Selecting Valid Genetic Instruments and Constructing Robust Confidence Intervals for Two-sample Mendelian Randomization Using Genome-wide Summary Statistics

Minhao Yao[1], Zijian Guo[2*], Zhonghua Liu[3*]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong SAR, China.

[2] Department of Statistics, Rutgers University, Piscataway, NJ, USA.

[3] Department of Biostatistics, Columbia University, New York, NY, USA.

[*] Correspondence to: Zijian Guo (zijguo@stat.rutgers.edu) or Zhonghua Liu (zl2509@cumc.columbia.edu)

February 21, 2023

Abstract

Mendelian randomization (MR) is a causal inference method that uses genetic variants as instrumental variables (IVs) to estimate the causal effect of a modifiable exposure on the outcome of interest, even in the presence of unmeasured confounding. However, the core assumptions defining a valid IV might be violated in practice due to the presence of population stratification and/or widespread horizontal pleiotropy, leading to biased inference and possibly misleading scientific conclusions. To address this pressing challenge, we propose a novel MR method that first Selects valid genetic instruments and then performs Post-selection Inference (MR-SPI) to construct robust confidence intervals based on two-sample genome-wide summary statistics. Specifically, our proposed MR-SPI method has four steps: (1)

selecting relevant IVs that are associated with the exposure variable using hard thresholding; (2) the selected relevant IVs vote for each other to be a valid IV if their ratio estimates of the causal effect are similar; (3) selecting valid IVs by majority/plurality voting or finding the maximum clique of the voting matrix; and (4) constructing robust confidence intervals using the searching and sampling technique. We conduct extensive simulations to assess the practical performance of our method. Finally, we apply our MR-SPI method to learn causal relationships among 146 exposure-outcome pairs.

Keywords: Causal inference; Invalid instruments; Majority rule; Mendelian randomization; Pleiotropy; Plurality rule; Post-selection inference

Availability: The R package **MR.SPI** is publicly available at https://github.com/MinhaoYaooo/MR-SPI.

1 Introduction

In epidemiological studies, it is essential to infer the causal relationship of a modifiable risk factor on a health outcome of interest (Rothman and Greenland, 2005; Vandenbroucke et al., 2016). Despite the fact that randomized controlled trials (RCTs) serve as the gold standard for causal inference, it is neither feasible nor ethical to perform RCTs for many harmful exposures. Mendelian randomization (MR), which leverages the random assortment of genes from parents to offspring to mimic RCTs to establish causality in observational studies (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008; Davey Smith and Hemani, 2014). MR uses genetic variants, typically single-nucleotide polymorphisms (SNPs), as instrumental variables (IVs) to assess the causal association between an exposure and an outcome (Davey Smith and Ebrahim, 2004). Recently, many MR methods have been developed to investigate causal relationships using genome-wide association study (GWAS) summary data that consist of effect estimates of SNP-exposure and SNP-outcome associations from two non-overlapping sets of samples, which are commonly referred to as the two-sample MR methods (Burgess et al., 2013; Pierce and Burgess, 2013; Lawlor, 2016; Slob and Burgess, 2020). Since summary statistics are often publicly available and provide abundant information of associations between genetic variants and complex traits, two-sample MR methods become increasingly popular (Lawlor, 2016; Zhao et al., 2020; Morrison et al., 2020; Xue et al., 2021; Cheng et al., 2022).

Conventional MR methods require the genetic variants included in the analysis to be valid for reliable causal inference. A genetic variant is called a valid IV if the following three core assumptions hold (Didelez and Sheehan, 2007; Lawlor et al., 2008):

(A1) **Relevance**: The genetic variant is associated with the exposure;

(A2) **Effective Random Assignment**: The genetic variant is not associated with any confounder of the exposure-outcome relationship;

(A3) **Exclusion Restriction:** The genetic variant affects the outcome only through the exposure.

Among the three IV assumptions, the first assumption (A1) can be tested empirically by selecting genetic variants associated with the exposure in GWAS. However, assumptions (A2) and (A3) cannot be empirically verified in general and may be violated in practice, which leads to a biased estimate of the causal effect. The violation of (A2) may occur due to the presence of population stratification (Lawlor et al., 2008; Sanderson et al., 2021). The violation of (A3) may occur in the presence of the horizontal pleiotropy (Lawlor et al., 2008; Solovieff et al., 2013), which is a widespread biological phenomenon that the genetic

variant affects the outcome through some other biological pathway that does not involve the exposure in view (Hemani et al., 2018; Sivakumaran et al., 2011; Parkes et al., 2013; Gratten and Visscher, 2016).

Recently, several two-sample MR methods have been proposed to handle invalid IVs under certain assumptions. The Instrument Strength Independent of Direct Effect (InSIDE) assumption has been proposed and used by multiple methods, for example, the random-effects inverse-variance weighted (IVW) method (Bowden et al., 2017), MR-Egger (Bowden et al., 2015) and MR-RAPS (Robust Adjusted Profile Score) (Zhao et al., 2020). The InSIDE assumption requires that the SNP-exposure effect is asymptotically independent of the horizontal pleiotropic effect when the number of IVs goes to infinity. However, the InSIDE assumption is often not plausible in practice (Burgess and Thompson, 2017), and thus the estimate of causal effect might be biased using random-effects IVW, MR-Egger or MR-RAPS (Slob and Burgess, 2020). Another strand of methods imposes assumptions on the proportion of invalid IVs included in the analysis. For example, the weighted median method (Bowden et al., 2016) and the Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO) test (Verbanck et al., 2018) are based on the majority rule condition, which allows up to 50% of the candidate IVs to be invalid. However, the weighted median method and MR-PRESSO might produce unreliable results when more than half of the candidate IVs are invalid (Slob and Burgess, 2020). The plurality rule condition, which only requires a plurality of the candidate IVs to be valid, is weaker than the majority rule condition (Guo et al., 2018; Burgess et al., 2020). It is also termed as the ZEro Modal Pleiotropy Assumption (ZEMPA) (Hartwig et al., 2017; Slob and Burgess, 2020), and has been applied to some existing two-sample MR methods, for example, the mode-based estimation (Hartwig et al., 2017), MRMix (Qi and Chatterjee, 2019) and the contamination mixture method (Burgess et al., 2020). Among the above methods, MRMix and the contamination mixture methods require additional distributional assumptions on the genetic associations or the ratio estimates to provide reliable causal inference. Despite many efforts, most of the current MR methods require an ad-hoc set of pre-determined genetic instruments, which is often obtained by selecting genetic variants with strong SNP-exposure associations in the GWAS data (Swerdlow et al., 2016). Since the traditional way of selecting IVs only requires the exposure data, hence the same selected set of IVs is used for assessing the causal relationships between the exposure in view and different outcomes. Obviously, this one-size-fits-all exposure-specific strategy for selecting IVs might not work well for different outcomes due to complicated genetic architecture; for example, the pattern of horizontal pleiotropy might vary with different outcomes. It is thus desirable to develop an automatic algorithm to select a set of valid IVs for the specific exposure-outcome pair.

In this paper, we propose a novel two-sample MR method and algorithm that can automatically Select valid IVs for the specific exposure-outcome pair and then performs Post-selection Inference (MR-SPI) for

the causal effect. More specifically, MR-SPI contains the following four steps: (i) select relevant SNP IVs that are associated with the exposure; (ii) each selected relevant IV first provides a ratio estimate for the causal effect, and then receives votes on itself to be valid from other relevant IVs whose degrees of violation of (A2) and (A3) are small (thus more likely to be valid) under this ratio estimate of the causal effect; (iii) select valid IVs according to majority/plurality voting or by finding the maximum clique of the voting matrix; and (iv) perform post-selection inference to construct a confidence interval for the causal effect robust to finite-sample IV selection error.

To the best of our knowledge, MR-SPI is the first two-sample MR method that utilizes both exposure and outcome data to systematically select a set of exposure-outcome pair specific SNP IVs. Moreover, our proposed selection procedure does not require additional distributional assumptions to model the genetic effects or ratio estimates (Qi and Chatterjee, 2019; Burgess et al., 2020). The rest of this paper is organized as follows. In Section 2, we describe the proposed IV selection procedure and post-selection inference method. In Section 3, we conduct extensive simulations to assess the empirical performance of the proposed MR-SPI method. In Section 4, we apply MR-SPI to infer the causal relationships among 146 exposure-outcome pairs. We conclude this paper with discussions in Section 5.

2 Methodology

2.1 Two-sample GWAS Summary Statistics

Suppose that we obtain p independent SNPs $\boldsymbol{Z} = (Z_1, \cdots, Z_p)^\intercal$ by using LD clumping that keeps one representative SNP per LD region (Purcell et al., 2007). We also assume that the SNPs are standardized such that $\mathbb{E}Z_j = 0$ and $\mathrm{Var}(Z_j) = 1$ for $1 \leq j \leq p$, as in Bulik-Sullivan et al. (2015). Let D denote the exposure and Y denote the outcome. Consider the following linear structural models (Kang et al., 2016; Guo et al., 2018):

$$D = \boldsymbol{Z}^\intercal \boldsymbol{\gamma} + \delta, \tag{1}$$

$$Y = D\beta + \boldsymbol{Z}^\intercal \boldsymbol{\pi} + e, \tag{2}$$

where β represents the causal effect of interest, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^\intercal$ represents the IV strength, and $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_p)^\intercal$ encodes the violation of assumption (A2) and (A3). If assumptions (A2) and (A3) hold for SNP j, then $\pi_j = 0$ and otherwise $\pi_j \neq 0$ (see Supplementary Section S1 for details). The error terms δ and e with variances σ_δ^2 and σ_e^2 respectively are possibly correlated due to unmeasured confounding

factors. By plugging the exposure model (1) into the outcome model (2), we obtain the reduced-form outcome model:

$$Y = \boldsymbol{Z}^{\intercal}(\beta\boldsymbol{\gamma} + \boldsymbol{\pi}) + \epsilon,$$

where $\epsilon = \beta\delta + e$. Let $\boldsymbol{\Gamma} = (\Gamma_1, \cdots, \Gamma_p)^{\intercal}$ denote the SNP-outcome associations, then we have $\boldsymbol{\Gamma} = \beta\boldsymbol{\gamma} + \pi$. If $\gamma_j \neq 0$, then SNP j is called a relevant IV. If both $\gamma_j \neq 0$ and $\pi_j = 0$, then SNP j is called a valid IV. Let $\mathcal{S} = \{j : \gamma_j \neq 0, 1 \leq j \leq p\}$ denote the set of all relevant IVs, and $\mathcal{V} = \{j : \gamma_j \neq 0 \text{ and } \pi_j = 0, 1 \leq j \leq p\}$ denote the set of all valid IVs. The majority rule condition can be expressed as $|\mathcal{V}| > \frac{1}{2}|\mathcal{S}|$ (Han, 2008; Kang et al., 2016), and the plurality rule condition can be expressed as $|\mathcal{V}| > \max_{c \neq 0} |\{j \in \mathcal{S} : \pi_j/\gamma_j = c\}|$ (Guo et al., 2018). If the plurality rule condition holds, then valid IVs with the same ratio of SNP-outcome effect to SNP-exposure effect will form a plurality. Based on this key fact, our proposed MR-SPI selects the largest group of SNPs with similar ratio estimates of the causal effect as valid IVs using a voting procedure to be described in detail in Section 2.2.

Let $\widehat{\gamma}_j$ and $\widehat{\Gamma}_j$ be the estimated marginal effects of SNP j on the exposure and the outcome, and $\widehat{\sigma}_{\gamma_j}$ and $\widehat{\sigma}_{\Gamma_j}$ be the corresponding standard errors respectively. Let $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \cdots, \widehat{\gamma}_p)^{\intercal}$ and $\widehat{\boldsymbol{\Gamma}} = (\widehat{\Gamma}_1, \cdots, \widehat{\Gamma}_p)^{\intercal}$. In the two-sample setting, the summary statistics $\{\widehat{\gamma}_j, \widehat{\sigma}_{\gamma_j}\}_{1 \leq j \leq p}$ and $\{\widehat{\Gamma}_j, \widehat{\sigma}_{\Gamma_j}\}_{1 \leq j \leq p}$ are calculated from two non-overlapping samples with sample sizes n_1 and n_2 respectively. When all the SNPs are independent of each other, the joint asymptotic distribution of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ is:

$$\begin{pmatrix} \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma} \end{pmatrix} \xrightarrow{d} N\left[\mathbf{0}, \begin{pmatrix} \frac{1}{n_1}\mathbf{V}^{\gamma} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\mathbf{V}^{\Gamma} \end{pmatrix}\right], \tag{3}$$

where the diagonal entries of $\mathbf{V}^{\gamma}$ and $\mathbf{V}^{\Gamma}$ are

$$\mathbf{V}^{\gamma}_{jj} = \mathrm{Var}(Z^2_{ij})\gamma_j^2 + \sum_{l \neq j} \gamma_l^2 + \sigma_{\delta}^2 \quad \text{and} \quad \mathbf{V}^{\Gamma}_{jj} = \mathrm{Var}(Z^2_{ij})\Gamma_j^2 + \sum_{l \neq j} \Gamma_l^2 + \sigma_{\epsilon}^2,$$

and the off-diagonal entries of $\mathbf{V}^{\gamma}$ and $\mathbf{V}^{\Gamma}$ are

$$\mathbf{V}^{\gamma}_{j_1 j_2} = \gamma_{j_1}\gamma_{j_2}, \quad \mathbf{V}^{\Gamma}_{j_1 j_2} = \Gamma_{j_1}\Gamma_{j_2} \quad \text{and} \quad j_1 \neq j_2.$$

The derivation of the limiting distribution (3) can be found in Supplementary Section S2. Therefore, with the summary statistics of the exposure and the outcome, we can estimate $\frac{1}{n_1}\mathbf{V}^{\gamma}$ and $\frac{1}{n_2}\mathbf{V}^{\Gamma}$ as:

$$\frac{1}{n_1}\widehat{\mathbf{V}}^{\gamma}_{j_1 j_2} = \begin{cases} \widehat{\sigma}^2_{\gamma_{j_1}} & \text{if } j_1 = j_2, \\ \frac{1}{n_1}\widehat{\gamma}_{j_1}\widehat{\gamma}_{j_2} & \text{if } j_1 \neq j_2. \end{cases} \quad \text{and} \quad \frac{1}{n_2}\widehat{\mathbf{V}}^{\Gamma}_{j_1 j_2} = \begin{cases} \widehat{\sigma}^2_{\Gamma_{j_1}} & \text{if } j_1 = j_2, \\ \frac{1}{n_2}\widehat{\Gamma}_{j_1}\widehat{\Gamma}_{j_2} & \text{if } j_1 \neq j_2. \end{cases} \tag{4}$$

After obtaining $\{\widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{V}}^{\gamma}, \widehat{\boldsymbol{\Gamma}}, \widehat{\mathbf{V}}^{\Gamma}\}$, we can perform the proposed IV selection procedure as illustrated in Figure 1.

2.2 Selecting Valid Instruments

The first step of MR-SPI is to select relevant SNPs with large IV strength using GWAS summary statistics for the exposure. Specifically, we estimate the set of relevant IVs $\mathcal{S}$ by:

$$\widehat{\mathcal{S}} = \left\{ 1 \leq j \leq p : \frac{|\widehat{\gamma}_j|}{\widehat{\sigma}_{\gamma_j}} > \Phi^{-1}\left(1 - \frac{\alpha^*}{2}\right) \right\}, \tag{5}$$

where $\widehat{\sigma}_{\gamma_j}$ is the standard error of $\widehat{\gamma}_j$ in the summary statistics, $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution, and α^* is the user-specified threshold with the default value of 1×10^{-6}. This step is equivalent to filtering the SNPs in the exposure data with p-value $< \alpha^*$, and is adopted by most of the current two-sample MR methods to select (relevant) genetic instruments for downstream MR analysis. Note that the selected genetic instruments may not satisfy the IV independence and exclusion restriction assumptions and thus maybe invalid. In contrast, our proposed MR-SPI further incorporates the outcome data to automatically select a set of valid genetic instruments from $\widehat{\mathcal{S}}$ for a specific exposure-outcome pair.

Under the plurality rule condition, valid genetic instruments with the same ratio of SNP-outcome effect to SNP-exposure effect (i.e., Γ_j/γ_j) will form a plurality and yield "similar" ratio estimates of the causal effect. Based on this key fact, MR-SPI selects a plurality of relevant IVs whose ratio estimates are "similar" to each other as valid IVs. Specifically, we propose the following two criteria to measure the similarity between the ratio estimates of two SNPs j and k:

C1: We say the kth SNP "votes for" the jth SNP to be a valid IV if, by assuming the jth SNP is valid, the kth SNP's degree of violation of (A2) and (A3) is small;

C2: We say the ratio estimates of two SNPs j and k are "similar" if they mutually vote for each other to be valid.

The ratio estimate of the jth SNP is defined as $\widehat{\beta}^{[j]} = \widehat{\Gamma}_j/\widehat{\gamma}_j$. By assuming the jth SNP is valid, the plug-in estimate of the kth SNP's degree of violation of (A2) and (A3) can be obtained by

$$\widehat{\pi}_k^{[j]} = \widehat{\Gamma}_k - \widehat{\beta}^{[j]}\widehat{\gamma}_k = (\widehat{\beta}^{[k]} - \widehat{\beta}^{[j]})\widehat{\gamma}_k, \tag{6}$$

as we have $\Gamma_k = \beta\gamma_k + \pi_k$ for the true causal effect β, and $\widehat{\Gamma}_k = \widehat{\beta}^{[k]}\widehat{\gamma}_k$ for the ratio estimate $\widehat{\beta}^{[k]}$ of

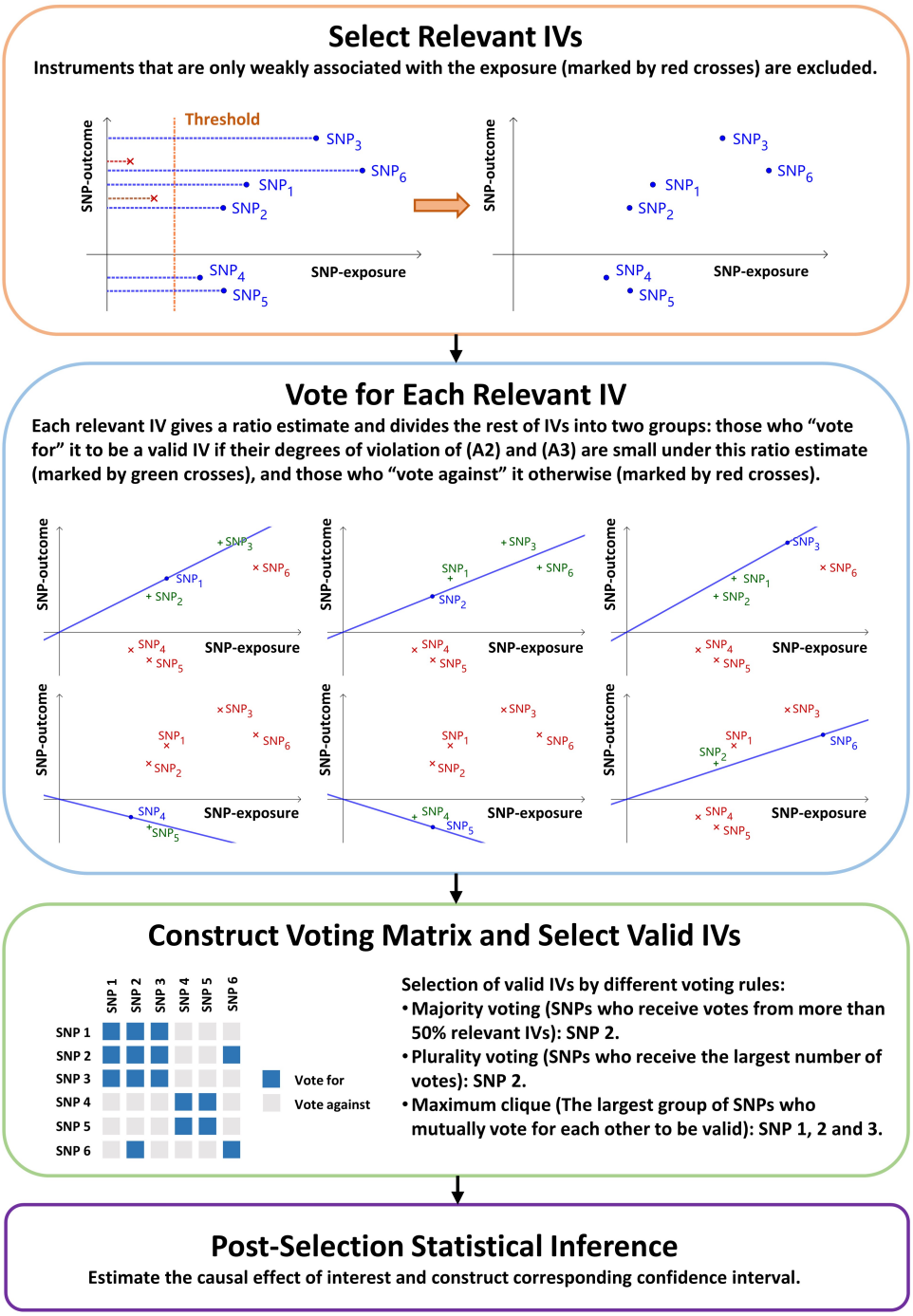

Figure 1: The framework of MR-SPI. In the first step, MR-SPI selects the relevant IVs with strong SNP-exposure associations. In the second step, each relevant IV receives votes on itself to be valid from other relevant IVs whose degrees of violation of (A2) and (A3) are small under its ratio estimate of causal effect. For example, by assuming SNP 1 is valid, the slope of the line connecting SNP 1 and the origin represents the ratio estimate of SNP 1, and SNP 2 and 3 vote for SNP 1 to be valid because they are close to that line, while SNP 4, 5 and 6 vote against it since they are far away from that line. In the third step, MR-SPI selects valid IVs according to some voting rule. In the inference step, MR-SPI estimates the causal effect and provides the corresponding confidence interval using the selected valid IVs.

the kth SNP. From equation (6), $\widehat{\pi}_k^{[j]}$ has two noteworthy implications. First, $\widehat{\pi}_k^{[j]}$ measures the difference between the ratio estimates of SNPs j and k (multiplied by the kth SNP-exposure effect estimate $\widehat{\gamma}_k$), and a small $\widehat{\pi}_k^{[j]}$ implies that the difference scaled by $\widehat{\gamma}_k$ is small. Second, $\widehat{\pi}_k^{[j]}$ represents the kth IV's degree of violation of (A2) and (A3) by regarding the jth SNP's ratio estimate $\widehat{\beta}^{[j]}$ as the true causal effect, thus a small $\widehat{\pi}_k^{[j]}$ implies a strong evidence that the kth IV supports the jth IV to be valid. Therefore, we say the kth IV votes for the jth IV to be valid if:

$$\frac{|\widehat{\pi}_k^{[j]}|}{\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]})} \leq \sqrt{\log \min(n_1, n_2)}, \tag{7}$$

where $\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]})$ is the standard error of $\widehat{\pi}_k^{[j]}$, which is given by:

$$\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]}) = \sqrt{\frac{1}{n_2}\left(\widehat{\mathbf{V}}_{kk}^{\Gamma} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)^2 \widehat{\mathbf{V}}_{jj}^{\Gamma} - 2\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\widehat{\mathbf{V}}_{jk}^{\Gamma}\right) + \frac{1}{n_1}(\widehat{\beta}^{[j]})^2\left(\widehat{\mathbf{V}}_{kk}^{\gamma} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)^2 \widehat{\mathbf{V}}_{jj}^{\gamma} - 2\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\widehat{\mathbf{V}}_{jk}^{\gamma}\right)},$$

and the term $\sqrt{\log \min(n_1, n_2)}$ in equation (7) ensures that the violation of (A2) and (A3) can be correctly detected with probability 1 as the sample sizes goes to infinity, as shown in Supplementary Section S3.

For each relevant IV in $\widehat{\mathcal{S}}$, we collect all relevant IVs' votes on whether it is a valid IV according to equation (7). Then we construct a voting matrix $\widehat{\mathbf{\Pi}} \in \mathbb{R}^{|\widehat{\mathcal{S}}|\times|\widehat{\mathcal{S}}|}$ to summarize the voting results and evaluate the similarity of two SNPs' ratio estimates according to criterion **C2**. Specifically, we define the (k, j) entry of $\widehat{\mathbf{\Pi}}$ as:

$$\widehat{\Pi}_{k,j} = \mathbf{1}\left(\max\left\{\frac{|\widehat{\pi}_k^{[j]}|}{\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]})}, \frac{|\widehat{\pi}_j^{[k]}|}{\widehat{\mathrm{SE}}(\widehat{\pi}_j^{[k]})}\right\} \leq \sqrt{\log \min(n_1, n_2)}\right). \tag{8}$$

From equation (8), we can see that the voting matrix $\widehat{\mathbf{\Pi}}$ is symmetric, and the entries of $\widehat{\mathbf{\Pi}}$ are binary: $\widehat{\Pi}_{k,j} = 1$ represents SNPs j and k vote for each other to be a valid IV, i.e., the ratio estimates of these two SNPs are close to each other; $\widehat{\Pi}_{k,j} = 0$ represents that they do not. For example, in Figure 1, $\widehat{\Pi}_{1,2} = 1$ since the ratio estimates of SNP 1 and 2 are similar, while $\widehat{\Pi}_{1,4} = 0$ because the ratio estimates of SNP 1 and 4 differ substantially.

After constructing the voting matrix $\widehat{\mathbf{\Pi}}$, we select the valid IVs by applying majority/plurality voting or finding the maximum clique of the voting matrix(Ouyang et al., 1997). Let $\mathbf{VM}_k = \sum_{j\in\widehat{\mathcal{S}}} \widehat{\Pi}_{k,j}$ be the total number of SNPs whose ratio estimates are similar to SNP k. For example, $\mathbf{VM}_1 = 3$ in Figure 1, since 3 SNPs (including SNP 1 itself) yield similar ratio estimates to SNP 1 according to criterion **C2**. A large $\mathbf{VM}_k$ implies a strong evidence that SNP k is a valid IV, since we assume that valid IVs form

a plurality of the relevant IVs. A relevant IV in $\widehat{\mathcal{S}}$ receives a majority voting if more than 50% SNPs in $\widehat{\mathcal{S}}$ vote for it to be valid, and receives a plurality voting if it receives the largest number of votes among all relevant SNPs in $\widehat{\mathcal{S}}$. Let $\widehat{\mathcal{V}}_M = \left\{k \in \widehat{\mathcal{S}} : \mathbf{VM}_k > |\widehat{\mathcal{S}}|/2\right\}$ denote the set of IVs with majority voting, and $\widehat{\mathcal{V}}_P = \left\{k \in \widehat{\mathcal{S}} : \mathbf{VM}_k = \max_{l \in \widehat{\mathcal{S}}} \mathbf{VM}_l\right\}$ denote the set of IVs with plurality voting, then the union $\widehat{\mathcal{V}} = \widehat{\mathcal{V}}_M \cup \widehat{\mathcal{V}}_P$ can be a robust estimate for $\mathcal{V}$ in practice. Alternatively, we can also find the maximum clique in the voting matrix as an estimate for $\mathcal{V}$. A clique in the voting matrix is a group of IVs who mutually vote for each other to be valid, and a maximum clique is a clique with the largest possible number of IVs. We will show in simulation studies that selecting valid IVs by finding the maximum clique in the voting matrix can empirically offer low false discovery rate (FDR) (Benjamini and Hochberg, 1995) and high true positive proportion (TPP).

2.3 Estimation and Inference of the Causal Effect

After selecting the set of valid genetic instruments $\widehat{\mathcal{V}}$ by majority/plurality voting or finding the maximum clique in the voting matrix, the causal effect β is estimated by:

$$\widehat{\beta}_{\text{SPI}} = \frac{\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}}{\widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}}, \tag{9}$$

and the variance of $\widehat{\beta}_{\text{SPI}}$ is estimated by:

$$\widehat{\text{Var}}\left(\widehat{\beta}_{\text{SPI}}\right) = \left(\frac{\mathbb{E}\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}}{\mathbb{E}\widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}}\right)^2 \left(\frac{\text{Var}(\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})}{(\mathbb{E}\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})^2} + \frac{\text{Var}(\widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})}{(\mathbb{E}\widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})^2} - 2\frac{\text{Cov}(\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}, \widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})}{(\mathbb{E}\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})(\mathbb{E}\widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})}\right), \tag{10}$$

where $\widehat{\gamma}_{\widehat{\mathcal{V}}}$ and $\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}$ are the estimates of SNP-exposure associations and SNP-outcome associations of the instruments in $\widehat{\mathcal{V}}$, respectively. The two expectation terms $\mathbb{E}\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}$ and $\mathbb{E}\widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}$ in equation (10) can be approximated by $\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}$ and $\widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}} + \sum_{j \in \widehat{\mathcal{V}}} \widehat{\sigma}_{\gamma_j}^2$, respectively. The variance-covariance matrix of $(\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}}, \widehat{\gamma}_{\widehat{\mathcal{V}}}^{\intercal} \widehat{\gamma}_{\widehat{\mathcal{V}}})$ can be obtained by the delta method (Oehlert, 1992), as shown in Supplementary Section S4. Let $\alpha \in (0, 1)$ be the significance level and $z_{1-\alpha/2}$ be the $(1 - \alpha/2)$-quantile of the standard normal distribution, then the $(1 - \alpha)$ confidence interval for β is given by:

$$\text{CI} = \left(\widehat{\beta}_{\text{SPI}} - z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\widehat{\beta}_{\text{SPI}})}, \widehat{\beta}_{\text{SPI}} + z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\widehat{\beta}_{\text{SPI}})}\right). \tag{11}$$

Under the plurality rule condition, as $\min\{n_1, n_2\} \to \infty$, we have:

$$\mathbb{P}\left\{\beta \in \left(\widehat{\beta}_{\text{SPI}} - z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\widehat{\beta}_{\text{SPI}})}, \widehat{\beta}_{\text{SPI}} + z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\text{Var}}(\widehat{\beta}_{\text{SPI}})}\right)\right\} \to 1 - \alpha. \tag{12}$$

The derivation of equation (12) can be found in Supplementary Section S5. Hence, MR-SPI provides a theoretical guarantee for the asymptotic coverage probability of the confidence interval under the plurality rule condition.

We summarize the procedure of selecting valid IVs and constructing the corresponding confidence interval by MR-SPI in Algorithm 1.

Algorithm 1: Selection of Valid Instruments and Inference by MR-SPI

input : GWAS summary statistics of independent SNPs $\{\widehat{\gamma}_j, \widehat{\sigma}_{\gamma_j}, \widehat{\Gamma}_j, \widehat{\sigma}_{\Gamma_j}\}_{1\le j\le p}$; Sample sizes n_1 for the exposure and n_2 for the outcome; Threshold α^* for selecting relevant IVs; Significance level $\alpha \in (0,1)$.

output: An estimate of the set of valid IVs $\widehat{\mathcal{V}}$, the causal effect estimate $\widehat{\beta}_{\mathrm{SPI}}$ and the corresponding confidence interval CI.

Estimate the variance-covariance matrices $\widehat{\mathbf{V}}^{\gamma}$ and $\widehat{\mathbf{V}}^{\Gamma}$ as in equation (4);
Select the set of relevant IVs $\widehat{\mathcal{S}}$ as in equation (5);
for $j \in \widehat{\mathcal{S}}$ **do**
  Calculate $\widehat{\beta}^{[j]} = \widehat{\Gamma}_j/\widehat{\gamma}_j$ and $\widehat{\pi}_k^{[j]} = \widehat{\Gamma}_k - \widehat{\beta}^{[j]}\widehat{\gamma}_k \quad \text{for} \quad k \in \widehat{\mathcal{S}}$;
  Each relevant IV $k \in \widehat{\mathcal{S}}$ votes for the jth IV to be valid if $|\widehat{\pi}_k^{[j]}|/\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]}) \le \sqrt{\log\min(n_1, n_2)}$;
end
Construct the symmetric voting matrix $\widehat{\mathbf{\Pi}} \in \mathbb{R}^{|\widehat{\mathcal{S}}|\times|\widehat{\mathcal{S}}|}$ as in equation (8);
Select the set of valid IVs $\widehat{\mathcal{V}}$ by majority voting, plurality voting or finding the maximum clique in the voting matrix;
Estimate the causal effect and construct the corresponding confidence interval as in equations (9) and (11) using IVs in $\widehat{\mathcal{V}}$.

2.4 A Robust Confidence Interval via Searching and Sampling

In finite-sample settings, the selected set of relevant IVs $\widehat{\mathcal{S}}$ might include some invalid IVs whose degrees of violation of (A2) and (A3) are small but nonzero, and we refer to them as "locally invalid IVs" Guo (2021). When locally invalid IVs exist and are incorrectly selected into $\widehat{\mathcal{V}}$, the confidence interval in equation (11) becomes unreliable, since its validity (i.e., the coverage probability attains the nominal level) requires that the invalid IVs are correctly identified. In practice, we can multiply the threshold $\sqrt{\log\min(n_1, n_2)}$ in the right-hand side of equation (7) by a scaling factor η to examine whether the confidence interval calculated by equation (11) is sensitive to the choice of the threshold. If the confidence interval varies substantially to the choice of the scaling fator η, then there might exist finite-sample IV selection error especially with locally invalid IVs. We demonstrate this issue with two numerical examples below: Figure 2(a) shows an example in which MR-SPI provides robust inference across difference values

of the scaling factor, while Figure 2(b) shows an example that MR-SPI might suffer from IV selection error, as the causal effect estimate and the corresponding confidence interval are sensitive to the choice of the scaling factor η. This issue motivates us to develop a more robust confidence interval.

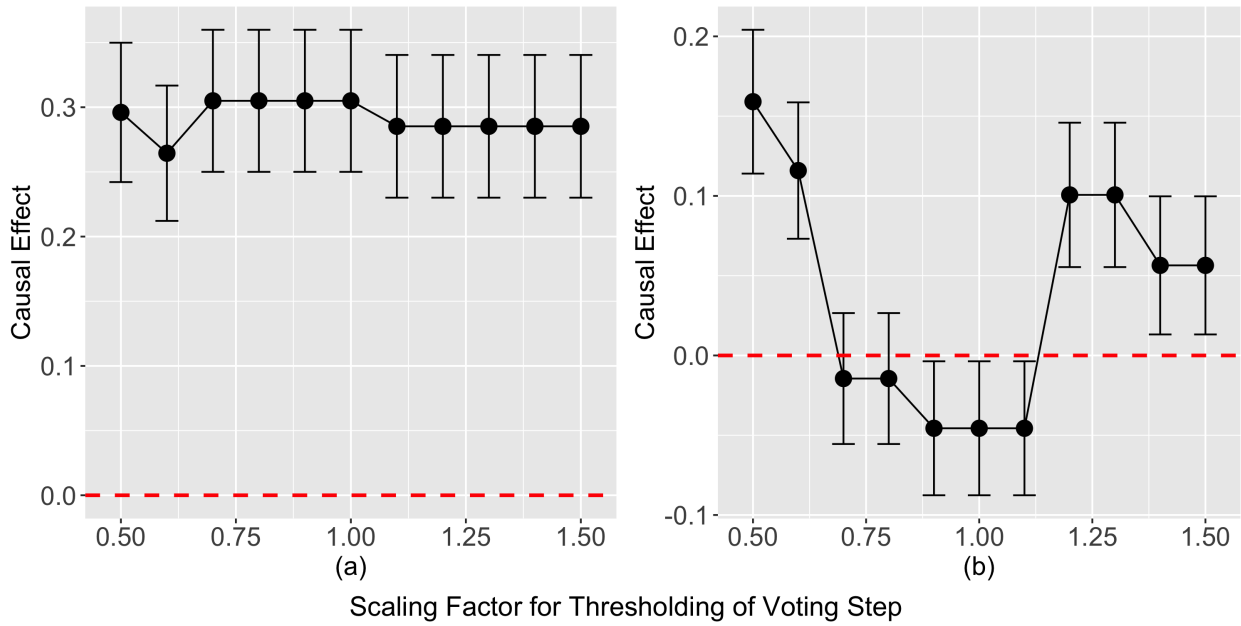

Figure 2: Two examples of sensitivity analysis of MR-SPI's voting procedure. (a) MR-SPI provides stable inference in this example, as the causal effect estimate and the corresponding confidence interval do not change significantly with the scaling factor. (b) MR-SPI might suffer from potential IV selection error in this example, since the causal effect estimate and the corresponding confidence interval are susceptible to the choice of the scaling factor. For example, MR-SPI provides a significantly positive estimate of the causal effect when $\eta = 0.5$, while the estimate becomes negative when $\eta = 1$.

To construct a confidence interval that is robust to finite-sample IV selection error, we borrow the idea of searching and sampling proposed by Guo (2021), with main steps described in Figure 3. The key idea is to sample the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ repeatedly from the following distribution:

$$\begin{pmatrix} \widehat{\boldsymbol{\gamma}}^{(m)} \\ \widehat{\boldsymbol{\Gamma}}^{(m)} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left[\begin{pmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\Gamma}} \end{pmatrix}, \begin{pmatrix} \frac{1}{n_1}\widehat{\mathbf{V}}^{\gamma} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\widehat{\mathbf{V}}^{\Gamma} \end{pmatrix}\right], \quad m = 1, \cdots, M, \tag{13}$$

where M is the number of sampling times. Since $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ follow distributions centered at $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$, there exists m^* such that $\widehat{\boldsymbol{\gamma}}^{(m^*)}$ and $\widehat{\boldsymbol{\Gamma}}^{(m^*)}$ are close enough to the true genetic effects $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ when the number of sampling times M is sufficiently large, and thus the confidence interval obtained by using $\widehat{\boldsymbol{\gamma}}^{(m^*)}$ and $\widehat{\boldsymbol{\Gamma}}^{(m^*)}$ instead of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\Gamma}}$ might have a larger probability of covering β.

For each sampling, we construct the confidence interval by searching over a grid of β values such that more than half of the instruments in $\widehat{\mathcal{V}}$ are detected as valid. As for the choice of grid, we start with the

smallest interval $[L, U]$ that contains all the following intervals:

$$\left(\widehat{\beta}^{[j]} - \sqrt{\log\min(n_1, n_2)\widehat{\mathrm{Var}}(\widehat{\beta}^{[j]})}, \widehat{\beta}^{[j]} + \sqrt{\log\min(n_1, n_2)\widehat{\mathrm{Var}}(\widehat{\beta}^{[j]})}\right) \quad \text{for} \quad j \in \widehat{\mathcal{V}},$$

where $\widehat{\beta}^{[j]}$ is the ratio estimate of the jth SNP, $\widehat{\mathrm{Var}}(\widehat{\beta}^{[j]}) = \left(\widehat{\mathbf{V}}^{\Gamma}_{jj}/n_2 + (\widehat{\beta}^{[j]})^2\widehat{\mathbf{V}}^{\gamma}_{jj}/n_1\right)/\widehat{\gamma}_j^2$ is the variance of $\widehat{\beta}^{[j]}$, and $\sqrt{\log\min(n_1, n_2)}$ serves the same purpose as in equation (7). Then we discretize $[L, U]$ into $\mathcal{B} = \{b_1, b_2, \cdots, b_K\}$ as the grid set such that $b_1 = L, b_K = U$ and $|b_{k+1} - b_k| = n^{-0.6}$ for $1 \leq k \leq K - 2$. We set the grid size $n^{-0.6}$ so that the error caused by discretization is smaller than the parametric rate $n^{-1/2}$.

For each grid value $b \in \mathcal{B}$ and sampling index $1 \leq m \leq M$, we propose an estimate of the degree of violation of (A2) and (A3) of the jth SNP by:

$$\widehat{\pi}_j^{(m)}(b) = \left(\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)}\right) \cdot \mathbf{1}\left(|\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)}| \geq \lambda\widehat{\rho}_j(b, \alpha)\right), \quad j \in \widehat{\mathcal{V}}, \tag{14}$$

where $\widehat{\rho}_j(b, \alpha) = \Phi^{-1}\left(1 - \frac{\alpha}{2|\widehat{\mathcal{V}}|}\right)\sqrt{\left(\widehat{\mathbf{V}}^{\Gamma}_{jj}/n_2 + b^2\widehat{\mathbf{V}}^{\gamma}_{jj}/n_1\right)}$ is a data-dependent threshold, $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the standard normal distribution, $\alpha \in (0, 1)$ is the significance level, and $\lambda = (\log\min\{n_1, n_2\}/M)^{\frac{1}{2|\widehat{\mathcal{V}}|}}$ ($\lambda < 1$ when M is sufficiently large) is a scaling factor to make the thresholding more stringent so that the confidence interval in each sampling is shorter, which we will see later in equation (15). Here, $\widehat{\pi}_j^{(m)}(b) = 0$ indicates that the jth SNP is detected as a valid IV in the mth sampling if we take $\{\widehat{\boldsymbol{\gamma}}^{(m)}, \widehat{\boldsymbol{\Gamma}}^{(m)}\}$ as the estimates of genetic effects and b as the true causal effect. Let $\widehat{\boldsymbol{\pi}}^{(m)}_{\widehat{\mathcal{V}}}(b) = (\widehat{\pi}_j^{(m)}(b))_{j\in\widehat{\mathcal{V}}}$, then we construct the mth sampling's confidence interval $\mathrm{CI}^{(m)}$ by searching for the smallest and largest $b \in \mathcal{B}$ such that more than half of SNPs in $\widehat{\mathcal{V}}$ are detected to be valid according to equation (14), i.e.,

$$\beta^{(m)}_{\min} = \min_{b\in\mathcal{B}:\|\widehat{\boldsymbol{\pi}}^{(m)}_{\widehat{\mathcal{V}}}(b)\|_0 < |\widehat{\mathcal{V}}|/2} b, \quad \beta^{(m)}_{\max} = \max_{b\in\mathcal{B}:\|\widehat{\boldsymbol{\pi}}^{(m)}_{\widehat{\mathcal{V}}}(b)\|_0 < |\widehat{\mathcal{V}}|/2} b, \quad \text{and} \quad \mathrm{CI}^{(m)} = \left(\beta^{(m)}_{\min}, \beta^{(m)}_{\max}\right). \tag{15}$$

From equations (14) and (15), we can see that, when λ is smaller, there will be fewer SNPs in $\widehat{\mathcal{V}}$ being detected as valid for a given $b \in \mathcal{B}$, which leads to less $b \in \mathcal{B}$ satisfying $\|\widehat{\boldsymbol{\pi}}^{(m)}_{\widehat{\mathcal{V}}}(b)\|_0 < |\widehat{\mathcal{V}}|/2$, thus the confidence interval in each sampling will be shorter. If there does not exist $b \in \mathcal{B}$ such that the majority of IVs in $\widehat{\mathcal{V}}$ are detected as valid, we set $\mathrm{CI}^{(m)} = \varnothing$. Let $\mathcal{M} = \{1 \leq m \leq M : \mathrm{CI}^{(m)} \neq \varnothing\}$ denote the set of all sampling indexes corresponding to non-empty searching confidence intervals, then the proposed

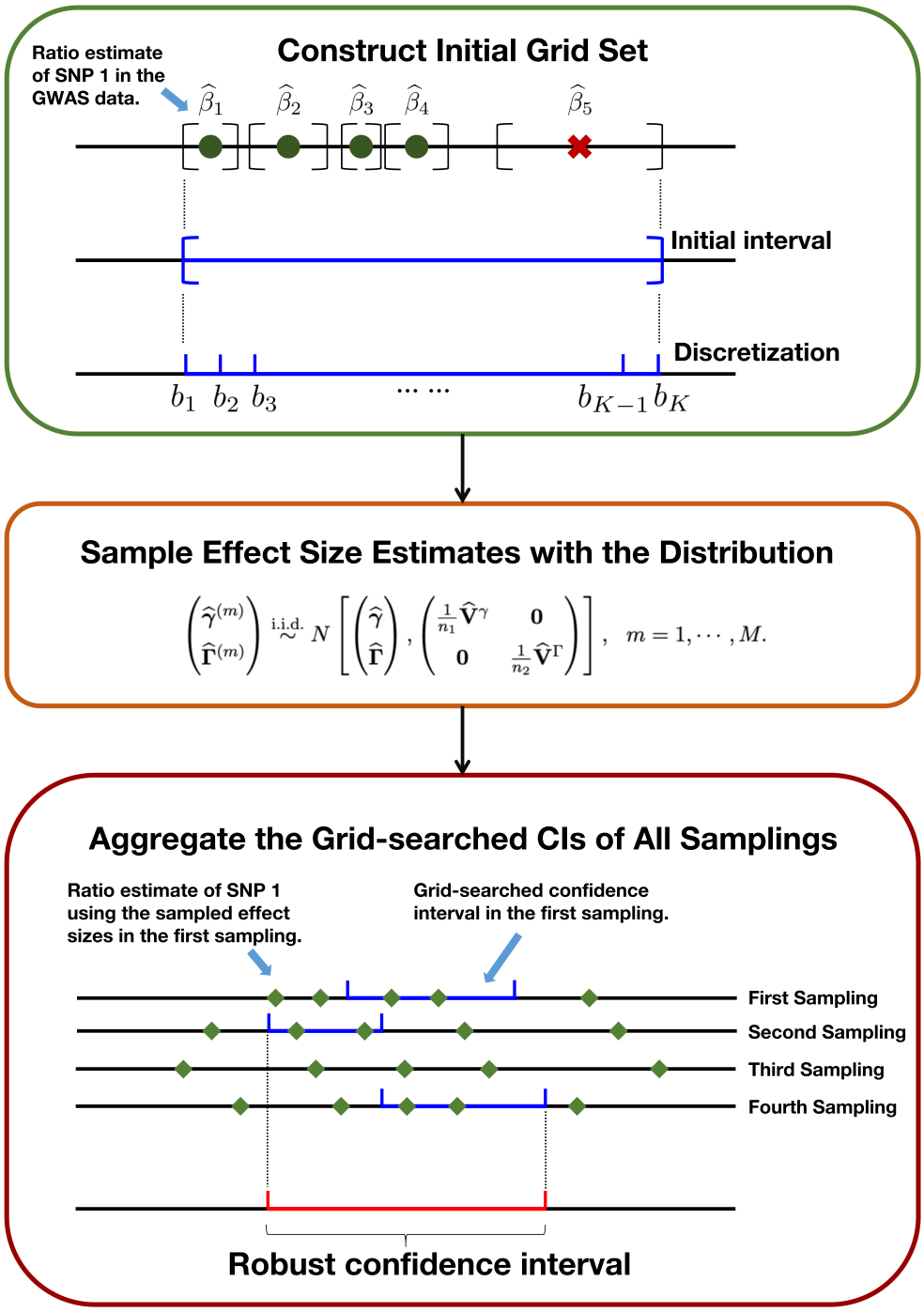

Figure 3: The procedure of constructing the robust confidence interval by MR-SPI. When locally invalid IVs exist in finite samples, MR-SPI might incorrectly select invalid IVs (marked by the red cross). In such cases, a robust confidence interval can be constructed to improve the coverage probability. First, we construct an initial interval using SNPs in $\widehat{\mathcal{V}}$ and discretize it to a grid set. Second, we repeatedly sample the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$. Third, we find a confidence interval for each sampling (marked by blue line segments) by grid search, and then aggregate these confidence intervals to construct the robust confidence interval (marked by the red line segment). Note that the confidence interval in the third sampling is empty since the majority rule is violated.

robust confidence interval is given by:

$$\mathrm{CI}^{\mathrm{robust}} = \left(\min_{m \in \mathcal{M}} \beta_{\min}^{(m)}, \max_{m \in \mathcal{M}} \beta_{\max}^{(m)} \right). \tag{16}$$

We summarize the procedure of constructing the proposed robust confidence interval in Algorithm 2.

Algorithm 2: Constructing A Robust Confidence Interval via Searching and Sampling

input : GWAS summary statistics of independent SNPs $\{\widehat{\gamma}_j, \widehat{\sigma}_{\gamma_j}, \widehat{\Gamma}_j, \widehat{\sigma}_{\Gamma_j}\}_{1 \le j \le p}$; Sample sizes n_1 for the exposure and n_2 for the outcome; Threshold α^* for selecting relevant IVs; Significance level $\alpha \in (0, 1)$; Sampling number M.

output: The robust confidence interval $\mathrm{CI}^{\mathrm{robust}}$.

```
Estimate the set of valid IVs V̂ as in Algorithm 1;
Construct the initial interval [L, U] as in equation (2.4) and obtain the corresponding grid set B;
for m ← 1 to M do
    Sample γ̂^(m) and Γ̂^(m) from the distribution in equation (13);
    Calculate {π̂^(m)_V̂(b)}_{b∈B} as in equation (14);
    Construct CI^(m) by grid search as in equation (15);
end
Construct the robust confidence interval CI^robust as in equation (16);
```

3 Simulation Studies

We conduct extensive simulations to evaluate the performance of MR-SPI in the presence of invalid IVs. We set the number of candidate IVs $p = 10$ and the sample sizes $n_1 = n_2 \in \{5000, 10000, 20000, 40000, 80000\}$. We generate the jth genetic instruments Z_j and X_j independently from a binomial distribution $\mathrm{Bin}(2, f_j)$, where $f_j \sim U(0.05, 0.50)$ is the minor allele frequency of SNP j. Then we generate the exposure $\boldsymbol{D} = (D_1, \cdots, D_{n_1})^\intercal$ and the outcome $\boldsymbol{Y} = (Y_1, \cdots, Y_{n_2})^\intercal$ according to models (1) and (2) respectively. Finally, we calculate the genetic associations and their corresponding standard errors for the exposure and the outcome, respectively. As for the parameters in models (1) and (2), we fix the causal effect $\beta = 1$, and we consider 4 scenarios for $\boldsymbol{\gamma} \in \mathbb{R}^p$ and $\boldsymbol{\pi} \in \mathbb{R}^p$:

S1 : set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\intercal$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_6, \mathbf{1}_4)^\intercal$.

S2 : set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\intercal$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_4, \mathbf{1}_3, -\mathbf{1}_3)^\intercal$.

S3 : set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\intercal$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_6, \mathbf{1}_2, 0.25, 0.25)^\intercal$.

S4 : set $\boldsymbol{\gamma} = 0.2 \cdot (\mathbf{1}_5, -\mathbf{1}_5)^\intercal$ and $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_4, \mathbf{1}_2, 0.25, \mathbf{1}_2, -0.25)^\intercal$.

Scenarios **S1** and **S3** satisfy the majority rule condition, while **S2** and **S4** only satisfy the plurality rule condition. In addition, **S3** and **S4** simulate the cases where locally invalid IVs exist, as we shrink some of the SNPs' violation degrees of (A2) and (A3) down to 0.25 times in these two scenarios. In total, we run 1000 replications in each scenario.

We compare the bias, empirical coverage and average length of 95% confidence interval of MR-SPI to the following competing methods: (i) the random-effects IVW method that performs random-effects meta-analysis to account for pleiotropy (Bowden et al., 2017), (ii) MR-RAPS that assumes pleiotropic effects are normally distributed and applies the maximum profile likelihood estimation to obtain the causal effect estimate (Zhao et al., 2020), (iii) MR-PRESSO that detects the invalid IVs by checking whether the residual sum of squares of the regression decreases substantially after omitting some genetic variants (Verbanck et al., 2018), (iv) the weighted median method that takes the weight median of the ratio estimates as the causal effect (Bowden et al., 2016), (v) the mode-based estimation that takes the mode of the smoothed empirical density function of the ratio estimates as the causal effect (Hartwig et al., 2017), (vi) MRMix that models the SNP-exposure and SNP-outcome effects with a bivariate normal mixture distribution (Qi and Chatterjee, 2019) and (vii) the contamination mixture method that models the ratio estimates of candidate IVs with a normal mixture distribution (Burgess et al., 2020). We exclude MR-Egger in this simulation since it is heavily biased under our simulation settings. Among the above methods, the random-effects IVW method and MR-RAPS require the InSIDE assumption, MR-PRESSO and the weighted median method require the majority rule condition, while MR-SPI, the mode-based estimation, MRMix and the contamination mixture method require the plurality rule condition (or ZEMPA assumption). For simplicity, we shall use IVW to represent the random-effects IVW method in the remaining of this paper.

We present the bias of those MR methods in Figure 4 and Supplementary Table S1. Generally, the proposed MR-SPI has small bias in all four scenarios. IVW and MR-RAPS are biased since the InSIDE assumption does not hold in our parameter settings. Interestingly, the biases of these two methods are smaller in scenarios **S3** and **S4** compared to **S1** and **S2**, as the degree of violation of (A2) and (A3) is generally smaller when some of the candidate IVs are only locally invalid. MR-PRESSO generally yields biased estimates as it fails to remove outliers in most of our settings. The weighted median estimator is biased when only the plurality rule condition holds (scenario **S2**), since it requires more than half of the candidate IVs to be valid. The biases of the mode-based estimation, MRMix and the contamination mixture method are all small, as these three methods only require the plurality rule condition to hold, which is satisfied in all four scenarios.

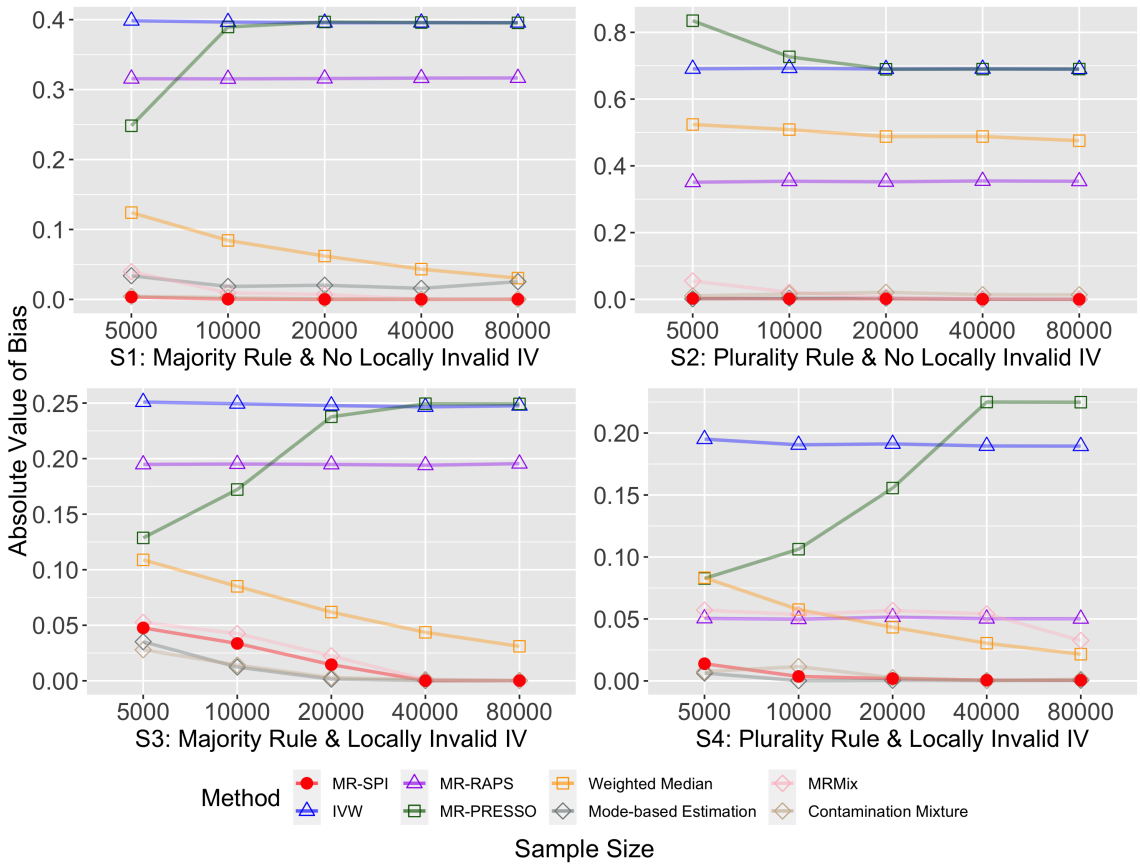

Figure 4: Absolute value of the bias of estimated causal effect using different MR methods under different scenarios.

Figure 5 and Supplementary Table S2 report the empirical coverage of the above methods. Under scenarios **S1** and **S2** where no locally invalid IV exists, the confidence interval of MR-SPI can attain 95% coverage level even when the sample sizes are small (e.g., $n_1 = n_2 = 5000$). In the presence of locally invalid IVs, i.e., under scenarios **S3** and **S4**, the empirical coverage of MR-SPI's confidence interval can still attain the nominal level when $n_1 = n_2 = 80000$, as MR-SPI can correctly distinguish the locally invalid IVs from the valid IVs when sample sizes are large enough. However, MR-SPI fails to identify those locally invalid IVs under **S3** and **S4** if the sample sizes are small, and therefore the empirical coverage of MR-SPI is lower than 95%. In such cases, we suggest using the robust confidence interval constructed by Algorithm 2, which attains the 95% coverage level and thus is less vulnerable to the IV selection error in finite samples. The empirical coverage of the weighted median method is lower than MR-SPI even in scenario **S1** where the majority rule condition holds, and the similar coverage issue of the median estimator also occurred in Guo et al. (2018). Compared to the confidence interval of MR-SPI, the mode-based estimation is generally more conservative with coverage above the nominal level in our simulation settings, which is the price to pay for being less affected by the invalid instruments, as

discussed in Hartwig et al. (2017). Both MRMix and the contamination mixture method cannot attain the 95% coverage level in most scenarios. These two methods make distributional assumptions for either the genetic associations or the ratio estimates, which might be violated in our simulation settings, and thus the coverage levels are below the nominal level.

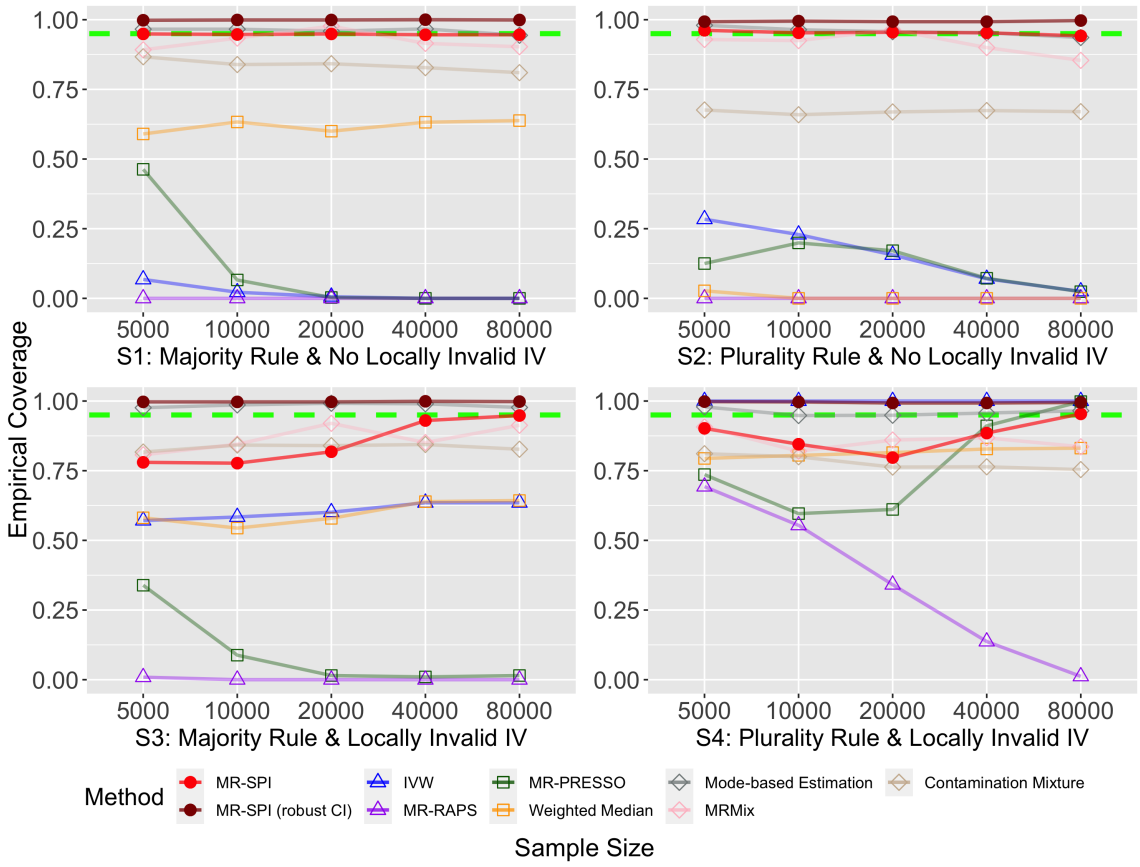

Figure 5: Empirical coverage of different MR methods under different scenarios. The green dashed line represents the nominal level (95%).

We report the average length of 95% confidence interval of the above methods in Figure 6 and Supplementary Table S3. Although MR-RAPS generally has the shortest confidence interval, it is biased and the coverage level is low, as the InSIDE assumption does not hold in our simulation settings. Among the methods except MR-RAPS, MR-SPI has the shortest confidence intervals under most scenarios. The average length of confidence interval of IVW is not significantly decreasing as the sample sizes increase, since we apply the random-effects IVW method here, which scales up the standard error of the causal effect estimate when heterogeneity in the ratio estimates exists (Bowden et al., 2017). In scenario **S4**, MR-PRESSO has longer confidence intervals as the sample sizes increase, since it tends to treat none of the candidate SNPs as outlier under this simulation setting, i.e., when the majority rule does not hold and locally invalid IV exists. When no outlier is identified, MR-PRESSO uses all candidate SNPs, and

thus the standard error of MR-PRESSO under **S4** will be close to that of IVW when the sample sizes are large.

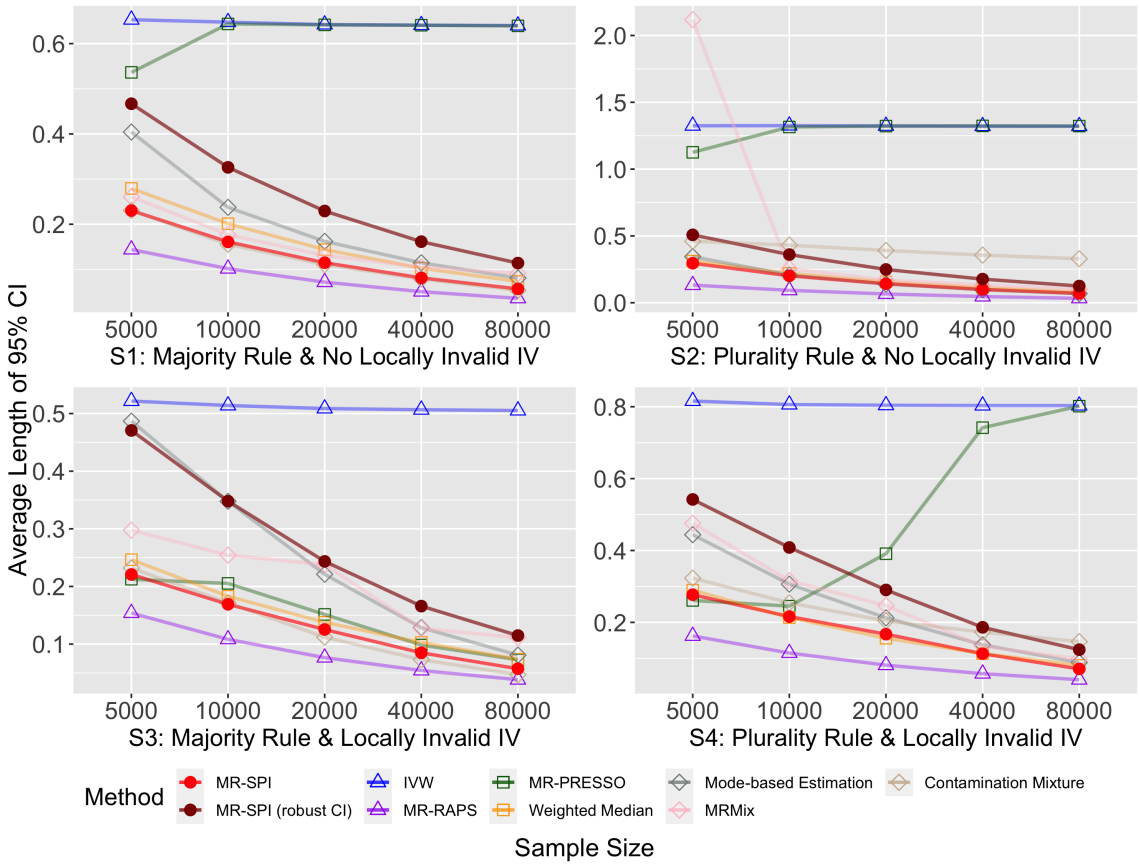

Figure 6: Average length of 95% confidence interval of the causal effect by different MR methods under different scenarios.

From previous simulation results, we notice that the average length of MR-SPI's robust confidence interval constructed by Algorithm 2 is always longer than the default confidence interval calculated by equation (11). Therefore, we conduct a simulation study to explore in what case the length of the robust confidence interval would be closer to the default confidence interval. Specifically, we set the sample sizes $n_1 = n_2 = 10000$ and the number of SNPs $p = 10$, and fix the IV strength to be $0.2 \cdot \mathbf{1}_{10}^{\intercal}$. We set $\boldsymbol{\pi} = \tau \cdot (\mathbf{0}_4, \mathbf{1}_3, -\mathbf{1}_3)^{\intercal}$, where $\tau \in [0, 0.3]$ is the violation strength to control the degree of violation of (A2) and (A3). We run 1000 replications in total and present the ratio of average lengths of robust confidence interval to default confidence interval in Figure 7. As the violation strength τ increases, the ratio of average lengths of these two confidence intervals decreases at first and then increases to a stable level. When the violation strength τ is small, i.e., the degrees of violation of (A2) and (A3) is small, the default confidence interval would not be much affected even if invalid IVs are incorrectly selected. When the violation strength τ is large, the invalid IVs are more likely to be detected by MR-SPI, and thus

the default confidence interval would also be less influenced by the invalid IVs. Compared to the default confidence interval, the robust confidence interval is less affected by the violation strength τ, indicating its robustness to the IV selection error. The minimum of this ratio is around 1.5, which is the price to pay for robustness in this numerical example.

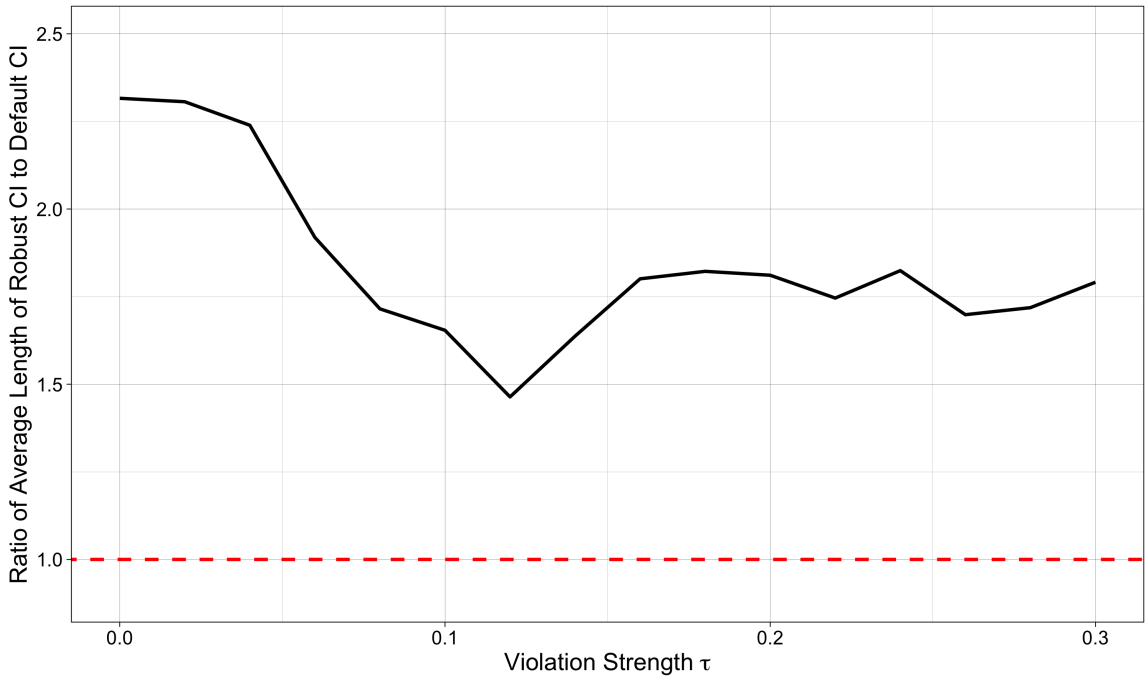

Figure 7: Ratio of the average length of 95% confidence interval of the robust confidence interval to the default confidence interval across a range of violation strength τ. The robust confidence interval is calculated by Algorithm 2. The default confidence interval is calculated by the limit distribution of the causal effect estimate using the selected set of valid instruments $\widehat{\mathcal{V}}$.

We also report the FDR and TPP of the selection of valid IVs by MR-SPI under scenarios **S1**-**S4** in Table 1. In our simulation, we select valid IVs by finding the maximum clique in the voting matrix. Generally, the TPP of MR-SPI is close to 1 under all scenarios, while the FDR of MR-SPI is close to 0 if no locally invalid IV exists and the plurality rule condition holds. Under scenarios **S3** and **S4**, MR-SPI fails to distinguish between valid IVs and those locally invalid IVs when the sample sizes are small (e.g. 5000), and therefore the FDR is relatively high. As the sample sizes increase, the FDR of MR-SPI would be close to 0 even in the presence of locally invalid IVs, suggesting that MR-SPI can correctly identify the locally invalid IVs when the sample sizes are large.

Table 1: The FDR and TPP of valid IV selection by MR-SPI under different scenarios and sample sizes. The FDR is close to 0 in the absence of locally invalid IV or when the sample sizes are large. The TPP is close to 1 under all scenarios.

Sample Sizes	FDR				TPP			
	S1	**S2**	**S3**	**S4**	**S1**	**S2**	**S3**	**S4**
5000	0.000	0.018	0.225	0.298	0.996	0.998	0.987	0.996
10000	0.000	0.009	0.198	0.251	0.998	1.000	0.981	0.997
20000	0.000	0.005	0.124	0.195	0.999	1.000	0.987	0.999
40000	0.000	0.006	0.022	0.072	1.000	1.000	0.997	0.999
80000	0.000	0.005	0.000	0.005	0.999	0.999	1.000	1.000

The simulation studies demonstrate that MR-SPI performs well compared to the other competing MR methods under the plurality rule condition. When no locally invalid IV exists, MR-SPI can select the valid IVs correctly and provide nearly unbiased estimates of the causal effect, and the confidence interval of MR-SPI can attain the nominal coverage level. If locally invalid IVs exist and are incorrectly selected for analysis in finite samples, the robust confidence interval of MR-SPI is recommended for use, which is less vulnerable to the IV selection error.

4 Real Data Applications

4.1 Two Benchmark Datasets

In this section, we apply the proposed MR-SPI method to two benchmark real datasets to evaluate its performance. These two real datasets serve as the benchmark because the exposure and the outcome are the same trait in each dataset, and thus the horizontal pleiotropic effects are expected to be zero. In the first benchmark real data analysis, both the exposure and the outcome are coronary artery disease (CAD), and we refer to it as the CAD-CAD dataset. Since both the exposure and the outcome are CAD, the causal effect is expected to be one. The exposure data come from the Coronary Artery Disease (C4D) Genetics Consortium (The Coronary Artery Disease (C4D) Genetics Consortium, 2011), and the outcome data come from the Coronary ARtery DIsease Genome-wide Replication and Meta-analysis (CARDIoGRAM) consortium (Schunkert et al., 2011). We first clump the SNPs in the exposure data using the software Plink (Purcell et al., 2007) with $r^2 < 1 \times 10^{-2}$ to obtain the independent genetic instruments, and then use 1×10^{-6} as the p-value threshold to select relevant instruments. In total, 5 relevant instruments are included for downstream analysis. We compare MR-SPI to the other 8 competing MR methods including IVW, MR-Egger, MR-RAPS, MR-PRESSO, the weighted median method, the mode-based estimation,

MRMix and the contamination mixture method, and the causal effect estimates and the corresponding 95% confidence intervals using the above methods are presented in Figure 8(a). Generally, the confidence intervals of MR-SPI, IVW and MR-Egger all cover 1, and MR-SPI provides the shortest confidence interval among these three methods. Besides, none of the relevant IVs is excluded in the voting step, which is in line with the expectation that horizontal pleiotropy might not exist in this dataset.

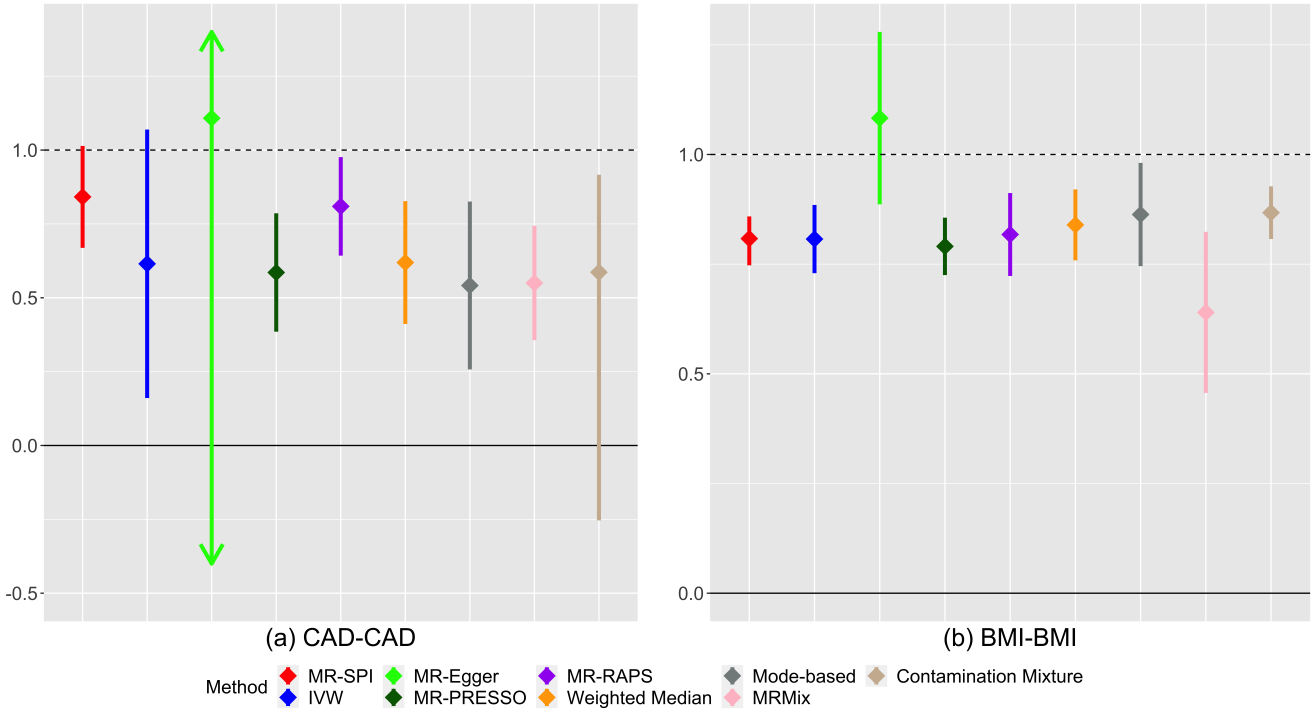

Figure 8: Point estimates and 95% confidence intervals for the causal effects of (a) CAD-CAD dataset and (b) BMI-BMI dataset using different MR methods. Confidence intervals are clipped to arrows if they exceed axis limits. CAD: coronary artery disease; BMI: body mass index.

The second real data analysis is based on the GWAS studies of body mass index (BMI), and we refer to it as the BMI-BMI dataset. The exposure data and the outcome data are the BMI GWAS data for physically active men and women respectively, both of which come from the GIANT consortium (Locke et al., 2015). After LD clumping and filtering SNPs with the same parameters as in the CAD-CAD dataset, 64 candidate SNPs are selected as relevant IVs and none of them is detected to be invalid by MR-SPI. The point estimates of the causal effect and corresponding 95% confidence intervals using the above methods are shown in Figure 8(b). Overall, all the above methods except MR-Egger provide causal effect estimates that are significantly below one. As discussed in previous studies (Locke et al., 2015), some significant loci of BMI might exhibit heterogeneity in genetic effects between men and women. Therefore, the "true" effect might not be equal to one in this dataset due to the gender difference in the genetic architecture of BMI.

4.2 Learning Causal Relationships of 146 Exposure-Outocme Pairs

In this section, we examine the causal relationships among 14 complex traits from four categories including ischemic stroke, cholesterol levels, heart disease, and Coronavirus disease 2019 (COVID-19) related traits. Since MR-SPI requires that the GWAS summary statistics of the exposure and the outcome come from two non-overlapping samples, we exclude the trait pairs whose exposure and outcome are in the same consortium. In addition, we also exclude trait pairs whose exposure and outcome are two similar phenotypes (for example, heart failure and coronary artery disease), and we finally get 146 pairwise exposure-outcome combinations. All the GWAS summary statistics used for MR analysis are publicly available with more detailed description of each dataset given in Supplementary Table S7.

We first perform LD clumping using the software Plink (Purcell et al., 2007) to obtain independent SNPs with $r^2 < 0.01$, and then use 1×10^{-6} as the threshold to select relevant IVs that are associated with each exposure trait. Among the 146 exposure-outcome pairs, MR-SPI detects invalid IVs in 16 exposure-outcome pairs. For example, MR-SPI detects one invalid SNP (rs616154, marked by red triangle) in the causal relationship from cardioembolic stroke (CES) to SARS-CoV-2 infection, as illustrated in Figure 9(a). SNP rs616154 is identified to be invalid since its ratio estimate of the causal effect is 0.525, which is far away from other SNPs' ratio estimates and thus no other relevant SNP votes for it to be a valid IV. We search for the human phenotypes that are strongly associated with SNP rs616154 using the PhenoScanner tool (Staley et al., 2016; Kamat et al., 2019), and find that this SNP is also associated with the Interleukin-6 (IL-6) levels which is a potential biomarker of COVID-19 progression (Ulhaq and Soraya, 2020), indicating that SNP rs616154 might exhibit horizontal pleiotropy in the relationship of CES on SARS-CoV-2 infection and thus is a potential invalid IV. After excluding SNP rs616154, the point estimate of the causal effect by MR-SPI (represented by the slope of the green solid line in Figure 9(a)) is nearly zero, suggesting that cardioembolic stroke might not be a risk factor for SARS-CoV-2 infection. The causal effect estimate of MR-PRESSO (represented by the slope of the blue dashed line in Figure 9(a)) is also close to zero, as MR-PRESSO detects SNP rs616154 as an outlier and excludes it from analysis. However, IVW and MR-RAPS include SNP rs616154 in the MR analysis, and thus their causal effect estimates (represented by the slopes of the black and orange dashed line in Figure 9(a), respectively) might be biased. As a contrast, Figure 9(b) illustrates the causal effect estimates in the relationship of heart failure (HF) on any ischemic stroke (AIS) by MR-SPI, IVW, MR-PRESSO and MR-RAPS. In this relationship, MR-SPI does not identify any invalid IV, and thus MR-SPI produces a causal effect estimate that is similar to IVW and MR-RAPS.

Figure 9 illustrates that the inclusion of invalid IVs might lead to misleading scientific findings,

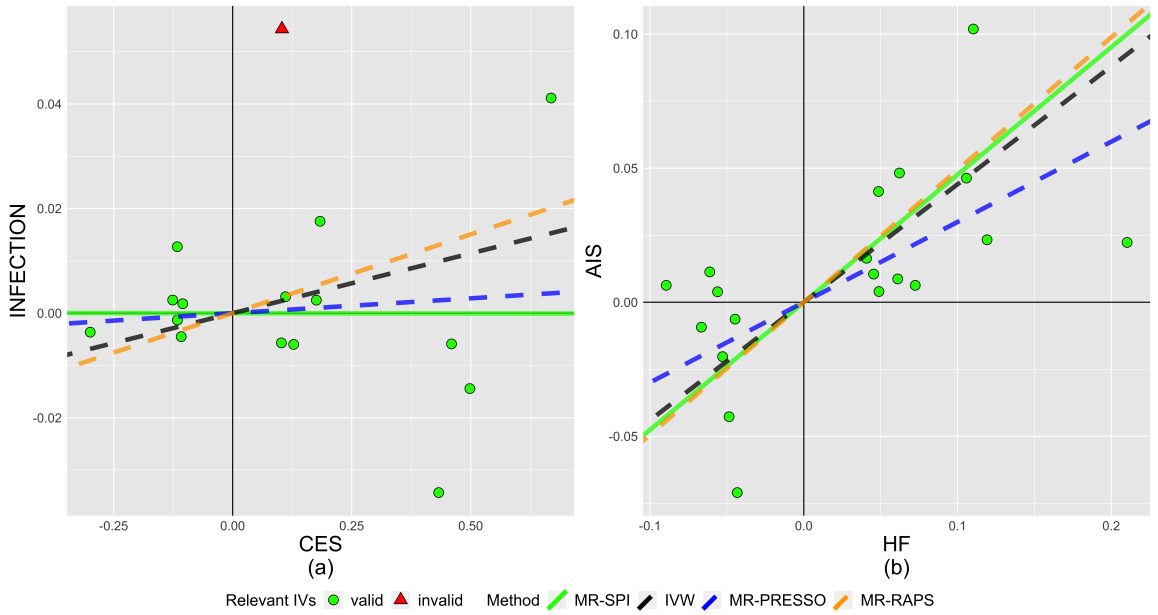

Figure 9: Scatter plot of (a) CES on AF, and (b) HF on AIS. The slope of the green solid line represents the causal effect estimate of MR-SPI. The slopes of the black, blue and orange dashed line represents the causal effect estimate of IVW, MR-PRESSO and MR-RAPS, respectively. Green circles represent the valid IVs and red triangles represent the invalid IVs detected by MR-SPI. CES: cardioembolic stroke; INFECTION: SARS-CoV-2 infection; HF: heart failure; AIS: any ischemic stroke.

and thus MR-SPI selects only valid IVs for downstream analysis to provide reliable causal inference. After excluding those invalid IVs, MR-SPI identifies 27 significant associations after Bonferroni correction for multiple comparison (Dunn, 1961), which are given in Figure 10. We also apply the other eight competing MR methods including IVW, MR-Egger, MR-RAPS, MR-PRESSO, the weighted median method, the mode-based estimation, MRMix and the contamination mixture method to infer the causal relationships among these exposure-outcome pairs, and the results are presented in Supplementary Figures S1-S8. Among the 146 exposure-outcome pairs, MR-SPI detects invalid IVs in 16 exposure-outcome pairs. Some of our findings are in line with previous studies, for example, an increase in LDL level might be associated with increased risk of CAD and HF (Trialists et al., 2012; Gencer et al., 2020). In addition, MR-SPI also detects significant associations that cannot be discovered by other competing MR methods. For example, MR-SPI suggests that SARS-CoV-2 infection might be a risk factor for HF ($\widehat{\beta} = 0.143, p\text{-value} = 1.435 \times 10^{-4}$), which cannot be identified by the other competing MR methods considered in this paper. Our finding is consistent with a former study that reported a significant increase in the risk of developing acute heart failure in patients with confirmed COVID-19 infection (Rey et al., 2020).

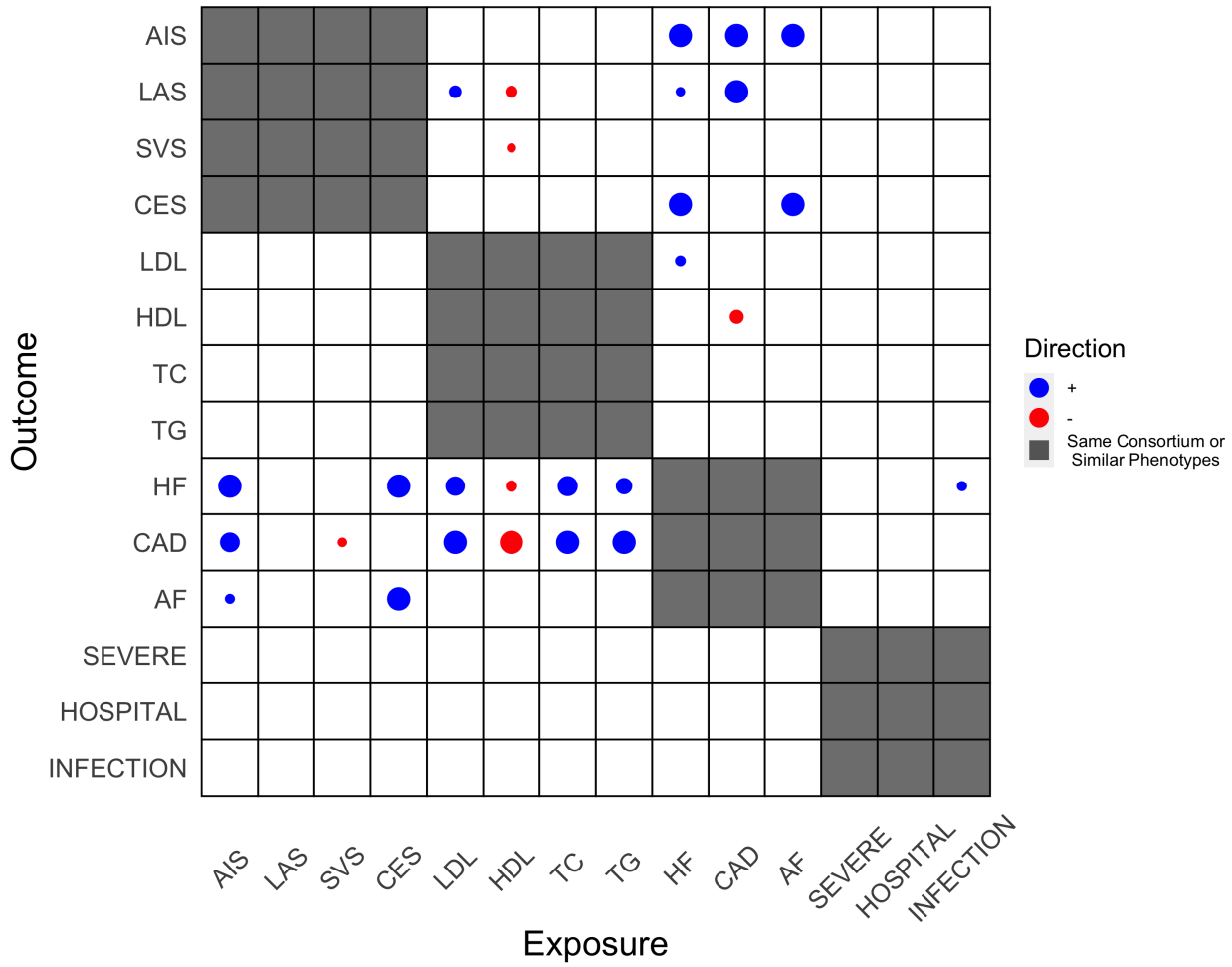

Figure 10: Point estimates of causal effects using MR-SPI. The significant positive and negative associations after Bonferroni correction are marked by blue filled circles and red filled circles, respectively. The radius of a circle is proportional to the $-\log_{10}(p\text{-value})$ of the corresponding exposure-outcome pair. Those pairs whose exposure and outcome come from the same consortium or are two similar phenotypes are marked as grey cells. AIS: any ischemic stroke; LAS: large-artery atherosclerotic stroke; SVS: small vessel stroke; CES: cardioembolic stroke; LDL: low-density lipoprotein; HDL: High-density lipoprotein; TC: total cholesterol; TG: triglycerides; HF: heart failure; CAD: coronary artery disease; AF: atrial fibrillation; SEVERE: severe COVID-19; HOSPITAL: COVID-19 hospitalization; INFECTION: SARS-CoV2 infection.

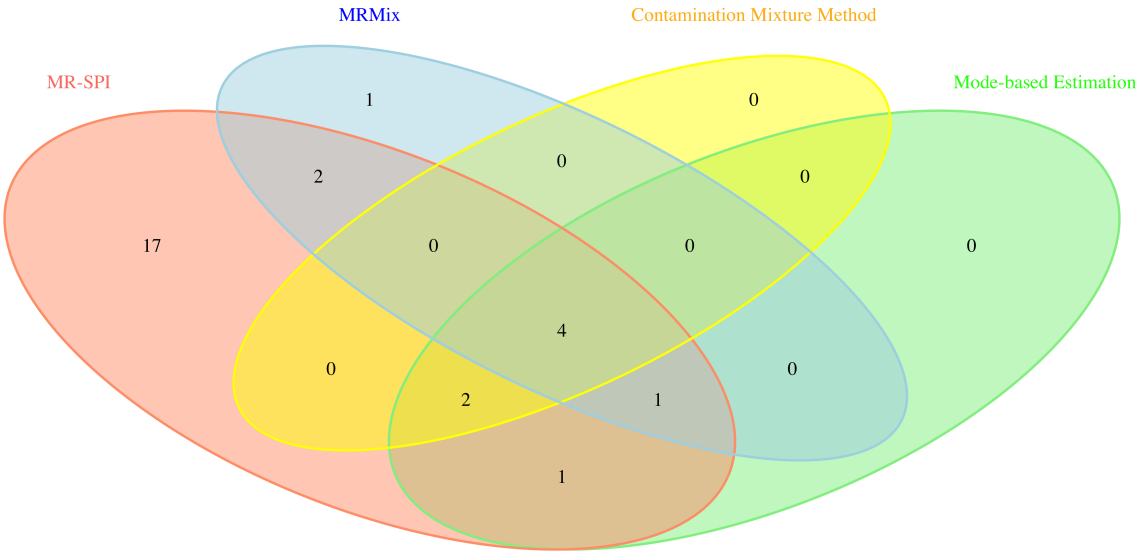

Figure 11: Venn diagram of significant associations detected by MR-SPI, the mode-based estimation, MRMix and the contamination mixture method after Bonferroni correction.

To demonstrate the similarities and differences in the results of MR-SPI and other MR methods, we plot the Venn diagrams that show the number of significant associations that are either shared or uniquely detected by these methods. We present the Venn diagram of the significant pairs using MR-SPI, the mode-based estimation, MRMix and the contamination mixture method in Figure 11, as these four methods are all based on the plurality rule condition. Venn diagrams that compare MR-SPI and the other competing MR methods can be found in Supplementary Figures S9 and S10. From Figure 11, MR-SPI generally detects more significant associations than the mode-based estimation, MRMix and the contamination mixture method among these 146 exposure-outcome pairs. Indeed, these three competing MR methods fail to discover some causal relationships that have been supported from previous literature. For example, the mode-based estimation, MRMix and the contamination mixture method fail to detect that an increased HDL level might be associated with a decreased risk of CAD, which is identified by MR-SPI ($\widehat{\beta} = -0.183, p\text{-value} = 3.730 \times 10^{-17}$) and has been supported with evidence by previous epidemiological studies (Krauss et al., 1987; Rader and Hovingh, 2014). Supplementary Figure S9 compares the significant relationships detected by MR-SPI and three MR methods that require InSIDE assumption (IVW, MR-Egger and MR-RAPS). MR-RAPS detects 17 significant associations that are not identified by MR-SPI, of which some associations might be spurious. For example, MR-RAPS suggests significant associations of AIS on low-density lipoprotein (LDL), high-density lipoprotein (HDL) and total

cholesterol (TC) level. However, the reverse association, i.e., cholestrol level on the risk of stroke, has been reported in previous epidemiological researches (Collaboration et al., 2007; Baigent et al., 2010). In Supplementary Figure S10, we compare MR-SPI with MR-PRESSO and the weighted median method that both require the majority rule condition. MR-PRESSO and the weighted median detect 14 and 10 significant associations, respectively, all of which are also identified by MR-SPI. Besides, MR-SPI identifies 11 more significant associations, most of which are in line with previous epidemiological studies, for example, HF might be a risk factor for ischemic stroke (Witt et al., 2006; Haeusler et al., 2011).

To deal with the issue of potential IV selection error, we also construct robust confidence intervals of these exposure-outcome pairs by MR-SPI according to Algorithm 2. Generally, MR-SPI discovers four significant associations whose robust confidence intervals do not include zero (CES on HF, LDL on CAD, TC on CAD, and atrial fibrillation (AF) on CES), and we compare the robust confidence intervals (represented by blue bars) with the default confidence intervals calculated by equation (11) (represented by red bars) in Figure 12. As shown in Figure 12, the robust confidence intervals are longer than the default confidence intervals of MR-SPI, indicating that locally invalid IVs might exist and might be incorrectly selected in these datasets. Therefore, we suggest using the robust confidence intervals for inference in these four relationships to provide more reliable causal findings.

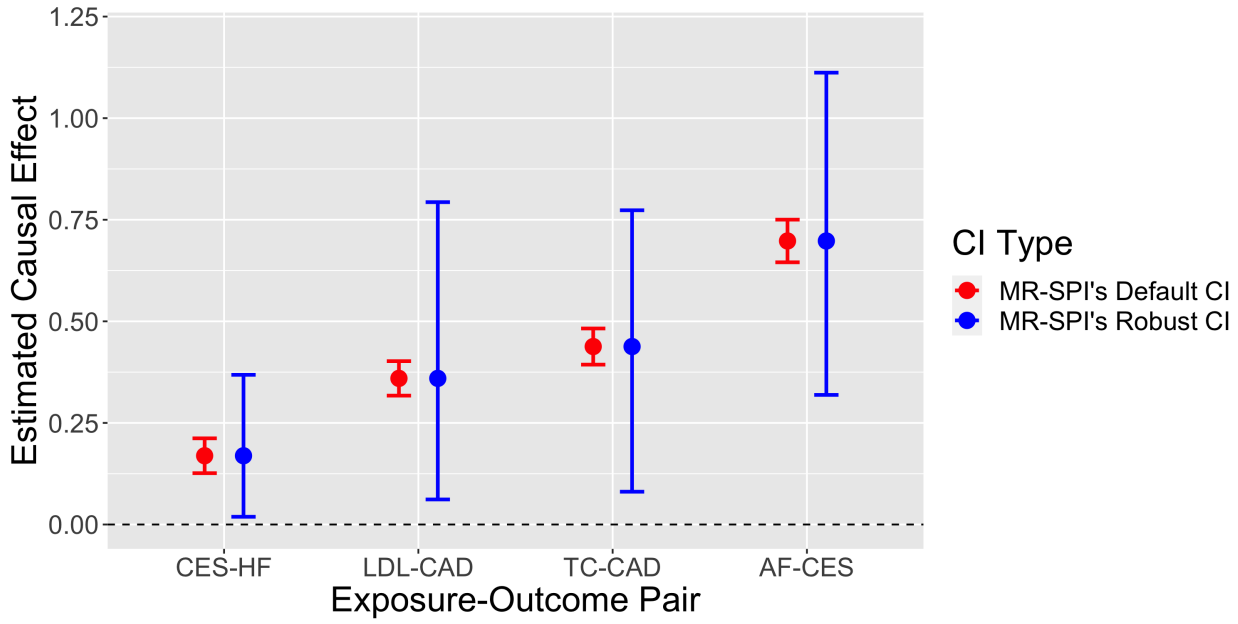

Figure 12: Significant associations detected by MR-SPI using the robust confidence interval. The red bars represent the default confidence interval calculated using the causal effect estimates and the corresponding standard errors by MR-SPI, and the blue bars represents the robust confidence interval constructed by the searching and sampling method.

5 Discussion

In this paper, we develop a novel two-sample MR method and algorithm, named MR-SPI, to automatically select valid SNPs from GWAS studies and perform post-selection inference. MR-SPI first selects relevant IVs with strong SNP-exposure associations, and then applies the voting procedure to select a plurality of the relevant IVs whose ratio estimates are similar to each other as valid IVs. In case that the causal effect estimate of MR-SPI is biased due to the selection of locally invalid IVs in finite samples, MR-SPI can provide a robust confidence interval constructed by the searching and sampling method, which is less vulnerable to IV selection error. Through extensive simulations and comprehensive real data analyses, we show that MR-SPI can be helpful to select valid instruments among genome-wide genetic variants for a specific exposure-outcome pair and provide reliable confidence interval for the causal effect even when many invalid instruments exist.

We emphasize two main advantages of MR-SPI in this paper. First, MR-SPI can incorporate both exposure and outcome data to automatically select a set of valid genetic instruments among genome-wide studies, and the selection procedure does not rely on additional distributional assumptions on the genetic effects. Therefore, MR-SPI is the first to offer such a practical approach to select valid instruments for a specific exposure-outcome pair from GWAS studies for MR analyses, which is especially helpful in the existence of wide-spread horizontal pleiotropy. Second, we propose a robust confidence interval for the causal effect using the searching and sampling method, which is less vulnerable to the IV selection error. Therefore, when locally invalid IVs are incorrectly selected and the causal effect estimate is biased in finite samples, we can still provide reliable inference for the causal effect with the robust confidence interval.

MR-SPI also has some limitations. Currently, MR-SPI can only perform causal inference using independent SNPs from two non-overlapping samples. As a future work, we plan to extend MR-SPI to include SNPs with LD structure from summary statistics of two possibly overlapping samples. Besides, the robust confidence interval is slightly more conservative than the confidence interval calculated from the limit distribution of the causal effect estimate, which is the price to pay for the robustness to the finite-sample IV selection error. Further studies are needed to construct less conservative confidence intervals that are robust to the IV selection error.

In conclusion, MR-SPI provides an automatic approach to select valid instruments among candidate SNPs and perform causal inference using two-sample GWAS summary statistics. Simulation studies and real data examples have shown that MR-SPI can provide reliable inference for the causal relationships even in the presence of invalid IVs. Our developed software is also user-friendly and highly efficient. We hope that MR-SPI can help researchers to detect more trustworthy causal mechanisms with increasingly

rich and publicly available GWAS datasets.

References

Baigent, C., Blackwell, L., Emberson, J., Holland, L., Reith, C., Bhala, N., Peto, R., Barnes, E., Keech, A., Simes, J., et al. (2010). Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials. *The Lancet*, 376(9753):1670–1681.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

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.

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.

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. and Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *European Journal of Epidemiology*, 32(5):377–389.

Cheng, Q., Zhang, X., Chen, L. S., and Liu, J. (2022). Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. *Nature Communications*, 13(1):1–13.

Collaboration, P. S. et al. (2007). Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55 000 vascular deaths. *The Lancet*, 370(9602):1829–1839.

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 Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42.

Davey Smith, G. and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98.

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.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Gencer, B., Marston, N. A., Im, K., Cannon, C. P., Sever, P., Keech, A., Braunwald, E., Giugliano, R. P., and Sabatine, M. S. (2020). Efficacy and safety of lowering LDL cholesterol in older patients: a systematic review and meta-analysis of randomised controlled trials. *The Lancet*, 396(10263):1637–1643.

Gratten, J. and Visscher, P. M. (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine*, 8(1):1–3.

Guo, Z. (2021). Post-selection problems for causal inference with invalid instruments: A solution using searching and sampling. *arXiv preprint arXiv:2104.06911*.

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.

Haeusler, K. G., Laufs, U., and Endres, M. (2011). Chronic heart failure and ischemic stroke. *Stroke*, 42(10):2977–2982.

Han, C. (2008). Detecting invalid instruments using L1-GMM. *Economics Letters*, 101(3):285–287.

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.

Hemani, G., Bowden, J., and Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics*, 27(R2):R195–R208.

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.

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.

Krauss, R., Williams, P., Brensike, J., Detre, K., Lindgren, F., Kelsey, S., Vranizan, K., and Levy, R. (1987). Intermediate-density lipoproteins and progression of coronary artery disease in hypercholesterolaemic men. *The Lancet*, 330(8550):62–66.

Lawlor, D. A. (2016). Commentary: Two-sample Mendelian randomization: opportunities and challenges. *International Journal of Epidemiology*, 45(3):908.

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.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.

Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.

Ouyang, Q., Kaplan, P. D., Liu, S., and Libchaber, A. (1997). DNA solution of the maximal clique problem. *Science*, 278(5337):446–449.

Parkes, M., Cortes, A., Van Heel, D. A., and Brown, M. A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics*, 14(9):661–673.

Pierce, B. L. and Burgess, S. (2013). Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184.

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.

Rader, D. J. and Hovingh, G. K. (2014). HDL and cardiovascular disease. *The Lancet*, 384(9943):618–625.

Rey, J. R., Caro-Codón, J., Rosillo, S. O., Iniesta, Á. M., Castrejón-Castrejón, S., Marco-Clement, I., Martín-Polo, L., Merino-Argos, C., Rodríguez-Sotelo, L., García-Veas, J. M., et al. (2020). Heart failure in COVID-19 patients: prevalence, incidence and prognostic implications. *European Journal of Heart Failure*, 22(12):2205–2215.

Rothman, K. J. and Greenland, S. (2005). Causation and causal inference in epidemiology. *American Journal of Public Health*, 95(S1):S144–S150.

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.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, 43(4):333–338.

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.

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.

Staley, J. R., Blackshaw, J., Kamat, M. A., Ellis, S., Surendran, P., Sun, B. B., Paul, D. S., Freitag, D., Burgess, S., Danesh, J., et al. (2016). PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics*, 32(20):3207–3209.

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.

The Coronary Artery Disease (C4D) Genetics Consortium (2011). A genome-wide association study in europeans and south asians identifies five new loci for coronary artery disease. *Nature Genetics*, 43(4):339–344.

Trialists, C. T. et al. (2012). The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *The Lancet*, 380(9841):581–590.

Ulhaq, Z. S. and Soraya, G. V. (2020). Interleukin-6 as a potential biomarker of covid-19 progression. *Medecine et maladies infectieuses*, 50(4):382.

Vandenbroucke, J. P., Broadbent, A., and Pearce, N. (2016). Causality and causal inference in epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45(6):1776–1786.

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.

Witt, B. J., Brown Jr, R. D., Jacobsen, S. J., Weston, S. A., Ballman, K. V., Meverden, R. A., and Roger, V. L. (2006). Ischemic stroke after heart failure: a community-based study. *American heart journal*, 152(1):102–109.

Xue, H., Shen, X., and Pan, W. (2021). Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal of Human Genetics*, 108(7):1251–1269.

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.