceptible to IV assumption violations due to within-family comparison.
Bias mitigation	Adjustment for principal components, matching, and negative control outcomes.	Natural control for genetic and environmental confounding within families.

8 Individual-level versus Summary-level Data

8.1 MR methods using individual-level data

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ be the vector of continuous outcomes of n subjects, $\mathbf{D} = (D_1, \dots, D_n)^\top$ be the vector of continuous exposures, and $\mathbf{Z} = (Z_{ij})_{n \times p}$ be the matrix of genetic instruments, where Z_{ij} is the genotype of j th genetic instrument of subject i . *Individual-level data MR* utilizes the dataset $\{\mathbf{Y}, \mathbf{D}, \mathbf{Z}\}$ containing individual-level measurements of outcomes, exposures, and genotypes. When individual-level data are available and all instruments are valid, two-stage least squares (2SLS) estimator $\hat{\beta}_{2\text{SLS}}$ defined in equation (8) is commonly employed to estimate the causal effect β for continuous outcomes under the ALICE model framework (Angrist and Pischke, 2009; Wooldridge, 2016). When all genetic instruments satisfy the three core IV assumptions A1-A3, the 2SLS estimator $\hat{\beta}_{2\text{SLS}}$ is consistent to the causal effect β (Wooldridge, 2016). When some genetic instruments violate one or more of the core IV assumptions, alternative individual-level data MR methods have been developed to estimate the causal effect β , for example, sisVIVE (Kang et al., 2016), TSHT (Guo et al., 2018), MR-GENIUS (Tchetgen Tchetgen et al., 2021), MR-MiSTERI (Liu et al., 2023), MRSquare (Sun et al., 2023a), GENIUS-MAWII (Ye et al., 2024), and MR-MAGIC (Zhang et al., 2025). See Sections 5 and 6, as well as Kang et al. (2024), for details.

8.2 MR methods using summary-level data

In contrast, *summary-level* MR utilizes summary statistics of marginal IV-exposure and IV-outcome associations derived from individual-level data (often not directly accessible) to perform causal inference. Let $\mathbf{Z}_{\cdot j} = (Z_{1j}, \dots, Z_{nj})^\top$ denote the genotype vector of j th genetic instrument, then the marginal estimates of γ_j and Γ_j are obtained through marginal regressions of \mathbf{D} and \mathbf{Y} on $\mathbf{Z}_{\cdot j}$, respectively:

$$\begin{aligned}\hat{\gamma}_j &= (\mathbf{Z}_{\cdot j}^\top \mathbf{Z}_{\cdot j})^{-1} \mathbf{Z}_{\cdot j}^\top \mathbf{D}, \\ \hat{\Gamma}_j &= (\mathbf{Z}_{\cdot j}^\top \mathbf{Z}_{\cdot j})^{-1} \mathbf{Z}_{\cdot j}^\top \mathbf{Y}.\end{aligned}$$

Then, the ratio estimator using j th genetic instrument is

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}.$$

When all p genetic instruments are valid, the following inverse-variance weighted (IVW) estimator (Burgess et al., 2013) combines ratio estimators from each genetic instrument:

$$\hat{\beta}_{\text{IVW}} = \frac{\sum_{j=1}^p \hat{\gamma}_j^2 \hat{\sigma}_{\Gamma,j}^{-2} \hat{\beta}_j}{\sum_{j=1}^p \hat{\gamma}_j^2 \hat{\sigma}_{\Gamma,j}^{-2}},$$

where $\hat{\sigma}_{\Gamma,j}$ is the standard error of $\hat{\Gamma}_j$. The IVW estimator upweights genetic instruments with stronger IV-exposure associations (i.e., larger $\hat{\gamma}_j^2$) and more precise IV-outcome associations (i.e., smaller $\hat{\sigma}_{\Gamma,j}^2$). In addition, $\hat{\beta}_{\text{IVW}}$ consistently estimates β when all instruments are valid and mutually independent. When some genetic instruments violate the core IV assumptions, the IVW estimator is no longer consistent. To address this, several summary-level data MR methods have been proposed to obtain robust causal effect estimates even in the violation of core IV assumptions (Bowden et al., 2015, 2016; Verbanck et al., 2018; Zhao et al., 2019b; Burgess et al., 2020; Xu et al., 2023; Yao et al., 2024). We provide a list of commonly used software implementations with links in Supplementary Section S2.

8.3 Comparison of individual-level and summary-level data in MR

Compared to summary-level data MR methods, individual-level data MR offers distinct methodological advantages in modeling nonlinear biological relationships and addressing invalid instrument

issues. First, individual-level data allow for the characterization of nonlinear relationship among genetic instruments, exposures and outcomes (Hall and Horowitz, 2005; Veitia et al., 2013; Staley and Burgess, 2017; Guo et al., 2022; Sulc et al., 2022; Sun et al., 2023a). For example, Guo et al. (2022) explicitly models both nonlinear IV-exposure associations and nonlinear violations of assumptions (A2) and (A3), and proposes the Two-Stage Curvature Identification (TSCI) method to identify and estimate the causal effect of interest using individual-level data. In contrast, summary-level data are calculated using linear or generalized linear models, and thus MR based on summary-level data lacks the capacity to detect nonlinear associations (Burgess, 2024). Second, individual-level data enable more flexible approaches for handling invalid IVs. For example, Sun et al. (2023a) proposes a class of G-estimators for the causal effect in the presence of multiple potentially invalid IVs by leveraging gene-gene interactions, and Liu et al. (2023) proposes novel identification assumptions for the average treatment effect on the treated (ATT) with a possibly invalid IV.

Conversely, summary-level data MR provides significant practical advantages in genomic data. First, publicly accessible genome-wide association studies (GWAS) summary statistics have become increasingly abundant (Yang et al., 2012; Zhu et al., 2016; Buniello et al., 2019), overcoming privacy concerns and logistic burdens that often limit access to individual-level genetic data (Kaufman et al., 2009; Naveed et al., 2015; Harmanci and Gerstein, 2016). Second, GWAS consortia routinely combine data from hundreds of thousands of participants, significantly enhancing statistical power to detect causal relationships (Swerdlow et al., 2016). In addition, platforms like MR-Base (Hemani et al., 2018b) further streamline analysis by enabling efficient harmonization of exposure and outcome summary statistics across multiple GWAS datasets.

9 One-sample versus Two-sample Design

9.1 Overview and conceptual differences

To facilitate comparison between one-sample and two-sample MR designs, we adopt the ALICE framework for two independent datasets (Angrist and Krueger, 1992, 1995; Zhao et al., 2019b). Let $s \in \{1, 2\}$ index two independent samples with sample sizes $n^{(1)}$ and $n^{(2)}$, respectively. Within each sample s , let $Y_i^{(s)}$ denote the outcome, $D_i^{(s)}$ denote the exposure, and $\mathbf{Z}_i^{(s)} = (Z_{i1}^{(s)}, \dots, Z_{ip}^{(s)})^\top \in \mathbb{R}^p$ denote the vector of p genetic instruments for subject i . The data $\{Y_i^{(s)}, D_i^{(s)}, \mathbf{Z}_i^{(s)}\}_{i=1}^{n^{(s)}}$ are generated

according to:

$$D_i^{(s)} = \sum_{j=1}^p \gamma_j^{(s)} Z_{ij}^{(s)} + \delta_i^{(s)},$$

$$Y_i^{(s)} = \beta^{(s)} D_i^{(s)} + \sum_{j=1}^p \pi_j^{(s)} Z_{ij}^{(s)} + \varepsilon_i^{(s)},$$

with error terms satisfying $\mathbb{E}(\delta_i^{(s)} | \mathbf{Z}_i^{(s)}) = 0$ and $\mathbb{E}(\varepsilon_i^{(s)} | \mathbf{Z}_i^{(s)}) = 0$. We now state the study objectives in one-sample and two-sample MR designs as follows (Zhao et al., 2019b):

- **Objective in one-sample MR design:** Given either the individual-level data $\{Y_i^{(s)}, D_i^{(s)}, \mathbf{Z}_i^{(s)}\}_{i=1}^{n^{(s)}}$, or the corresponding summary-level data of IV-exposure and IV-outcome associations from sample s , how to estimate the causal effect $\beta^{(s)}$?
- **Objective in two-sample MR design:** Given the individual-level data $\{Y_i^{(1)}, \mathbf{Z}_i^{(1)}\}_{i=1}^{n^{(1)}}$ from the first sample and $\{D_i^{(2)}, \mathbf{Z}_i^{(2)}\}_{i=1}^{n^{(2)}}$ from the second sample (or their corresponding summary-level data), how to estimate the causal effects $\beta^{(1)}$ and/or $\beta^{(2)}$?

As noted in Zhao et al. (2019b), to enable the identification and estimation of the causal effect in two-sample designs, we further need to impose the following assumptions (Angrist and Krueger, 1992, 1995; Zhao et al., 2019b):

Assumption C1 (Homogeneity in parameters). $\beta^{(1)} = \beta^{(2)} = \beta$, $\gamma_j^{(1)} = \gamma_j^{(2)} = \gamma_j$, and $\pi_j^{(1)} = \pi_j^{(2)} = \pi_j$ for $j = 1, \dots, p$.

Assumption C2 (Homogeneity in the distribution of error terms). $(\delta_i^{(1)}, \varepsilon_i^{(1)}) \stackrel{d}{=} (\delta_{i'}^{(2)}, \varepsilon_{i'}^{(2)})$ for $i \in \{1, \dots, n^{(1)}\}$ and $i' \in \{1, \dots, n^{(2)}\}$, where $\stackrel{d}{=}$ indicates that the random vectors have the same distribution.

Remark 7. Under assumptions C1 and C2, the only source of heterogeneity between the two samples arises from differences in the distribution of instruments (Zhao et al., 2019b). In the context of MR, such heterogeneity may reflect differences in genetic ancestry, sampling design, or genotyping platforms, which can lead to differences in allele frequencies or linkage disequilibrium patterns.

9.2 Weak IV biases in one-sample and two-sample MR estimations

In this section, we focus on the comparison between one-sample and two-sample MR designs, and for simplicity we assume all genetic instruments are valid IVs, i.e., $\pi_j^{(1)} = \pi_j^{(2)} = 0$ for $j = 1, \dots, p$.

Under assumptions [C1](#) and [C2](#) and by assuming all genetic instruments are valid IVs, the above ALICE model becomes

$$\begin{aligned} D_i^{(s)} &= \sum_{j=1}^p \gamma_j Z_{ij}^{(s)} + \delta_i^{(s)}, \\ Y_i^{(s)} &= \beta D_i^{(s)} + \varepsilon_i^{(s)}, \end{aligned}$$

and the reduced-form outcome model becomes

$$Y_i^{(s)} = \sum_{j=1}^p \Gamma_j Z_{ij}^{(s)} + e_i^{(s)},$$

where $\Gamma_j = \beta \gamma_j$ and $e_i^{(s)} = \beta \delta_i^{(s)} + \varepsilon_i^{(s)}$. Within sample s , let $\mathbf{Y}^{(s)} = (Y_1^{(s)}, \dots, Y_{n^{(s)}}^{(s)})^\top$ be the vector of outcomes, $\mathbf{D}^{(s)} = (D_1^{(s)}, \dots, D_{n^{(s)}}^{(s)})^\top$ be the vector of exposures, and $\mathbf{Z}^{(s)} = (\mathbf{Z}_1^{(s)}, \dots, \mathbf{Z}_{n^{(s)}}^{(s)})^\top \in \mathbb{R}^{n^{(s)} \times p}$ be the matrix of genetic instruments. For simplicity, we further assume that genetic instruments are (1) standardized such that $\mathbb{E}(Z_{ij}^{(s)}) = 0$ and $\text{Var}(Z_{ij}^{(s)}) = 1$ for $j = 1, \dots, p$ ([Bulik-Sullivan et al., 2015](#)), and (2) mutually independent after LD clumping ([Purcell et al., 2007](#)). We now analyze the weak instrument biases of one-sample and two-sample 2SLS estimators under this setup.

Let $\widehat{\beta}_{2\text{SLS}}^{(s)}$ denote the one-sample 2SLS estimator using individual-level data from sample s . According to [Hahn and Hausman \(2002\)](#), the weak instrument bias of $\widehat{\beta}_{2\text{SLS}}^{(s)}$ can be approximated as follows:

$$\mathbb{E} \left(\widehat{\beta}_{2\text{SLS}}^{(s)} \right) - \beta \approx \frac{\sigma_{\delta, \varepsilon}}{n^{(s)} \|\boldsymbol{\gamma}\|^2 / p + \sigma_\delta^2},$$

where $\sigma_{\delta, \varepsilon}$ is the covariance between $\boldsymbol{\delta}^{(s)}$ and $\boldsymbol{\varepsilon}^{(s)}$, σ_δ^2 is the variance of $\boldsymbol{\delta}^{(s)}$, and $\|\boldsymbol{\gamma}\|^2 = \sum_{j=1}^p \gamma_j^2$. We also provide the derivation of this approximate bias in Supplementary Section [S1](#). On the other hand, the approximate bias of ordinary least square (OLS) estimator $\widehat{\beta}_{\text{OLS}}^{(s)}$ using sample s is

$$\mathbb{E} \left(\widehat{\beta}_{\text{OLS}}^{(s)} \right) - \beta \approx \frac{\sigma_{\delta, \varepsilon}}{\|\boldsymbol{\gamma}\|^2 + \sigma_\delta^2}.$$

With weak instruments, both the one-sample 2SLS and OLS estimators are biased when the error terms in the exposure and outcome models are correlated, i.e., $\sigma_{\delta, \varepsilon} \neq 0$. Importantly, the direction of the bias for $\widehat{\beta}_{2\text{SLS}}^{(s)}$ is the same as that for $\widehat{\beta}_{\text{OLS}}^{(s)}$, implying that the one-sample 2SLS estimator tends to be biased towards the OLS estimator with weak instruments.

In the two-sample design, the IV-exposure associations are first estimated in the first sample and then used to construct fitted exposures in the second sample. The causal effect is subsequently estimated by regression the outcome on these fitted exposures in the second sample. This estimation

strategy is also known as the Split-Sample Instrumental Variable (SSIV) estimation ([Angrist and Krueger, 1995](#)). Under our setting, the two-sample 2SLS estimator, denoted as $\hat{\beta}_{\text{SSIV}}$, has the following approximation ([Angrist and Krueger, 1995](#)):

$$\mathbb{E}(\hat{\beta}_{\text{SSIV}}) \approx \beta \times \frac{\|\gamma\|^2}{\|\gamma\|^2 + p\sigma_\delta^2/n^{(1)}}.$$

This expression shows that the two-sample 2SLS estimator is attenuated toward zero by a factor that depends on the first-stage sample size $n^{(1)}$, the number of IVs p , the IV strengths, and the variance of the error in exposure model σ_δ^2 . Unlike in the one-sample setting, where weak instruments bias the 2SLS estimator toward the confounded OLS estimate, the two-sample 2SLS estimator is biased toward zero when instruments are weak.

Then, we consider the case where only marginal association estimates and their standard errors are available from a single sample. Specifically, in sample s , the marginal estimates of IV-exposure and IV-outcome associations of j th genetic instrument are $\hat{\gamma}_j^{(s)} = \left((\mathbf{Z}_{\cdot j}^{(s)})^\top \mathbf{Z}_{\cdot j}^{(s)}\right)^{-1} (\mathbf{Z}_{\cdot j}^{(s)})^\top \mathbf{D}^{(s)}$ and $\hat{\Gamma}_j^{(s)} = \left((\mathbf{Z}_{\cdot j}^{(s)})^\top \mathbf{Z}_{\cdot j}^{(s)}\right)^{-1} (\mathbf{Z}_{\cdot j}^{(s)})^\top \mathbf{Y}^{(s)}$, respectively. Let $\hat{\sigma}_{\Gamma,j}^{(s)}$ denote the standard error of $\hat{\Gamma}_j^{(s)}$. Then, the one-sample IVW estimator for the causal effect β using summary statistics from sample s is given by

$$\hat{\beta}_{\text{IVW}}^{(s)} = \frac{\sum_{j=1}^p \hat{\Gamma}_j^{(s)} \hat{\gamma}_j^{(s)} / (\hat{\sigma}_{\Gamma,j}^{(s)})^2}{\sum_{j=1}^p (\hat{\gamma}_j^{(s)})^2 / (\hat{\sigma}_{\Gamma,j}^{(s)})^2},$$

and the bias of $\hat{\beta}_{\text{IVW}}^{(s)}$ can be approximated as

$$\mathbb{E}(\hat{\beta}_{\text{IVW}}^{(s)}) - \beta \approx \frac{\sigma_{\delta,\varepsilon}}{n^{(s)}} \times \frac{\sum_{j=1}^p 1/(\hat{\sigma}_{\Gamma,j}^{(s)})^2}{\sum_{j=1}^p (\gamma_j^2 + (\sigma_{\gamma,j}^{(s)})^2) / (\hat{\sigma}_{\Gamma,j}^{(s)})^2},$$

where $(\sigma_{\gamma,j}^{(s)})^2$ is the variance of $\hat{\gamma}_j^{(s)}$. As with the one-sample 2SLS estimator, this weak instrument bias arises due to the correlation between error terms in the exposure and outcome models, and tends toward the OLS estimator.

Finally, we consider the two-sample IVW estimator that combines the the IV-exposure association estimates $\{\hat{\gamma}_j^{(1)}\}_{j=1}^p$ from the first sample and the IV-outcome association estimates $\{\hat{\Gamma}_j^{(2)}\}_{j=1}^p$ from the second sample ([Burgess et al., 2013](#); [Bowden et al., 2017](#)), which is given by

$$\hat{\beta}_{\text{IVW}}^{(1,2)} = \frac{\sum_{j=1}^p \hat{\Gamma}_j^{(2)} \hat{\gamma}_j^{(1)} / (\hat{\sigma}_{\Gamma,j}^{(2)})^2}{\sum_{j=1}^p (\hat{\gamma}_j^{(1)})^2 / (\hat{\sigma}_{\Gamma,j}^{(2)})^2},$$

and the expectation of the two-sample IVW estimator can be approximated as follows (Zhao et al., 2020; Ye et al., 2021):

$$\mathbb{E} \left(\widehat{\beta}_{\text{IVW}}^{(1,2)} \right) \approx \beta \times \frac{\sum_{j=1}^p \gamma_j^2 / (\widehat{\sigma}_{\Gamma,j}^{(2)})^2}{\sum_{j=1}^p (\gamma_j^2 + (\sigma_{\gamma,j}^{(1)})^2) / (\widehat{\sigma}_{\Gamma,j}^{(2)})^2}.$$

This reveals that the two-sample IVW estimator is biased toward zero with weak instruments, similar to the two-sample 2SLS estimator.

Conclusion 5. *In the presence of weak IVs, one-sample MR estimators tends to be biased towards the confounded OLS estimator, whereas two-sample MR estimators tends to be biased towards zero.*

Remark 8. *To handle the weak IV bias in two-sample summary-level data MR analysis, Xu et al. (2023) proposes a novel penalized inverse-variance weighted (pIVW) estimator that adjusts the IVW estimator through a penalized likelihood approach.*

9.3 Advantages, limitations, and recommendations for practice

In summary, both the 2SLS estimator with individual-level data and the IVW estimator with summary-level data in one-sample study designs are biased toward the OLS estimator with weak instruments. In contrast, both 2SLS and IVW estimators in two-sample study designs tends to be biased toward zero. In addition, in two-sample study design, the identification of the causal effect and interpretation of the estimate require assumptions C1 and C2 on parameters and error terms in addition to core IV assumptions. Neither study design is universally superior; the choice between one-sample and two-sample study design depends on data availability (single versus two independent datasets), interpretability (population-specific versus generalizable estimates), and bias trade-offs (weak IV bias toward confounded OLS estimates versus toward zero).

10 Selecting Genetic IVs using the Anna Karenina Principle

A critical step in MR analysis is the selection of appropriate genetic variants to serve as instruments. In practice, variants are often chosen based on their strength of association with the exposure, typically using GWAS summary statistics. Thresholds such as $p < 5 \times 10^{-8}$ (genome-wide significance) or $p < 1 \times 10^{-6}$ are commonly applied, although the precise cut-off varies across studies (Panagiotou et al., 2012; Swerdlow et al., 2016; Kanai et al., 2016; Sanderson et al., 2022a).

The more difficult challenge lies in distinguishing valid from invalid instruments when the core assumptions may be violated. Different MR methods address this issue by introducing additional identification assumptions. For example, MR-PRESSO (Verbanck et al., 2018) identifies valid IVs under the *majority rule*.

Motivated by the Anna Karenina Principle (AKP) (Yao et al., 2024), which states that "all happy families are alike, but every unhappy family is unhappy in its own way," we view valid instruments as a coherent group that share the same properties, while each invalid instrument may fail validity in a distinct manner. Building on this intuition, MR-SPI adopts the *plurality rule assumption* A8, which requires only that the largest group of instruments corresponds to the valid set, even if valid instruments do not form a majority.

The SPI procedure operationalizes this idea by identifying the largest cluster of variants that conform to the IV assumptions, while allowing for heterogeneous violations among the rest. This strategy reflects the AKP perspective: validity is uniform, but invalidity can be idiosyncratic. By leveraging this asymmetry, MR-SPI provides a principled and robust approach to selecting instruments from GWAS summary data. More specifically, the selection steps are as follows (Figure 5):

- Step 1. **Linkage disequilibrium (LD) clumping.** Obtain a set of approximately independent genetic variants by removing those in high LD with one another using PLINK (Purcell et al., 2007).
- Step 2. **Selecting relevant genetic IVs.** From the LD-clumped genetic variants, retain those showing strong associations with the exposure using a specific p -value thresholding (e.g., 5×10^{-8}).
- Step 3. **Voting on genetic IV validity.** For each relevant genetic IV j , compute the ratio estimate $\hat{\beta}_j = \hat{\Gamma}_j / \hat{\gamma}_j$. Then, treating $\hat{\beta}_j$ as the true causal effect, for every other relevant genetic IV k , calculate its estimated degrees of violation of assumptions A2 and A3, defined as $\hat{\pi}_j^{[k]} = \hat{\Gamma}_k - \hat{\beta}_j \hat{\gamma}_k$. A small $|\hat{\pi}_j^{[k]}|$ suggests that k th genetic IV is also likely to be a valid IV by assuming j th genetic IV is valid, and thus k th genetic IV "votes for" j th genetic IV to be valid.
- Step 4. **Selecting valid genetic IVs from the obtained voting matrix.** Select valid genetic IVs by finding the maximum clique of the voting matrix that encodes whether two relevant IVs mutually vote for each other to be valid.

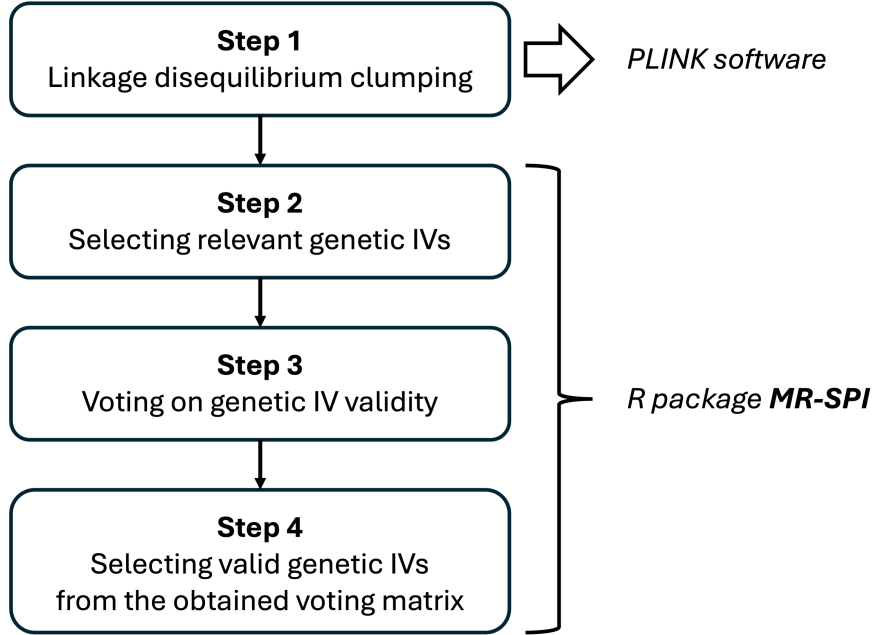


Figure 5: IV selection procedure of MR-SPI.

In practice, Step 1 can be performed via PLINK by applying a predefined pairwise correlation threshold (e.g., $r^2 < 0.01$) within a specified genomic window (e.g., 1Mb). Steps 2-4 can be implemented using the R package **MR-SPI** (<https://github.com/MinhaoYaooo/MR-SPI>). The instruments selected through these steps can then be used for downstream MR analyses (e.g., IVW method). If the plurality rule assumption A8 holds, this procedure can help exclude invalid IVs and enhance the robustness of MR findings.

Several considerations are important when selecting genetic instruments. Ideally, the sample used for IV selection should be independent of the samples used for estimating causal effect to minimize the winner’s curse (Jiang et al., 2023; Ma et al., 2023). For example, Zhao et al. (2019a) proposes a three-sample MR design to eliminate the bias due to the winner’s curse. It is also advisable to use external resources, such as PhenoScanner (Kamat et al., 2019), to screen candidate genetic IVs for associations with potential confounders or secondary traits, thereby improving instrument validity. Careful attention to these issues enhances the reliability and reproducibility of MR findings.

11 Applications in Real Datasets

11.1 Application 1: assessing the causal effect of body mass index on diastolic blood pressure using one-sample individual-level data from UK Biobank

In this section, we apply several one-sample MR methods to assess the causal effect of body mass index (BMI) on diastolic blood pressure (DBP). This analysis utilizes data from the UK Biobank (UKB) cohort study, a biomedical database comprising genetic and phenotypic information from approximately 500,000 UK participants (Sudlow et al., 2015; Bycroft et al., 2018). Participants who reported using anti-hypertensive medication or had missing data were excluded, resulting in a final sample of 254,502 individuals. Following Sun et al. (2023a), we selected the top 10 independent single-nucleotide polymorphisms (SNPs) most strongly associated with BMI after applying linkage disequilibrium (LD) clumping with $r^2 < 0.01$. These SNPs are rs1558902, rs6567160, rs543874, rs13021737, rs10182181, rs2207139, rs11030104, rs10938397, rs13107325, and rs3810291.

We compare the following methods for estimating the causal effect β : (1) two-stage least squares (2SLS) (Wooldridge, 2016); (2) Two-Stage Hard Thresholding (TSHT) (Guo et al., 2018); (3) Confidence Interval method for Instrumental Variable (CIIV) (Windmeijer et al., 2021); (4) Some Invalid Some Valid IV Estimator (sisVIVE) (Kang et al., 2016); (5) MRSquare (Sun et al., 2023a); MR Mixed-Scale Treatment Effect Robust Identification (MR-MiSTERI) (Liu et al., 2023); and MR with MAny weak Genetic Interactions for Causality (MR-MAGIC) (Zhang et al., 2025). For sisVIVE, we choose the tuning parameter via 10-fold cross-validation. For MRSquare, we set the minimum number of valid instruments to be 6. The results are summarized in Figure 6.

From Figure 6, all methods suggest a positive causal effect of BMI on DBP, with point estimates ranging from 0.1677 to 0.4037. The 2SLS method yields the smallest estimate ($\hat{\beta}_{2SLS} = 0.1677$; 95% CI: 0.0456–0.2898), likely due to the inclusion of invalid instruments in the analysis. Unlike 2SLS, which assumes all instruments are valid, the other methods account for potential invalid IVs in various ways. Notably, TSHT, CIIV, and sisVIVE implement procedures to select valid instruments from candidate ones. In this application, TSHT identifies two invalid IVs (rs10182181 and rs13107325), CIIV identifies three (rs10182181, rs13107325, and rs3810291), and sisVIVE identifies one (rs13107325). Instrument rs13107325 is consistently identified as an invalid IV by all three

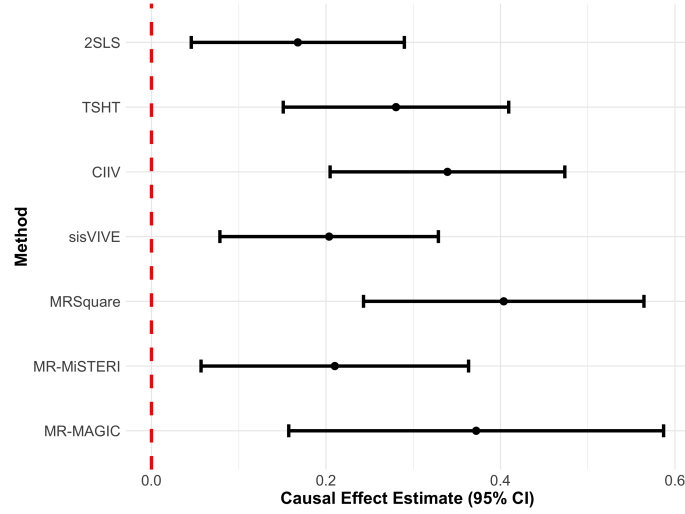


Figure 6: Point estimates and 95% confidence intervals for the causal effect of body mass index (BMI) on diastolic blood pressure (DBP) in the UK Biobank data, obtained using different one-sample MR methods.

methods.

11.2 Application 2: performing xMR analysis to identify plasma proteins associated with the risk of Alzheimer’s disease using two-sample summary-level data

The increasingly available large-scale multi-omics data (e.g., epigenomics, transcriptomics, proteomics, and metabolomics data) enable us to perform omics MR (xMR, firstly coined in [Yao et al. \(2024\)](#)) to detect putative causal omics biomarkers for complex traits and diseases, thereby uncovering the underlying causal mechanisms. For a detailed, step-by-step tutorial on implementing commonly used xMR methods, please refer to [Yao and Liu \(2025\)](#).

In this section, we apply several two-sample MR methods to perform xMR analysis, aiming to identify putative causal plasma proteins associated with the risk of Alzheimer’s disease. For the exposure, we use UK Biobank Pharma Proteomics Project (UKB-PPP) summary statistics on 1,463 plasma proteins measured in 54,306 individuals ([Sun et al., 2023b](#)). For the outcome, we use

summary statistics from a meta-analysis of GWASs for clinically diagnosed AD and AD by proxy, comprising 455,258 samples in total (Jansen et al., 2019). Genetic instruments for each protein are selected by applying a Bonferroni-corrected threshold of $p\text{-value} < 3.40 \times 10^{-11}$, followed by LD clumping at threshold $r^2 < 0.01$, as described in Sun et al. (2023b). We compare the following two-sample MR methods in this application: (1) inverse-variance weighted (IVW) method (Bowden et al., 2017); (2) MR-Egger regression (Bowden et al., 2015); (3) MR using the Robust Adjusted Profile Score (MR-RAPS) (Zhao et al., 2020); (4) MR Pleiotropy RESidual Sum and Outlier test (MR-PRESSO) (Verbanck et al., 2018); (5) the weighted median method (Bowden et al., 2016); (6) the mode-based estimation (Hartwig et al., 2017); (7) MRMix (Qi and Chatterjee, 2019); (8) the contamination mixture method (Burgess et al., 2020); and (9) MR with valid IV Selection and Post-selection Inference (MR-SPI) (Yao et al., 2024).

Proteins identified to be significantly associated with Alzheimer’s disease after Bonferroni correction (Bland and Altman, 1995) are summarized in Figure S1. In Figure S1(A), we report the number of significant plasma proteins detected by each method, which ranges from 0 (MR-PRESSO) to 14 (MRMix). MR-PRESSO detects no significant proteins, likely because the small number of candidate IVs per protein limits its power to perform the outlier test and detect invalid instruments. In Figure S1(B), we list the 11 plasma proteins identified by at least two methods. Notably, the seven proteins identified by MR-SPI correspond to the top seven proteins ranked by the number of supporting methods. For a more detailed discussion of these seven proteins including gene ontology analysis and AlphaFold3-based structural prediction (Abramson et al., 2024), please refer to Yao et al. (2024).

12 Future directions

12.1 Binary and survival outcomes

In epidemiological research, the binary and survival outcomes are common, yet current MR methods predominantly focus on continuous outcomes. For one-sample MR with individual-level data, establishing identification conditions for MR analysis with binary or survival outcomes is essential, particularly in the presence of invalid instruments (Clarke and Windmeijer, 2010; Deng et al., 2023; Liu et al., 2025). For two-sample MR with summary-level data, the current standard

practice is to directly apply existing two-sample MR methods under the ALICE model framework for continuous outcomes to analyze summary statistics of binary or survival outcomes, while the interpretation of the causal effect estimate obtained by this direct application is unclear and requires further justification (Zhao et al., 2019b).

12.2 Longitudinal studies

A longitudinal study is a research design involving repeated measures of the same variables over prolonged periods of time, widely used in epidemiology and social science to track trends and establish causal relationships. However, longitudinal studies aimed at estimating causal effects may be subject to bias arising from unmeasured confounding and/or time-varying confounding variables. MR analysis can mitigate such unmeasured and time-varying confounding bias by leveraging genetic variations as IVs, though it typically relies on data measured at a single time point. Developing MR methods tailored for longitudinal studies holds promise for estimating time-varying causal effects, thereby providing novel biological insights into lifetime health trajectories (Labrecque and Swanson, 2019; Morris et al., 2022; Sanderson et al., 2022b).

12.3 Multivariate MR

Multivariate MR extends the classical MR framework to estimate the causal effects of multiple exposures on an outcome simultaneously (Burgess and Thompson, 2015; Sanderson et al., 2019). This approach is particularly valuable when exposures are biologically correlated, and has been applied to disentangle complex causal relationships, for example, identifying metabolite biomarkers for age-related macular degeneration (Zuber et al., 2020). Recent studies have begun to address challenges arising from invalid instruments in multivariate MR framework (Liang et al., 2022; Lin et al., 2023; Chan et al., 2024). However, methods robust to the violation of core IV assumptions when handling high-dimensional exposures (e.g., omics biomarkers) are still lacking.

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 52008. The authors express sincere thanks to Dr. Paul Albert for initiating this tutorial and for his constructive feedback during its development.

References

- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500.
- Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63.
- Andrews, D. and Stock, J. H. (2005). Inference with weak instruments. NBER Technical Working Papers 0313, National Bureau of Economic Research, Inc.
- Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753.
- Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336.
- Angrist, J. D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):C52–C83.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.
- Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, 87(418):328–336.
- Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist’s companion*. Princeton University Press.
- Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340.

- Bateson, W. and Mendel, G. (2013). *Mendel's principles of heredity*. Courier Corporation.
- Biffen, R. H. (1905). Mendel's laws of inheritance and wheat breeding. *The Journal of Agricultural Science*, 1(1):4–48.
- Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ*, 310(6973):170.
- Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525.
- Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314.
- Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*, 36(11):1783–1802.
- Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, 91(433):14–28.
- Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I., et al. (2020). Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications*, 11(1):1–13.
- Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., and Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295.
- Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012.

- Burgess, S. (2024). Towards more reliable non-linear Mendelian randomization investigations. *European Journal of Epidemiology*, 39(5):447–449.
- Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665.
- Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1):1–11.
- Burgess, S., Swanson, S. A., and Labrecque, J. A. (2021). Are Mendelian randomization investigations immune from bias due to reverse causation? *European Journal of Epidemiology*, 36:253–257.
- Burgess, S. and Thompson, S. G. (2015). Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, 181(4):251–260.
- Burgess, S., Thompson, S. G., and Collaboration, C. C. G. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3):755–764.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O’Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.
- Castle, W. E. (1903). Mendel’s law of heredity. *Science*, 18(456):396–406.
- Chan, L. S., Malakhov, M. M., and Pan, W. (2024). A novel multivariable Mendelian randomization framework to disentangle highly correlated exposures with application to metabolomics. *The American Journal of Human Genetics*, 111(9):1834–1847.
- Chari, S. and Dworkin, I. (2013). The conditional nature of genetic interactions: the consequences of wild-type backgrounds on mutational interactions in a genome-wide modifier screen. *PLoS Genetics*, 9(8):e1003661.
- Chen, W.-M. and Abecasis, G. R. (2007). Family-based association tests for genomewide association scans. *The American Journal of Human Genetics*, 81(5):913–926.

- Christ, C. F. (1966). *Econometric models and methods*. John Wiley & Sons, Inc.
- Clarke, P. S. and Windmeijer, F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, 11(4):756–770.
- Craig, P., Katikireddi, S. V., Leyland, A., and Popham, F. (2017). Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annual Review of Public Health*, 38:39–56.
- Davey Smith, G. and Ebrahim, S. (2003). ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22.
- Davey Smith, G. and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98.
- Davey Smith, G., Holmes, M. V., Davies, N. M., and Ebrahim, S. (2020). Mendel’s laws, Mendelian randomization and causal inference in observational data: substantive and nomenclatural issues. *European Journal of Epidemiology*, 35(2):99–111.
- Davies, N. M., Hemani, G., Neiderhiser, J. M., Martin, H. C., Mills, M. C., Visscher, P. M., Yengo, L., Young, A. S., and Keller, M. C. (2024). The importance of family-based sampling for biobanks. *Nature*, 634(8035):795–803.
- Davies, N. M., Howe, L. J., Brumpton, B., Havdahl, A., Evans, D. M., and Davey Smith, G. (2019). Within family Mendelian randomization studies. *Human Molecular Genetics*, 28(R2):R170–R179.
- Deng, Y., Tu, D., O’Callaghan, C. J., Jonker, D. J., Karapetis, C. S., Shapiro, J., Liu, G., and Xu, W. (2023). A Bayesian approach for two-stage multivariate Mendelian randomization with mixed outcomes. *Statistics in Medicine*, 42(13):2241–2256.
- Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330.
- Dijksterhuis, E. J. (2014). *Archimedes*. Princeton University Press.
- DiNardo, J. (2010). Natural experiments and quasi-natural experiments. In *Microeconometrics*, pages 139–153. Springer.

- Dunning, T. (2012). *Natural experiments in the social sciences: A design-based approach*. Cambridge University Press.
- Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. *JAMA*, 318(19):1925–1926.
- Ference, B. A., Holmes, M. V., and Smith, G. D. (2021). Using Mendelian randomization to improve the design of randomized trials. *Cold Spring Harbor Perspectives in Medicine*, 11(7):a040980.
- Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd, Edinburgh.
- Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica: Journal of the Econometric Society*, pages 979–1001.
- Goldstein, C. E., Weijer, C., Brehaut, J. C., Fergusson, D. A., Grimshaw, J. M., Horn, A. R., and Taljaard, M. (2018). Ethical issues in pragmatic randomized controlled trials: a review of the recent literature identifies gaps in ethical argumentation. *BMC Medical Ethics*, 19:1–10.
- Guindo-Martínez, M., Amela, R., Bonàs-Guarch, S., Puiggròs, M., Salvoro, C., Miguel-Escalada, I., Carey, C. E., Cole, J. B., Rüeger, S., Atkinson, E., et al. (2021). The impact of non-additive genetic associations on age-related complex diseases. *Nature Communications*, 12(1):2436.
- Guo, Z. (2023). Causal inference with invalid instruments: post-selection problems and a solution using searching and sampling. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):959–985.
- Guo, Z., Kang, H., Tony Cai, T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):793–815.
- Guo, Z., Zheng, M., and Bühlmann, P. (2022). Robustness against weak or invalid instruments: Exploring nonlinear treatment models with machine learning. *arXiv preprint arXiv:2203.12808*.
- Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.
- Hahn, J. and Hausman, J. (2002). Notes on bias in estimators for simultaneous equation models. *Economics Letters*, 75(2):237–241.

- Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929.
- Han, C. (2008). Detecting invalid instruments using L1-GMM. *Economics Letters*, 101(3):285–287.
- Han, S. and Zhou, X.-H. (2023). Defining estimands in clinical trials: a unified procedure. *Statistics in Medicine*, 42(12):1869–1887.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.
- Harmanci, A. and Gerstein, M. (2016). Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature Methods*, 13(3):251–256.
- Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998.
- Haycock, P. C., Burgess, S., Wade, K. H., Bowden, J., Relton, C., and Smith, G. D. (2016). Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American Journal of Clinical Nutrition*, 103(4):965–978.
- Heckman, J. J. and Robb Jr, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1-2):239–267.
- Hellman, S. and Hellman, D. S. (2017). Of mice but not men: problems of the randomized clinical trial. In *Research Ethics*, pages 201–205. Routledge.
- Hemani, G., Bowden, J., and Davey Smith, G. (2018a). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics*, 27(R2):R195–R208.
- Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018b). The MR-Base platform supports systematic causal inference across the human phenome. *eLife*, 7:e34408.
- Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist’s dream? *Epidemiology*, 17(4):360–372.

- Hernán, M. A. and Robins, J. M. (2020). *Causal inference: what if*. CRC PRESS.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.
- Holland, P. W. (1988). Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50.
- Howe, L. J., Nivard, M. G., Morris, T. T., Hansen, A. F., Rasheed, H., Cho, Y., Chittoor, G., Ahlskog, R., Lind, P. A., Palviainen, T., et al. (2022). Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature Genetics*, 54(5):581–592.
- Hwang, L.-D., Davies, N. M., Warrington, N. M., and Evans, D. M. (2021). Integrating family-based and Mendelian randomization designs. *Cold Spring Harbor Perspectives in Medicine*, 11(3):a039503.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.
- Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.
- Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer’s disease risk. *Nature Genetics*, 51(3):404–413.

- Jiang, T., Gill, D., Butterworth, A. S., and Burgess, S. (2023). An empirical investigation into the impact of winner’s curse on estimates from Mendelian randomization. *International Journal of Epidemiology*, 52(4):1209–1219.
- Kahan, B. C., Hindley, J., Edwards, M., Cro, S., and Morris, T. P. (2024). The estimands framework: a primer on the ICH E9 (R1) addendum. *BMJ*, 384.
- Kamat, M. A., Blackshaw, J. A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A. S., and Staley, J. R. (2019). PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics*, 35(22):4851–4853.
- Kanai, M., Tanaka, T., and Okada, Y. (2016). Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set. *Journal of Human Genetics*, 61(10):861–866.
- Kang, H., Guo, Z., Liu, Z., and Small, D. (2024). Identification and inference with invalid instruments. *Annual Review of Statistics and Its Application*, 12.
- Kang, H., Lee, Y., Cai, T. T., and Small, D. S. (2022). Two robust tools for inference about causal effects with invalid instruments. *Biometrics*, 78(1):24–34.
- Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144.
- Katan, M. (1986). Apolipoprotein e isoforms, serum cholesterol, and cancer. *The Lancet*, 327(8479):507–508.
- Kaufman, D. J., Murphy-Bollinger, J., Scott, J., and Hudson, K. L. (2009). Public opinion about the importance of privacy in biobank research. *The American Journal of Human Genetics*, 85(5):643–654.
- Keene, O. N., Lynggaard, H., Englert, S., Lanius, V., and Wright, D. (2023). Why estimands are needed to define treatment effects in clinical trials. *BMC Medicine*, 21(1):276.
- Kleckner, N. (1996). Meiosis: how could it work? *Proceedings of the National Academy of Sciences*, 93(16):8167–8174.

- Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484.
- Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsen, B. J., Young, A. I., Thorgeirsson, T. E., Benonisdottir, S., Oddsson, A., Halldorsson, B. V., Masson, G., et al. (2018). The nature of nurture: Effects of parental genotypes. *Science*, 359(6374):424–428.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.
- Labrecque, J. A. and Swanson, S. A. (2019). Interpretation and potential biases of Mendelian randomization estimates with time-varying exposures. *American Journal of Epidemiology*, 188(1):231–238.
- LaPierre, N., Fu, B., Turnbull, S., Eskin, E., and Sankararaman, S. (2023). Leveraging family data to design Mendelian randomization that is provably robust to population stratification. *Genome Research*, 33(7):1032–1041.
- Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163.
- Leatherdale, S. T. (2019). Natural experiment methodology for research: a review of how different methods can support real-world research. *International Journal of Social Research Methodology*, 22(1):19–35.
- Lewis, J. A. (1999). Statistical principles for clinical trials (ich e9): an introductory note on an international guideline. *Statistics in Medicine*, 18(15):1903–1942.
- Liang, X., Sanderson, E., and Windmeijer, F. (2022). Selecting valid instrumental variables in linear models with multiple exposure variables: adaptive lasso and the median-of-medians estimator. *arXiv preprint arXiv:2208.05278*.
- Lin, Z., Xue, H., and Pan, W. (2023). Robust multivariable Mendelian randomization based on constrained maximum likelihood. *The American Journal of Human Genetics*, 110(4):592–605.

- Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388.
- Little, R. J. and Lewis, R. J. (2021). Estimands, estimators, and estimates. *JAMA*, 326(10):967–968.
- Liu, Z., Sun, B., Ye, T., Richardson, D., and Tchetgen, E. T. (2025). Quasi instrumental variable methods for stable hidden confounding and binary outcome. *arXiv preprint arXiv:2508.16096*.
- Liu, Z., Ye, T., Sun, B., Schooling, M., and Tchetgen, E. (2023). Mendelian randomization mixed-scale treatment effect robust identification and estimation for causal inference. *Biometrics*, 79(3):2208–2219.
- Lundberg, I., Johnson, R., and Stewart, B. M. (2021). What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3):532–565.
- Ma, X., Wang, J., and Wu, C. (2023). Breaking the winner’s curse in Mendelian randomization: Rerandomized inverse variance weighted estimator. *The Annals of Statistics*, 51(1):211–232.
- Mikusheva, A. and Sun, L. (2022). Inference with many weak instruments. *The Review of Economic Studies*, 89(5):2663–2686.
- Morris, T. T., Heron, J., Sanderson, E. C., Davey Smith, G., Didelez, V., and Tilling, K. (2022). Interpretation of Mendelian randomization using a single measure of an exposure that varies over time. *International Journal of Epidemiology*, 51(6):1899–1909.
- Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J.-P., Malin, B. A., and Wang, X. (2015). Privacy in the genomic era. *ACM Computing Surveys (CSUR)*, 48(1):1–44.
- Neyman, J. (1923). Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51.
- Nkomo, V. T., Gardin, J. M., Skelton, T. N., Gottdiener, J. S., Scott, C. G., and Enriquez-Sarano, M. (2006). Burden of valvular heart diseases: a population-based study. *The Lancet*, 368(9540):1005–1011.

- Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., Smith, G. D., and Sterne, J. A. (2012). Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research*, 21(3):223–242.
- Panagiotou, O. A., Ioannidis, J. P., and Project, G.-W. S. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1):273–286.
- Paré, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women’s Genome Health Study. *PLoS Genetics*, 6(6):e1000981.
- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96 – 146.
- Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787.
- Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575.
- Qi, G. and Chatterjee, N. (2019). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications*, 10(1):1–10.
- Richardson, T. S. and Robins, J. M. (2014). ACE bounds; SEMs with equilibrium conditions. *Statistical Science*, 29(3):363–366.
- Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and Methods*, 23(8):2379–2412.

- Rosenbaum, P. R., Rosenbaum, P., and Briskman (2010). *Design of observational studies*, volume 10. Springer.
- Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of Econometrics*, 2:881–935.
- Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26.
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.
- Sanderson, E., Davey Smith, G., Windmeijer, F., and Bowden, J. (2019). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, 48(3):713–727.
- 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. (2022a). Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):6.
- Sanderson, E., Richardson, T. G., Hemani, G., and Davey Smith, G. (2021). The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *International Journal of Epidemiology*, 50(4):1350–1361.
- Sanderson, E., Richardson, T. G., Morris, T. T., Tilling, K., and Davey Smith, G. (2022b). Estimation of causal effects of a time-varying exposure at multiple time points through multivariable Mendelian randomization. *PLoS Genetics*, 18(7):e1010290.

- Sanson-Fisher, R. W., D’Este, C. A., Carey, M. L., Noble, N., and Paul, C. L. (2014). Evaluation of systems-oriented public health interventions: alternative research designs. *Annual Review of Public Health*, 35:9–27.
- Schlesselman, J. J. (1982). *Case-control studies: design, conduct, analysis*, volume 2. Oxford university press.
- Sedgwick, P. (2013). Prospective cohort studies: advantages and disadvantages. *BMJ*, 347.
- Sedgwick, P. (2014). Retrospective cohort studies: advantages and disadvantages. *BMJ*, 348.
- Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618.
- Slob, E. A. and Burgess, S. (2020). A comparison of robust Mendelian randomization methods using summary data. *Genetic Epidemiology*, 44(4):313–329.
- Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058.
- Smith, G. D., Paternoster, L., and Relton, C. (2017). When will Mendelian randomization become relevant for clinical practice and public health? *JAMA*, 317(6):589–591.
- Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495.
- Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.
- Staley, J. R. and Burgess, S. (2017). Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genetic Epidemiology*, 41(4):341–352.
- Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

- Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear IV regression. NBER Technical Working Papers 0284, National Bureau of Economic Research, Inc.
- Stolberg, H. O., Norman, G., and Trop, I. (2004). Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 25(1):1.
- Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, 12(3):e1001779.
- Sulc, J., Sjaarda, J., and Kutalik, Z. (2022). Polynomial Mendelian randomization reveals non-linear causal effects for obesity-related traits. *Human Genetics and Genomics Advances*, 3(3).
- Sun, B., Liu, Z., and Tchetgen Tchetgen, E. (2023a). Semiparametric efficient G-estimation with invalid instrumental variables. *Biometrika*, 110(4):953–971.
- Sun, B. B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.-H., Richardson, T. G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S. G., et al. (2023b). Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*, 622(7982):329–338.
- Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V., White, J., Mindell, J. S., Kivimaki, M., Brunner, E. J., Whittaker, J. C., et al. (2016). Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology*, 45(5):1600–1616.
- Szklo, M. (1998). Population-based cohort studies. *Epidemiologic Reviews*, 20(1):81–90.
- Tchetgen Tchetgen, E., Sun, B., and Walter, S. (2021). The GENIUS approach to robust Mendelian randomization inference. *Statistical Science*, 36(3):443–464.
- Terza, J. V., Basu, A., and Rathouz, P. J. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543.
- Thanassoulis, G. and O’Donnell, C. J. (2009). Mendelian randomization: nature’s randomized trial in the post-genome era. *JAMA*, 301(22):2386–2388.

- VanderWeele, T. J. (2016). Commentary: on causes, causal inference, and potential outcomes. *International Journal of Epidemiology*, 45(6):1809–1816.
- VanderWeele, T. J., Tchetgen, E. J. T., Cornelis, M., and Kraft, P. (2014). Methodological challenges in Mendelian randomization. *Epidemiology*, 25(3):427–435.
- Veitia, R. A., Bottani, S., and Birchler, J. A. (2013). Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. *Trends in Genetics*, 29(7):385–393.
- Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, 50(5):693–698.
- Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.
- Wan, F., Colditz, G. A., and Sutcliffe, S. (2021). Matched versus unmatched analysis of matched case-control studies. *American Journal of Epidemiology*, 190(9):1859–1866.
- Wang, A., Liu, W., and Liu, Z. (2022). A two-sample robust Bayesian Mendelian Randomization method accounting for linkage disequilibrium and idiosyncratic pleiotropy with applications to the COVID-19 outcomes. *Genetic Epidemiology*, 46(3-4):159–169.
- Wang, H., Zhang, F., Zeng, J., Wu, Y., Kemper, K. E., Xue, A., Zhang, M., Powell, J. E., Goddard, M. E., Wray, N. R., et al. (2019). Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank. *Science Advances*, 5(8):eaaw3538.
- Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):531–550.
- Windmeijer, F., Liang, X., Hartwig, F. P., and Bowden, J. (2021). The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):752–776.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

- Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. South-Western Cengage Learning.
- Wright, P. G. (1928). *The tariff on animal and vegetable oils*. Macmillan.
- Xu, S., Fung, W. K., and Liu, Z. (2021). MRCIP: a robust Mendelian randomization method accounting for correlated and idiosyncratic pleiotropy. *Briefings in Bioinformatics*, 22(5):bbab019.
- Xu, S., Wang, P., Fung, W. K., and Liu, Z. (2023). A novel penalized inverse-variance weighted estimator for Mendelian randomization with applications to COVID-19 outcomes. *Biometrics*, 79(3):2184–2195.
- Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4):369–375.
- Yao, M. and Liu, Z. (2025). An introduction to causal inference methods with multi-omics data. *Current Protocols*, 5(6):e70168.
- Yao, M., Miller, G. W., Vardarajan, B. N., Baccarelli, A. A., Guo, Z., and Liu, Z. (2024). Deciphering proteins in Alzheimer’s disease: A new Mendelian randomization method integrated with AlphaFold3 for 3D structure prediction. *Cell Genomics*, 4(12).
- Ye, T., Liu, Z., Sun, B., and Tchetgen Tchetgen, E. (2024). GENIUS-MAWII: For robust Mendelian randomization with many weak invalid instruments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):1045–1067.
- Ye, T., Shao, J., and Kang, H. (2021). Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *The Annals of Statistics*, 49(4):2079–2100.
- Zhang, D., Yao, M., Liu, Z., and Sun, B. (2025). Mr-magic: Robust causal inference using many weak genetic interactions. *arXiv preprint arXiv:2504.13565*.
- Zhao, Q., Chen, Y., Wang, J., and Small, D. S. (2019a). Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization. *International Journal of Epidemiology*, 48(5):1478–1492.

- Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48(3):1742–1769.
- Zhao, Q., Wang, J., Spiller, W., Bowden, J., and Small, D. S. (2019b). Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science*, 34(2):317–333.
- Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, 48(5):481–487.
- Zuber, V., Colijn, J. M., Klaver, C., and Burgess, S. (2020). Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. *Nature Communications*, 11(1):29.

Supplementary Materials

S1 Weak IV bias in one-sample and two-sample MR

To analyze the finite-sample behavior of MR estimators in one-sample and two-sample design, we derive the weak IV bias expressions under the ALICE model, assuming all instruments are valid and homogeneity conditions (C1)–(C2) hold. Throughout, we denote the sample size of sample $s \in \{1, 2\}$ as $n^{(s)}$, and we assume genetic instruments are standardized and mutually independent, i.e., $\mathbb{E}(Z_{ij}^{(s)}) = 0$, $\text{Var}(Z_{ij}^{(s)}) = 1$, and $\text{Cov}(Z_{ij}^{(s)}, Z_{ik}^{(s)}) = 0$ for $j \neq k$.

S1.1 Bias of the one-sample 2SLS estimator

The one-sample 2SLS estimator in sample s is defined as

$$\widehat{\beta}_{2\text{SLS}}^{(s)} = \left(\mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \mathbf{D}^{(s)} \right)^{-1} \mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \mathbf{Y}^{(s)},$$

where $\mathbf{P}_{\mathbf{Z}^{(s)}} = \mathbf{Z}^{(s)} (\mathbf{Z}^{(s)\top} \mathbf{Z}^{(s)})^{-1} \mathbf{Z}^{(s)\top}$ is the projection matrix onto the column space of $\mathbf{Z}^{(s)}$. Substituting the structural models $\mathbf{D}^{(s)} = \mathbf{Z}^{(s)} \boldsymbol{\gamma} + \boldsymbol{\delta}^{(s)}$ and $\mathbf{Y}^{(s)} = \beta \mathbf{D}^{(s)} + \boldsymbol{\varepsilon}^{(s)}$, we obtain

$$\widehat{\beta}_{2\text{SLS}}^{(s)} = \beta + \left(\mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \mathbf{D}^{(s)} \right)^{-1} \mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \boldsymbol{\varepsilon}^{(s)}.$$

To approximate the bias, we take expectations of the numerator and denominator separately:

$$\mathbb{E} \left[\widehat{\beta}_{2\text{SLS}}^{(s)} \right] \approx \beta + \frac{\mathbb{E} \left[\mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \boldsymbol{\varepsilon}^{(s)} \right]}{\mathbb{E} \left[\mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \mathbf{D}^{(s)} \right]}.$$

For the numerator, note that $\mathbf{D}^{(s)} = \mathbf{Z}^{(s)} \boldsymbol{\gamma} + \boldsymbol{\delta}^{(s)}$, and that $\mathbb{E}[\mathbf{Z}^{(s)\top} \boldsymbol{\varepsilon}^{(s)}] = 0$. Therefore,

$$\mathbb{E} \left[\mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \boldsymbol{\varepsilon}^{(s)} \right] = \mathbb{E} \left[\boldsymbol{\delta}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \boldsymbol{\varepsilon}^{(s)} \right] = \text{tr} \left(\mathbf{P}_{\mathbf{Z}^{(s)}} \mathbb{E} \left[\boldsymbol{\delta}^{(s)} \boldsymbol{\varepsilon}^{(s)\top} \right] \right) = \sigma_{\delta, \varepsilon} \cdot p.$$

For the denominator, we expand:

$$\mathbb{E} \left[\mathbf{D}^{(s)\top} \mathbf{P}_{\mathbf{Z}^{(s)}} \mathbf{D}^{(s)} \right] = \boldsymbol{\gamma}^\top \mathbb{E}[\mathbf{Z}^{(s)\top} \mathbf{Z}^{(s)}] \boldsymbol{\gamma} + \sigma_\delta^2 \cdot \text{tr}(\mathbf{P}_{\mathbf{Z}^{(s)}}) = n^{(s)} \|\boldsymbol{\gamma}\|^2 + p \sigma_\delta^2.$$

Hence, the approximate bias is

$$\mathbb{E} \left[\widehat{\beta}_{2\text{SLS}}^{(s)} \right] - \beta \approx \frac{\sigma_{\delta, \varepsilon}}{\frac{n^{(s)} \|\boldsymbol{\gamma}\|^2}{p} + \sigma_\delta^2}.$$

S1.2 Bias of the two-sample 2SLS estimator

Let sample 1 be used to estimate the IV–exposure associations and sample 2 for the second stage. The first-stage estimator from sample 1 is:

$$\hat{\gamma}^{(1)} = \left(\mathbf{Z}^{(1)\top} \mathbf{Z}^{(1)} \right)^{-1} \mathbf{Z}^{(1)\top} \mathbf{D}^{(1)}.$$

Define the predicted exposure in sample 2 by

$$\hat{\mathbf{D}}^{(2,1)} = \mathbf{Z}^{(2)} \hat{\gamma}^{(1)}.$$

Then, the two-sample 2SLS estimator is

$$\hat{\beta}_{\text{SSIV}} = \left(\hat{\mathbf{D}}^{(2,1)\top} \hat{\mathbf{D}}^{(2,1)} \right)^{-1} \hat{\mathbf{D}}^{(2,1)\top} \mathbf{Y}^{(2)}.$$

We now compute the expectation of this estimator under the ALICE model with standardized instruments and valid IVs. Recall that

$$\mathbf{D}^{(1)} = \mathbf{Z}^{(1)} \boldsymbol{\gamma} + \boldsymbol{\delta}^{(1)}, \quad \mathbf{Y}^{(2)} = \beta \mathbf{Z}^{(2)} \boldsymbol{\gamma} + \mathbf{e}^{(2)}, \quad \text{where } \mathbf{e}^{(2)} = \beta \boldsymbol{\delta}^{(2)} + \boldsymbol{\varepsilon}^{(2)}.$$

Then, the numerator of $\hat{\beta}_{\text{SSIV}}$ is

$$\mathbb{E} \left[\hat{\mathbf{D}}^{(2,1)\top} \mathbf{Y}^{(2)} \right] = \mathbb{E} \left[\hat{\gamma}^{(1)\top} \mathbf{Z}^{(2)\top} \mathbf{Y}^{(2)} \right].$$

Because samples 1 and 2 are independent, the expectation factorizes as

$$\mathbb{E} \left[\hat{\gamma}^{(1)\top} \right] \cdot \mathbb{E} \left[\mathbf{Z}^{(2)\top} \mathbf{Y}^{(2)} \right] = \boldsymbol{\gamma}^\top \cdot \beta \cdot \mathbb{E} \left[\mathbf{Z}^{(2)\top} \mathbf{Z}^{(2)} \right] \cdot \boldsymbol{\gamma}.$$

Since $\mathbf{Z}^{(2)}$ is standardized and the instruments are mutually independent, $\mathbb{E} \left[\mathbf{Z}^{(2)\top} \mathbf{Z}^{(2)} \right] = n^{(2)} \mathbf{I}_p$.

Therefore,

$$\mathbb{E} \left[\hat{\mathbf{D}}^{(2,1)\top} \mathbf{Y}^{(2)} \right] = \beta \cdot n^{(2)} \|\boldsymbol{\gamma}\|^2.$$

We next compute the denominator:

$$\mathbb{E} \left[\hat{\mathbf{D}}^{(2,1)\top} \hat{\mathbf{D}}^{(2,1)} \right] = \mathbb{E} \left[\hat{\gamma}^{(1)\top} \mathbf{Z}^{(2)\top} \mathbf{Z}^{(2)} \hat{\gamma}^{(1)} \right].$$

To compute this, we decompose the estimation error in $\hat{\gamma}^{(1)}$:

$$\hat{\gamma}^{(1)} = \boldsymbol{\gamma} + \left(\mathbf{Z}^{(1)\top} \mathbf{Z}^{(1)} \right)^{-1} \mathbf{Z}^{(1)\top} \boldsymbol{\delta}^{(1)}.$$

Using independence and standardization, we have

$$\mathbb{E} \left[\left\| \left(\mathbf{Z}^{(1)\top} \mathbf{Z}^{(1)} \right)^{-1} \mathbf{Z}^{(1)\top} \boldsymbol{\delta}^{(1)} \right\|^2 \right] = \frac{p\sigma_{\delta}^2}{n^{(1)}}.$$

Thus,

$$\mathbb{E} \left[\hat{\boldsymbol{\gamma}}^{(1)\top} \mathbf{Z}^{(2)\top} \mathbf{Z}^{(2)} \hat{\boldsymbol{\gamma}}^{(1)} \right] = n^{(2)} \|\boldsymbol{\gamma}\|^2 + n^{(2)} \frac{p\sigma_{\delta}^2}{n^{(1)}}$$

Putting the numerator and denominator together, we have

$$\mathbb{E} \left(\hat{\beta}_{\text{SSIV}} \right) \approx \beta \cdot \frac{\|\boldsymbol{\gamma}\|^2}{\|\boldsymbol{\gamma}\|^2 + \frac{p\sigma_{\delta}^2}{n^{(1)}}}.$$

S1.3 Bias of the One-Sample IVW Estimator

We now derive the bias of the one-sample inverse-variance weighted (IVW) estimator under the ALICE model. Let $\hat{\gamma}_j^{(s)}$ and $\hat{\Gamma}_j^{(s)}$ denote the marginal SNP-exposure and SNP-outcome associations for instrument j in sample s , defined as

$$\hat{\gamma}_j^{(s)} = \frac{1}{n^{(s)}} \mathbf{Z}_{\cdot j}^{(s)\top} \mathbf{D}^{(s)}, \quad \hat{\Gamma}_j^{(s)} = \frac{1}{n^{(s)}} \mathbf{Z}_{\cdot j}^{(s)\top} \mathbf{Y}^{(s)},$$

where $\mathbf{Z}_{\cdot j}^{(s)}$ is the j th column of $\mathbf{Z}^{(s)}$, and $n^{(s)}$ is the sample size. The IVW estimator is given by

$$\hat{\beta}_{\text{IVW}}^{(s)} = \frac{\sum_{j=1}^p \hat{\Gamma}_j^{(s)} \hat{\gamma}_j^{(s)} / (\hat{\sigma}_{\Gamma, j}^{(s)})^2}{\sum_{j=1}^p (\hat{\gamma}_j^{(s)})^2 / (\hat{\sigma}_{\Gamma, j}^{(s)})^2},$$

where $\hat{\sigma}_{\Gamma, j}^{(s)}$ denotes the estimated standard error of $\hat{\Gamma}_j^{(s)}$, treated as known and non-random in this analysis.

We compute the expectation of the IVW estimator by separately expanding the numerator and denominator. Using the identity

$$\mathbb{E}[\hat{\beta}_{\text{IVW}}^{(s)}] \approx \frac{\sum_j \mathbb{E}[\hat{\Gamma}_j^{(s)} \hat{\gamma}_j^{(s)}] / (\hat{\sigma}_{\Gamma, j}^{(s)})^2}{\sum_j \mathbb{E}[(\hat{\gamma}_j^{(s)})^2] / (\hat{\sigma}_{\Gamma, j}^{(s)})^2},$$

we first expand the numerator using the identity

$$\mathbb{E}[\hat{\Gamma}_j^{(s)} \hat{\gamma}_j^{(s)}] = \mathbb{E}[\hat{\Gamma}_j^{(s)}] \mathbb{E}[\hat{\gamma}_j^{(s)}] + \text{Cov}(\hat{\Gamma}_j^{(s)}, \hat{\gamma}_j^{(s)}) = \beta_j \gamma_j^2 + \text{Cov}(\hat{\Gamma}_j^{(s)}, \hat{\gamma}_j^{(s)}).$$

Under the ALICE model, it can be shown that

$$(\sigma_{\gamma, j}^{(s)})^2 = \text{Var}(\hat{\gamma}_j^{(s)}) = \frac{1}{n^{(s)}} \left((\nu_j^{(s)} - 1)^2 \gamma_j^2 + \|\boldsymbol{\gamma}\|^2 + \sigma_{\delta}^2 \right),$$

$$\text{Cov}(\widehat{\Gamma}_j^{(s)}, \widehat{\gamma}_j^{(s)}) = \frac{1}{n^{(s)}} \left((\nu_j^{(s)} - 1)^2 \beta \gamma_j^2 + \beta \|\boldsymbol{\gamma}\|^2 + \beta \sigma_\delta^2 + \sigma_{\delta\varepsilon} \right),$$

where $\nu_j^{(s)} = \mathbb{E}[(Z_{ij}^{(s)})^4]$ is the kurtosis of the instrument j in sample s , and $\sigma_{\delta\varepsilon} = \text{Cov}(\delta_i^{(s)}, \varepsilon_i^{(s)})$ reflects unmeasured confounding.

Combining these expressions, we have

$$\mathbb{E}[\widehat{\Gamma}_j^{(s)} \widehat{\gamma}_j^{(s)}] = \beta \cdot \left(\gamma_j + (\sigma_{\gamma,j}^{(s)})^2 \right) + \frac{1}{n^{(s)}} \sigma_{\delta\varepsilon},$$

and so the expected value of the numerator becomes

$$\sum_j \frac{\mathbb{E}[\widehat{\Gamma}_j^{(s)} \widehat{\gamma}_j^{(s)}]}{(\widehat{\sigma}_{\Gamma,j}^{(s)})^2} = \beta \sum_j \frac{\gamma_j^2 + (\sigma_{\gamma,j}^{(s)})^2}{(\widehat{\sigma}_{\Gamma,j}^{(s)})^2} + \frac{\sigma_{\delta\varepsilon}}{n^{(s)}} \sum_j \frac{1}{(\widehat{\sigma}_{\Gamma,j}^{(s)})^2}.$$

For the denominator, we use the variance expansion again:

$$\mathbb{E}[(\widehat{\gamma}_j^{(s)})^2] = \gamma_j^2 + (\sigma_{\gamma,j}^{(s)})^2.$$

Therefore, putting numerator and denominator together, we obtain the approximate expectation:

$$\mathbb{E}[\widehat{\beta}_{\text{IVW}}^{(s)}] \approx \beta + \frac{\sigma_{\delta\varepsilon}}{n^{(s)}} \cdot \frac{\sum_{j=1}^p 1/(\widehat{\sigma}_{\Gamma,j}^{(s)})^2}{\sum_{j=1}^p (\gamma_j^2 + (\sigma_{\gamma,j}^{(s)})^2) / (\widehat{\sigma}_{\Gamma,j}^{(s)})^2}.$$

S1.4 Bias of the two-sample IVW estimator

We derive the bias of the two-sample inverse-variance weighted (IVW) estimator under the ALICE model using marginal association estimates from two independent samples. Let $\widehat{\gamma}_j^{(1)}$ and $\widehat{\Gamma}_j^{(2)}$ denote the marginal SNP-exposure and SNP-outcome associations for variant j , estimated from sample 1 and sample 2, respectively. The estimator is given by

$$\widehat{\beta}_{\text{IVW}}^{(1,2)} = \frac{\sum_{j=1}^p \widehat{\Gamma}_j^{(2)} \widehat{\gamma}_j^{(1)} / (\widehat{\sigma}_{\Gamma,j}^{(2)})^2}{\sum_{j=1}^p (\widehat{\gamma}_j^{(1)})^2 / (\widehat{\sigma}_{\Gamma,j}^{(2)})^2},$$

where $\widehat{\sigma}_{\Gamma,j}^{(2)}$ denotes the standard error of $\widehat{\Gamma}_j^{(2)}$. Because the samples are independent, the expectation of the product of estimates factorizes as

$$\mathbb{E}[\widehat{\Gamma}_j^{(2)} \widehat{\gamma}_j^{(1)}] = \mathbb{E}[\widehat{\Gamma}_j^{(2)}] \cdot \mathbb{E}[\widehat{\gamma}_j^{(1)}] = \beta \gamma_j^2.$$

Thus, the expectation of the numerator becomes

$$\mathbb{E} \left[\sum_{j=1}^p \frac{\widehat{\Gamma}_j^{(2)} \widehat{\gamma}_j^{(1)}}{(\widehat{\sigma}_{\Gamma,j}^{(2)})^2} \right] = \beta \sum_{j=1}^p \frac{\gamma_j^2}{(\widehat{\sigma}_{\Gamma,j}^{(2)})^2}.$$

To compute the expectation of the denominator, we have

$$\mathbb{E}[(\hat{\gamma}_j^{(1)})^2] = \gamma_j^2 + (\sigma_{\gamma,j}^{(1)})^2.$$

The expectation of the denominator is then

$$\sum_{j=1}^p \frac{\mathbb{E}[(\hat{\gamma}_j^{(1)})^2]}{(\hat{\sigma}_{\Gamma,j}^{(2)})^2} = \sum_{j=1}^p \frac{\gamma_j^2 + (\sigma_{\gamma,j}^{(1)})^2}{(\hat{\sigma}_{\Gamma,j}^{(2)})^2}.$$

Putting these together, we have

$$\begin{aligned} \mathbb{E}[\hat{\beta}_{\text{IVW}}^{(1,2)}] &= \mathbb{E}\left[\frac{\sum_{j=1}^p \hat{\Gamma}_j^{(2)} \hat{\gamma}_j^{(1)} / (\hat{\sigma}_{\Gamma,j}^{(2)})^2}{\sum_{j=1}^p (\hat{\gamma}_j^{(1)})^2 / (\hat{\sigma}_{\Gamma,j}^{(2)})^2}\right] \approx \frac{\mathbb{E}\left[\sum_{j=1}^p \hat{\Gamma}_j^{(2)} \hat{\gamma}_j^{(1)} / (\hat{\sigma}_{\Gamma,j}^{(2)})^2\right]}{\mathbb{E}\left[\sum_{j=1}^p (\hat{\gamma}_j^{(1)})^2 / (\hat{\sigma}_{\Gamma,j}^{(2)})^2\right]} \\ &= \beta \cdot \frac{\sum_{j=1}^p \gamma_j^2 / (\hat{\sigma}_{\Gamma,j}^{(2)})^2}{\sum_{j=1}^p (\gamma_j^2 + (\sigma_{\gamma,j}^{(1)})^2) / (\hat{\sigma}_{\Gamma,j}^{(2)})^2} = \beta \times \frac{1}{1 + \frac{\sum_{j=1}^p (\sigma_{\gamma,j}^{(1)})^2 / (\hat{\sigma}_{\Gamma,j}^{(2)})^2}{\sum_{j=1}^p \gamma_j^2 / (\hat{\sigma}_{\Gamma,j}^{(2)})^2}}. \end{aligned}$$

S2 MR software

In this section, we list commonly used software for MR analysis within the R environment.

S2.1 One-sample MR software with individual-level data

1. “ivreg”: perform two-stage least squares (2SLS) estimation. (<https://github.com/zeileis/ivreg>)
2. “OneSampleMR”: integrate several one-sample MR methods including multiplicative structural mean model (Hernán and Robins, 2006), two-stage predictor substitution estimators (Terza et al., 2008), and two-stage residual inclusion estimators (Terza et al., 2008). (<https://github.com/remlapmot/OneSampleMR>)
3. “sisVIVE”: estimate the causal effect with Some Invalid Some Valid IV Estimator (sisVIVE) (Kang et al., 2016). (<https://github.com/hyunseungkang/sisVIVE>)
4. “RobustIV”: perform causal inference with possibly invalid IVs using Two-Stage Hard Thresholding (TSHT) (Guo et al., 2018) and Searching-and-Sampling (Guo, 2023) methods. (<https://github.com/zijguo/RobustIV>)
5. “CIIV”: Confidence Interval method for Instrumental Variable estimation (Windmeijer et al., 2021). (<https://github.com/xlbristol/CIIV>)

6. “TSCI”: Two-Stage Curvature Identification with machine learning (Guo et al., 2022). (<https://github.com/dlcarl/TSCI>)
7. “MRMiSTERI”: Mendelian Randomization Mixed-Scale Treatment Effect Robust Identification and estimation for causal inference (Liu et al., 2023). (<https://github.com/zhonghualiu/MRMiSTERI>)
8. “MRSquared”: semiparametric efficient G-estimation with invalid instrumental variables (Sun et al., 2023a). (<https://github.com/zhonghualiu/MRSquared>)
9. “mr.genius”: robust MR analysis with many weak invalid instruments (Ye et al., 2024). (<https://github.com/tye27/mr.genius>)
10. “MR-MAGIC”: Mendelian Randomization with MAny weak Genetic Interactions for Causality (Zhang et al., 2025). (<https://github.com/zhangd17-web/MR-MAGIC>)

S2.2 Two-sample MR software with summary-level data

1. “TwoSampleMR”: integrate several two-sample MR methods using GWAS summary statistics, for example, inverse-variance weighted (IVW) method, MR-Egger regression, weighted median method, and mode-based estimation. (<https://github.com/MRCIEU/TwoSampleMR>)
2. “MendelianRandomization”: implement various two-sample MR methods with summary-level data. (<https://github.com/cran/MendelianRandomization>)
3. “MR-PRESSO”: Mendelian Randomization Pleiotropy RESidual Sum and Outlier test (Verbanck et al., 2018). (<https://github.com/rondolab/MR-PRESSO>)
4. “MRMix”: MRMix uses a normal mixture model to account for the widespread horizontal pleiotropy. (Qi and Chatterjee, 2019)(<https://github.com/gqi/MRMix>)
5. “mr.raps”: MR analysis using Robust Adjusted Profile Score (Zhao et al., 2020). (<https://github.com/qingyuanzhao/mr.raps>)
6. “mr.divw”: debiased IVW estimator with many weak instruments (Ye et al., 2021). (<https://github.com/tye27/mr.divw>)
7. “MRCIP”: MR analysis accounting for correlated and idiosyncratic pleiotropy (Xu et al., 2021). (<https://github.com/siqixu/MRCIP>)
8. “RBMRe”: Robust Bayesian MR analysis accounting for linkage disequilibrium and idiosyncratic pleiotropy (Wang et al., 2022). (<https://github.com/AnqiWang2021/RBMRe>)
9. “mr.pivw”: penalized IVW estimator for MR analysis (Xu et al., 2023). (<https://github.com/tye27/mr.pivw>)

[com/siqixu/mr.pivw](https://github.com/siqixu/mr.pivw))

10. “MR-SPI”: MR analysis that first Selects valid genetic variants and then performs Post-selection Inference (Yao et al., 2024). (<https://github.com/MinhaoYaooo/MR-SPI>)

S3 Glossary of MR assumptions and terms

We summarize below the key assumptions employed throughout this paper in the following Table S1.

Table S1: Glossary of MR assumptions used throughout the paper

Assumption	Name	Description
A1	IV relevance	The genetic variant is associated with the exposure.
A2	IV independence	The genetic variant is not associated with unmeasured confounders of the exposure-outcome relationship.
A3	Exclusion restriction	The genetic variant affects the outcome only through the exposure.
A4.1	Constant treatment effect	The treatment effect is constant across subjects.
A4.2	Additive homogeneity	The average treatment effect is the same across different levels of instrument for both treated and untreated groups.
A4.3	Monotonicity	No defiers in the population.
A4.4	No additive $U_i - Z_i$ interaction	No modification of IV-treatment association by unmeasured confounders on the additive scale.
A4.5	No additive $U_i - d$ interaction	No modification of the effect of the treatment on the outcome by unmeasured confounders on the additive scale.
A5	Confounding control	The potential outcomes are independent of the treatment status and the IV conditional on unmeasured confounders.
A6	Instrument strength independent of the direct effect (InSIDE)	The IV–exposure association is asymptotically independent of the degree of IV invalidity.
A7	Majority rule	The number of valid IVs exceeds half of the relevant IVs.
A8	Plurality rule	Valid IVs form the largest group among relevant IVs.

Table S1 (continued): Glossary of Assumptions

Label	Name	Description
B1	Homogeneous ATT	The ATT is homogeneous across levels of the instrument on the additive scale.
B2	Homogeneous confounding bias	The confounding bias is homogeneous across levels of the instrument on the odds ratio scale.
B3	Outcome heteroscedasticity	The residual variance for the outcome varies with the instrument.
C1	Homogeneity in parameters	The causal effect, IV-associations and degree of IV invalidity are equal across two samples.
C2	Homogeneity in error distribution	The joint distribution of errors in exposure and outcome models is the same across two samples.

We also summarize terms and abbreviations in the following Table S2.

Table S2: List of abbreviations

Abbreviation	Full Name
2SLS	two-stage least squares
AD	Alzheimer's disease
AKP	Anna Karenina Principle
ALICE	additive linear constant effect model
ATE	average treatment effect
ATT	average treatment effect on the treated
BMI	body mass index
CI	confidence interval
CIIV	confidence interval method for instrumental variable
DBP	diastolic blood pressure
dIVW	debiased inverse-variance weighted
GENIUS	G-estimation under no interaction with unmeasured selection
GENIUS-MAWII	G-estimation under no interaction with unmeasured selection leveraging many weak invalid instruments
InSIDE	instrument strength independent of the direct effect

Continued on next page

Table S2 (continued)

Abbreviation	Full Name
ITE	individual treatment effect
IV	instrumental variable
IVW	inverse-variance weighted
LATE	local average treatment effect
MR	Mendelian randomization
MR-MAGIC	Mendelian randomization with many weak genetic interactions for causality
MR-MiSTERI	Mendelian randomization mixed-scale treatment effect robust identification
MR-PRESSO	Mendelian randomization pleiotropy residual sum and outlier test
MR-RAPS	Mendelian randomization using the robust adjusted profile score
MR-SPI	Mendelian randomization with valid IV selection and post-selection inference
OLS	ordinary least squares
pIVW	penalized inverse-variance weighted
RCT	randomized controlled trial
sisVIVE	some invalid some valid IV estimator
SNP	single-nucleotide polymorphism
SSIV	split-sample instrumental variable
TSHT	two-stage hard thresholding
UKB	UK Biobank
UKB-PPP	UK Biobank Pharma Proteomics Project
xMR	multi-omics Mendelian randomization
ZEMPA	zero modal pleiotropy assumption

S4 Supplementary Figures

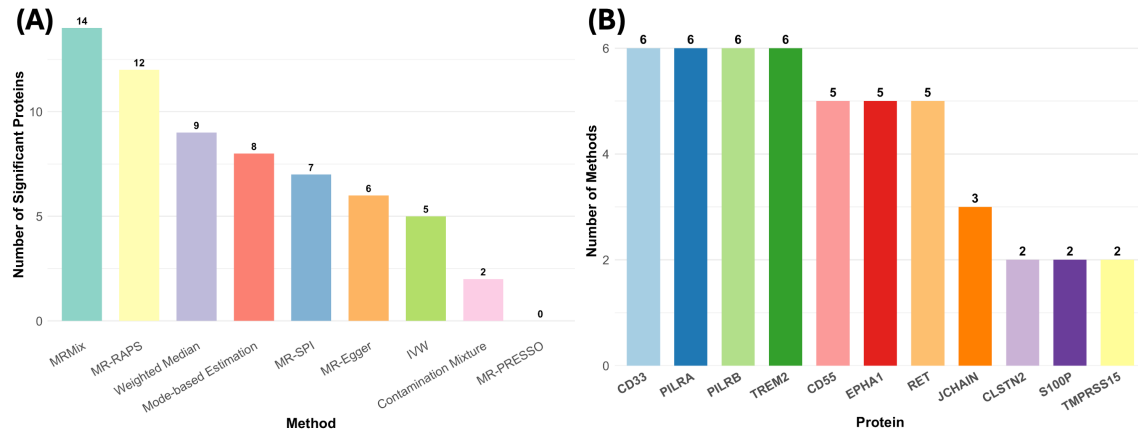


Figure S1: (A) Number of plasma proteins significantly associated with Alzheimer's disease identified by each MR method. (B) Proteins identified as significant by at least two methods, with the count of MR methods that identify each protein.