Robust Mendelian Randomization Analysis by Automatically Selecting Valid Genetic Instruments for Inferring Causal Relationships and Identifying Omics Biomarkers

Minhao Yao[1], Gary W. Miller[2], Badri N. Vardarajan[3], Andrea A. Baccarelli[2], Zijian Guo[4*], Zhonghua Liu[5*]

[1] *Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong SAR, China.*
[2] *Department of Environmental Health Sciences, Columbia University, New York, NY, USA.*
[3] *Taub Institute on Alzheimer's Disease and the Aging Brain, Department of Neurology, Columbia University, New York, NY, USA.*
[4] *Department of Statistics, Rutgers University, Piscataway, NJ, USA.*
[5] *Department of Biostatistics, Columbia University, New York, NY, USA.*

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

Abstract

Mendelian randomization (MR) uses genetic variants as instrumental variables (IVs) to estimate the causal effect of a modifiable exposure on the outcome of interest to remove unmeasured confounding bias. However, some genetic variants might be invalid IVs due to violations of core IV assumptions, for example, in the presence of population stratification or/and widespread horizontal pleiotropy. Inclusion of invalid genetic IVs for MR analysis might lead to biased causal effect estimate and misleading scientific conclusions. To address this challenge, we propose a novel MR method that first Selects valid genetic IVs and then performs Post-selection Inference (MR-SPI) based on two-sample genome-wide summary statistics. Extensive Monte Carlo simulation studies demonstrate the superior performance of MR-SPI. We apply MR-SPI to analyze 146 exposure-outcome pairs to establish putative causal relationships. We further analyze 912 plasma proteins using the UK Biobank proteomics data in 54,306 UK Biobank participants and identify 7 proteins significantly associated with the risk of Alzheimer's disease.

1 Introduction

In epidemiological studies, it is essential to infer the causal effect of a modifiable risk factor on a health outcome of interest[1,2]. Even though randomized controlled trials (RCTs) serve as the gold standard for causal inference, it is often neither feasible nor ethical to perform RCTs for many harmful exposures. Mendelian randomization (MR) leverages the random assortment of genes from parents to offspring to mimic RCTs to establish causality in observational studies[3,4,5]. MR uses genetic variants, typically single-nucleotide polymorphisms (SNPs), as instrumental variables (IVs) to assess the causal association between an exposure and an outcome[6]. Recently, many MR methods have been developed to investigate causal relationships using genome-wide association study (GWAS) summary statistics 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[7,8,9,10]. 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[9,11,12,13].

Conventional MR methods require the genetic variants included in the analysis to be valid IVs for reliable causal inference. A genetic variant is called a valid IV if the following three core assumptions hold[4,14]:

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

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

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

Among the three core IV assumptions (A1)-(A3), only 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 may lead to a biased estimate of the causal effect. Violation of (A2) may occur due to the presence of population stratification[4,15]. Violation of (A3) may occur in the presence of the horizontal pleiotropy[4,16], which is a widespread biological phenomenon that the genetic variant affects the outcome through other biological pathways that do not involve the exposure in view[17,18].

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 adopted by multiple methods, for example, the random-effects inverse-variance weighted (IVW) method[19], MR-Egger[20] and MR-RAPS (Robust Adjusted Profile Score)[11]. The InSIDE assumption requires that the SNP-exposure effect is asymptotically independent of the horizontal pleiotropic effect when the number of SNPs goes to infinity. However, the InSIDE assumption is often implausible in practice[21], and thus the estimate of causal effect might be biased using random-effects IVW, MR-Egger or MR-RAPS[10]. Another strand of methods imposes assumptions on the proportion of invalid IVs included in the analysis. For example, the weighted median method[22] and the Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO) test[23] are based on the majority rule condition that 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[10]. Besides, the MR-PRESSO outlier test requires that the InSIDE assumption holds and that the pleiotropic effects of genetic instruments have zero mean[23]. The plurality rule condition, which only requires a plurality of the candidate IVs to be valid, is weaker than the majority rule condition[24,25], and is also termed as the ZEro Modal Pleiotropy Assumption (ZEMPA)[10,26]. The plurality rule condition (or ZEMPA assumption) has been applied to some existing two-sample MR methods, for example, the mode-based estimation[26], MRMix[27] and the contamination mixture method[25]. Among the aforementioned 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 GWAS[28]. 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 because the genetic architecture may vary across outcomes; 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 a specific exposure-outcome pair.

In this paper, we propose a novel two-sample MR method and algorithm that can automatically

Select valid IVs for a 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 assumptions (A2) and (A3) are smaller than a threshold as in equation (4) (thus more likely to be valid) under this ratio estimate of the causal effect; (iii) select valid IVs that receive a majority/plurality of votes, or by finding the maximum clique of the voting matrix that encodes whether two relevant IVs mutually vote for each other; to be valid IVs and (iv) perform post-selection inference to construct a confidence interval for the causal effect that is 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 automatically select a valid set of exposure-outcome pair specific SNP IVs. Moreover, our proposed selection procedure does not require additional distributional assumptions, for example, normal mixture distributions, to model the SNP-trait associations or ratio estimates[25,27], and thus is more robust to possible violations of parametric distributional assumptions. Extensive simulations show that our MR-SPI method outperforms other competing MR methods under the plurality rule condition. We apply MR-SPI to infer the causal relationships among 146 exposure-outcome pairs involving COVID-19 related traits, ischemic stroke, cholesterol levels and heart disease, and detect significant associations among them. Furthermore, we employ MR-SPI to perform omics MR (xMR) with 912 plasma proteins using UK Biobank proteomics data in 54,306 UK Biobank participants and discover 7 proteins significantly associated with the risk of Alzheimer's disease.

2 Results

2.1 MR-SPI selects valid genetic instruments by voting procedure

MR-SPI is an automatic procedure to select valid genetic instruments and perform robust causal inference using two-sample GWAS summary data. In brief, MR-SPI contains the following four steps, as illustrated in Figure 1:

(i) select relevant SNPs strongly associated with the exposure in the GWAS summary data;

(ii) each relevant SNP provides a ratio estimate of the causal effect, and all the other relevant SNPs votes for it to be a valid IV if their degrees of violation of assumptions (A2) and (A3) are smaller than a threshold as in equation (4) under this ratio estimate of the causal effect;

(iii) select valid IVs by majority/plurality voting or by finding the maximum clique of the voting matrix that encodes whether two relevant IVs mutually vote for each other to be valid;

(iv) estimate the causal effect of interest using the selected valid IVs and construct a confidence interval for the causal effect that is robust to IV selection error in finite samples.

Current two-sample MR methods only use step (i) to select (relevant) genetic instruments for downstream MR analysis, while the selected genetic instruments might violate assumptions (A2) and (A3), leading to possibly unreliable scientific findings. To address this issue, MR-SPI automatically select valid genetic instruments for a specific exposure-outcome pair by further incorporating the outcome data. Our key idea of selecting valid genetic instruments is that, under the plurality rule condition, valid IVs will form the largest group of relevant IVs and give "similar" ratio estimates (see Online Methods). Specifically, we propose the following two criteria to measure the similarity between the ratio estimates of two SNPs j and k in step (ii):

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 assumptions (A2) and (A3) is smaller than a data-dependent threshold as in equation (4);

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

In step (iii), we construct a symmetric and binary voting matrix to encode the votes that each relevant SNP receives from other relevant SNPs: the (k, j) entry of the voting matrix is 1 if SNPs j and k mutually vote for each other to be valid, and 0 otherwise. There are two ways to select valid genetic instruments based on the voting matrix (see Online Methods): (1) we can select relevant SNPs who receive majority voting or plurality voting as valid IVs; (2) we can use SNPs in the maximum clique of the voting matrix as valid IVs[29]. Our simulation studies show that the maximum clique method can empirically offer lower false discovery rate (FDR)[30] and higher true positive proportion (TPP) as shown in Table 1.

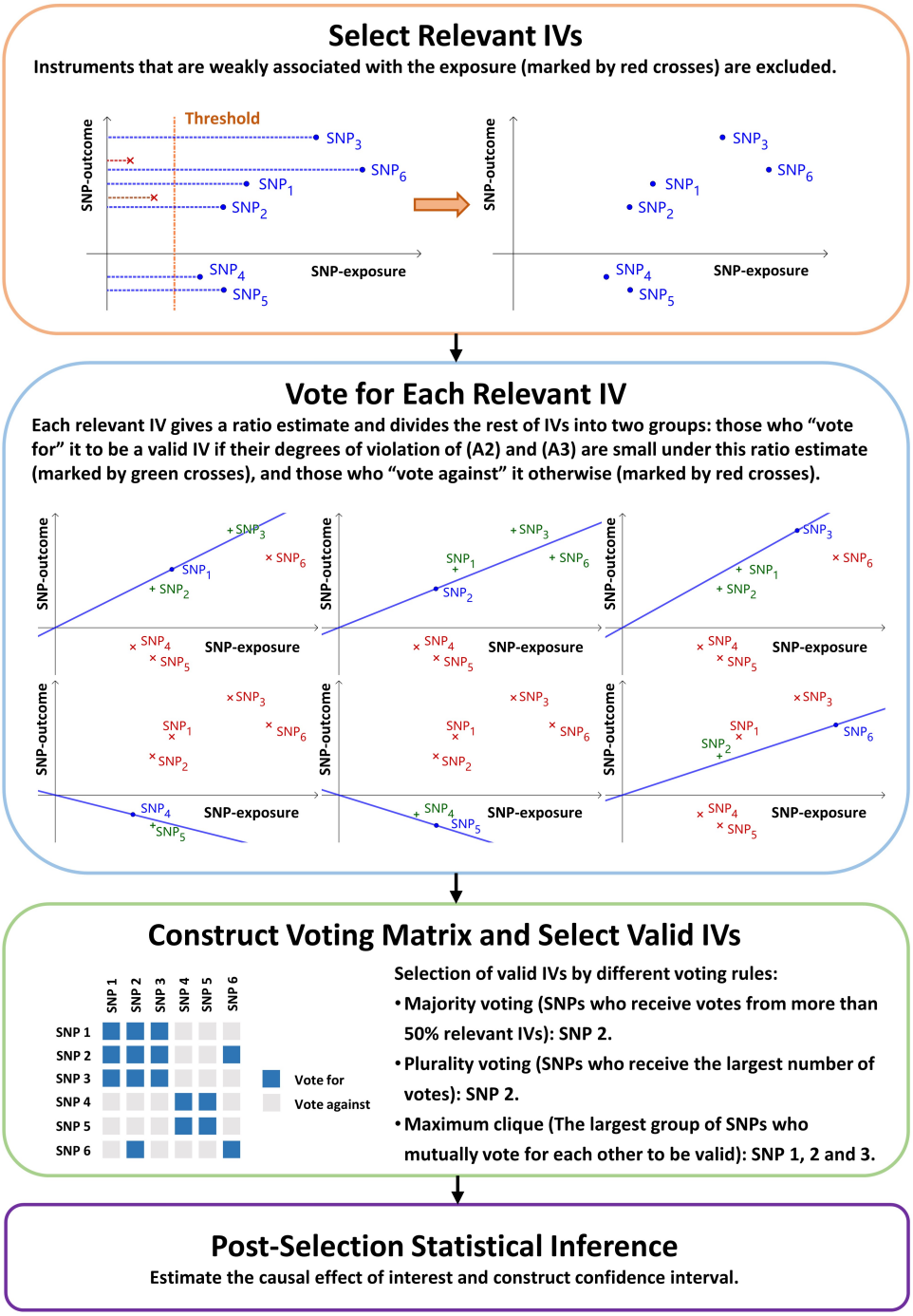


Figure 1: The framework of MR-SPI. In the first step, MR-SPI selects the relevant IVs with strong SNP-exposure associations. In the second step, each relevant IV provides a ratio estimate of the causal effect and receives votes on itself to be valid from the other relevant IVs whose degrees of violation of (A2) and (A3) are small under this ratio estimate. 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 SNPs 2 and 3 vote for SNP 1 to be valid because they are close to that line, while SNPs 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 majority voting, plurality voting or by finding the maximum clique. In the inference step, MR-SPI estimates the causal effect and constructs the confidence interval using the selected valid IVs.

In step (iv), we estimate the causal effect and construct a confidence interval for this causal effect using the selected valid genetic instruments. In finite samples, some invalid IVs with small (but still nonzero) degrees of violation of assumptions (A2) and (A3) might be incorrectly selected as valid IVs in finite-sample settings, and we refer to them as "locally invalid IVs" [31]. We then propose to construct a robust confidence interval with guaranteed nominal coverage even in the presence of IV selection error in finite-sample settings, as described in Figure 6 and Online Methods.

2.2 Comparing MR-SPI to other competing methods in simulation studies

We conduct extensive simulations to evaluate the performance of MR-SPI in the presence of invalid IVs. We simulate data in a two-sample setting under four setups: (**S1**) majority rule condition holds, and no locally invalid IVs exist; (**S2**) plurality rule condition holds, and no locally invalid IVs exist; (**S3**) majority rule condition holds, and locally invalid IVs exist; (**S4**) plurality rule condition holds, and locally invalid IVs exist. More detailed simulation settings are described in Online Methods.

We compare the percent bias, empirical coverage and average lengths of 95% confidence intervals of MR-SPI to the following competing methods: (i) the random-effects IVW method that performs random-effects meta-analysis to account for pleiotropy [19], (ii) MR-RAPS that assumes pleiotropic effects are normally distributed and applies the maximum profile likelihood estimation to obtain the causal effect estimate [11], (iii) MR-PRESSO that detects the SNPs that substantially reduce the residual sum of squares of the regression when omitted from the analysis as outliers [23], (iv) the weighted median method that takes the weighted median of the ratio estimates as the causal effect [22], (v) the mode-based estimation that takes the mode of the smoothed empirical density function of the ratio estimates as the causal effect [26], (vi) MRMix that models the SNP-exposure and SNP-outcome effects with a bivariate normal mixture distribution [27], and (vii) the contamination mixture method that models the ratio estimates of SNPs with a normal mixture distribution [25]. We exclude MR-Egger in this simulation since it is heavily biased under our simulation settings. Among those 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 the ZEMPA assumption). For simplicity, we shall use IVW to represent the random-effects IVW method here and after.

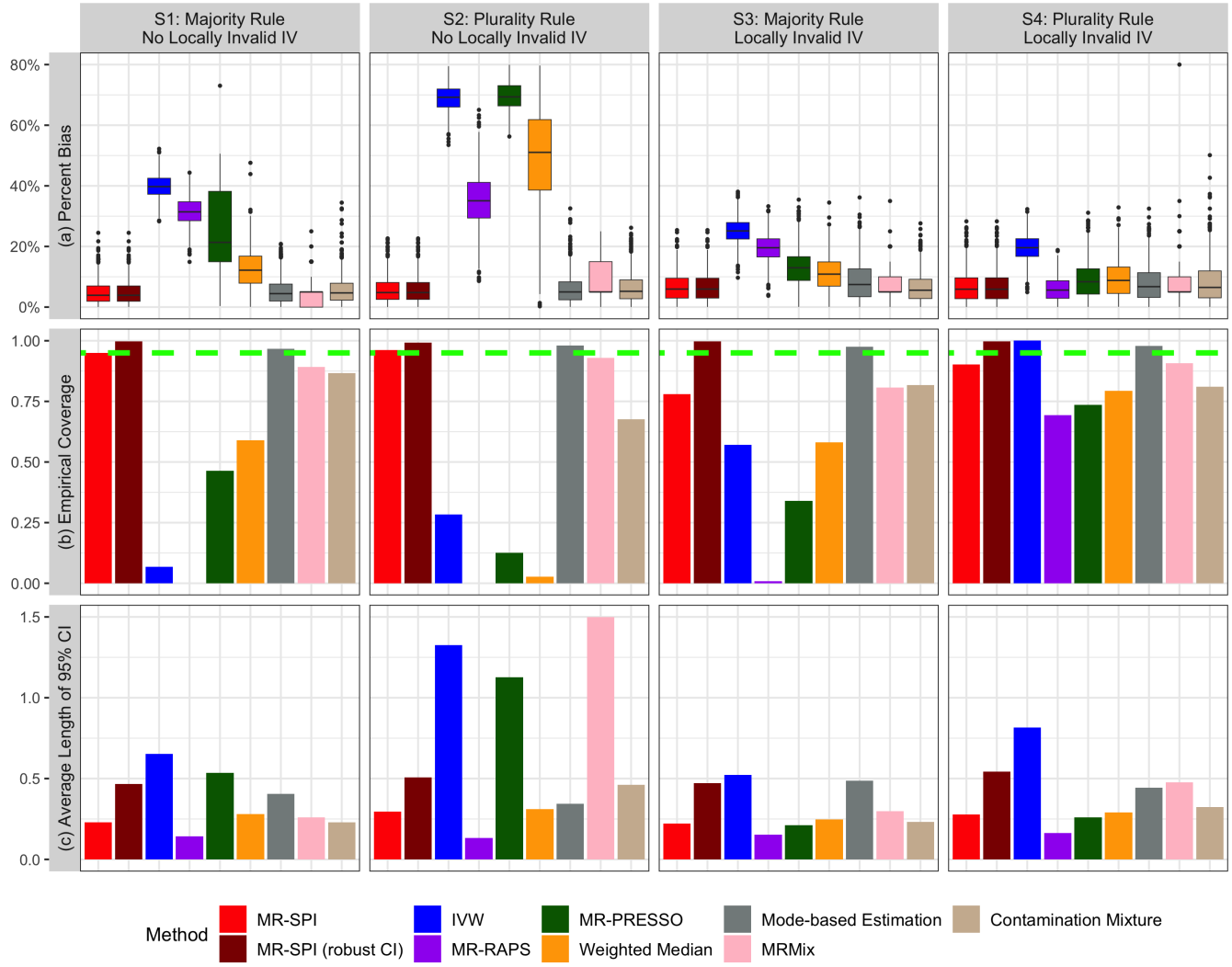


Figure 2: Performance of MR-SPI and the other competing MR methods in simulated data with sample sizes of 5000. (a) Boxplot of the percent bias in causal effect estimates. (b) Empirical coverage of 95% confidence intervals. The green dashed line in (b) represents the nominal level (95%). (c) Average lengths of 95% confidence intervals.

In Figure 2(a), we present the percent bias of those MR methods in simulated data with sample sizes of 5000 for both the exposure and the outcome. Moreover, Supplementary Figure S1(a) and Supplementary Table S1 provide a comparison of percent bias of those MR methods across different sample sizes ($n = 5000, 10000, 2000, 40000, 80000$). Generally, the proposed MR-SPI has small bias in all four settings. IVW and MR-RAPS are biased since the InSIDE assumption does not hold in our simulation settings. When the InSIDE assumption holds and the pleiotropic effects have zero mean, both IVW and MR-RAPS can give a nearly unbiased estimate of the causal effect [11,19]. Besides, the biases of these two methods are smaller in settings (**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 yields biased estimates as it fails to remove outliers in most of our settings. As discussed in Verbanck et al.[23], MR-PRESSO performs well when (1) both the majority rule condition and the InSIDE assumption hold, and (2) the pleiotropic effects have zero mean. The weighted median estimator is biased when only the plurality rule condition holds, since it requires more than half of the candidate IVs to be valid. The mode-based estimation, MRMix and the contamination mixture method are all nearly unbiased, as these three methods only require the plurality rule condition to hold, which is satisfied in all the four simulation settings.

Figure 2(b) reports the empirical coverage of the confidence intervals of those methods in simulated data with sample sizes of 5000. Additional results for empirical coverage of those methods under different sample sizes ($n = 5000, 10000, 2000, 40000, 80000$) can be found in Supplementary Figure S1(b) and Supplementary Table S2. Under settings (**S1**) and (**S2**) where locally invalid IV does not exist, the confidence interval of MR-SPI can attain 95% coverage level even when the sample sizes are small (e.g., 5000). In the presence of locally invalid IVs, i.e., under settings (**S3**) and (**S4**), the empirical coverage of the confidence interval of MR-SPI can still attain the nominal level when sample sizes are 80000, as MR-SPI can correctly distinguish locally invalid IVs from 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 (e.g., 5000), and therefore the empirical coverage of MR-SPI is lower than 95%. In such cases, our proposed robust confidence interval constructed by Algorithm 2 in Online Methods, can attain 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 setting (**S1**) where the majority rule condition holds. For example, when the sample sizes are 20000, the empirical coverage of the weighted median method is 0.638. Compared to the confidence interval of MR-SPI, the confidence interval of 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[26]. Both MRMix and the contamination mixture method cannot attain the 95% coverage level in all the four simulation settings. 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. However, when the underlying distributional assumption is satisfied, both MRMix and the contamination mixture

method can attain the nominal level under the plurality rule condition, as shown in additional simulations in Supplementary Section S7.

We report the average lengths of 95% confidence intervals of MR-SPI and the other competing methods under sample sizes = 5000 in Figure 2(c), and additional results for various sample sizes are provided in Supplementary Figure S1(c) and Supplementary Table S3. Although MR-RAPS generally has the shortest confidence interval, it is biased and the coverage level is close to zero, because the InSIDE assumption does not hold in our simulation settings. Among the methods except MR-RAPS, MR-SPI generally has the shortest confidence interval under all the four simulation settings. The average length of confidence interval of IVW is not 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 there exists heterogeneity in the ratio estimates[19]. In setting (**S4**), MR-PRESSO has longer confidence interval 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.

In Table 1, we report (1) the FDR that is defined by the proportion of invalid IVs in the set of SNPs selected by MR-SPI, and (2) the TPP that is defined by the proportion of valid IVs selected by MR-SPI in the true set of valid IVs. In our simulation, we select valid IVs by finding the maximum clique in the voting matrix. The TPP of MR-SPI is close to 1 under all settings, and the FDR of MR-SPI is close to 0 if locally invalid IV does not exist and the plurality rule condition holds. Under settings (**S3**) and (**S4**), MR-SPI might incorrectly select those locally invalid IVs when the sample sizes are small (e.g., 5000). As the sample sizes increase, the FDR of MR-SPI would be close to 0 even in the presence of locally invalid IVs. For example, the FDR is 0.005 under setting (**S4**) when the sample size is 80000. Therefore, even when locally invalid IV exists, MR-SPI can still correctly identify valid IVs if the sample sizes are large enough.

Table 1: The FDR and TPP of valid IV selection by MR-SPI under different settings 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 settings.

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 better compared to the other competing MR methods under the plurality rule condition. When locally invalid IV does not exist, MR-SPI can select 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. In practice, we can perform a sensitivity analysis of the causal effect estimate to the threshold in the voting step (see Online Methods and Supplementary Figure S12). If the causal effect estimate is sensitive to the choice of the threshold, then MR-SPI might suffer from the finite-sample IV selection error, and thus the robust confidence interval of MR-SPI is recommended for use in this case.

2.3 Evaluation of the performance of MR-SPI using two benchmark datasets

In this section, we apply the proposed MR-SPI method to two benchmark datasets to evaluate its performance. These two 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. We first apply MR-SPI to the dataset in which 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[32], and the outcome data come from the Coronary ARtery DIsease Genome-wide Replication and Meta-analysis (CARDIoGRAM) consortium[33]. We first perform linkage disequilibrium (LD)-based clumping to SNPs in the exposure data using the software Plink[34] with $r^2 < 0.01$ to obtain independent genetic instruments, and then use 1×10^{-6} as the p-value threshold to select relevant instruments. In total, five relevant instruments are included for downstream analysis. We compare MR-SPI to the other eight com-

peting MR methods including IVW, MR-Egger, MR-RAPS, MR-PRESSO, the weighted median method, the mode-based estimation, MRMix and the contamination mixture method. The causal effect estimates and the corresponding 95% confidence intervals using those methods are presented in Figure 3(a). Generally, the confidence intervals of MR-SPI, IVW and MR-Egger all cover 1, and MR-SPI provides the shortest confidence interval. In addition, none of the relevant IVs is excluded in the voting step, which is in line with the expectation that horizontal pleiotropy should not exist in this dataset.

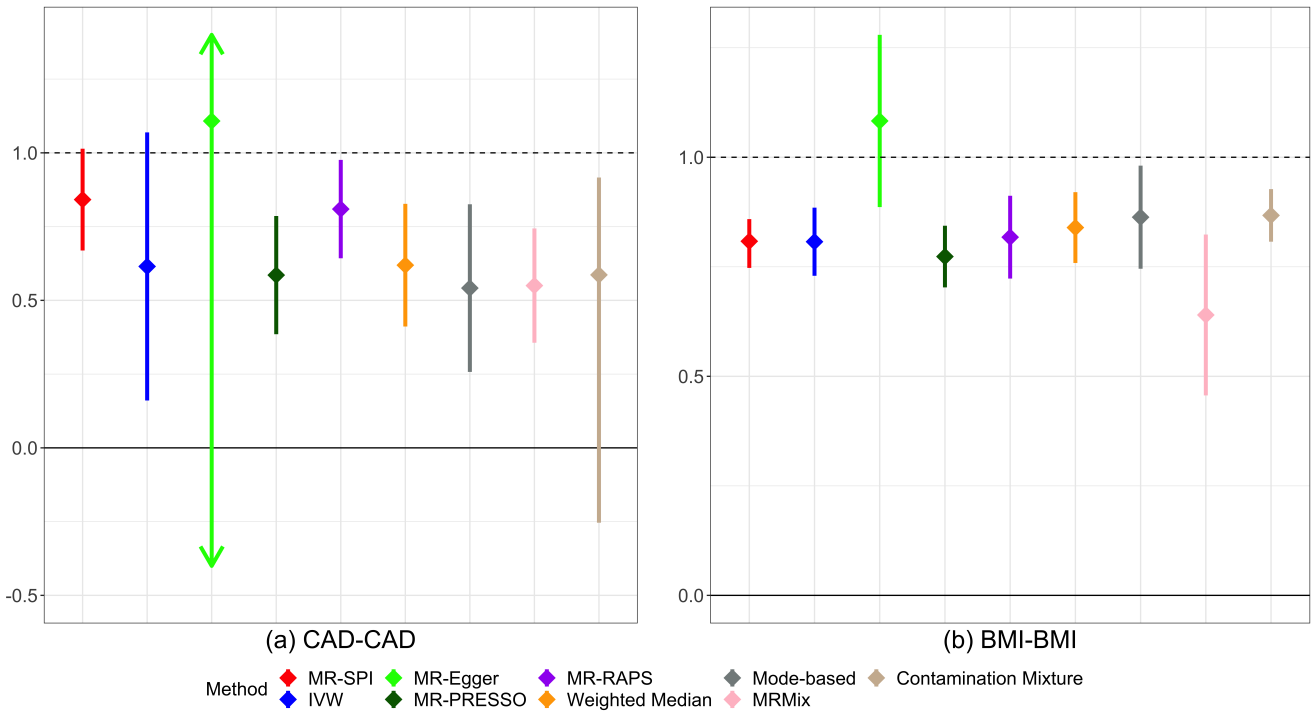


Figure 3: Point estimates and 95% confidence intervals for the causal effects of (a) CAD-CAD dataset and (b) BMI-BMI dataset using different MR methods. Confidence intervals are clipped to arrows if they exceed axis limits. CAD: coronary artery disease; BMI: body mass index.

Next, we apply MR-SPI to the dataset where the exposure data and the outcome data are the body mass index (BMI) GWAS data for physically active men and women respectively, and we refer to this dataset as the BMI-BMI dataset. In the BMI-BMI dataset, both the exposure data and the outcome data come from the GIANT consortium[35]. 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 MR-SPI and the other competing methods are shown in Figure 3(b). Overall, all the above methods except MR-Egger provide causal

effect estimates that are below one. As discussed in previous studies[35], 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 difference in the genetic architecture of BMI between men and women.

2.4 Learning causal relationships of 146 exposure-outcome pairs

In this section, we examine the causal relationships between complex traits and diseases 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 obtain 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 S4.

We first perform LD clumping using the software Plink[34] to obtain independent SNPs with $r^2 < 0.01$, and then use 1×10^{-6} as the p-value 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 the left panel of Figure 4(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[36,37], and find that this SNP is also associated with the Interleukin-6 (IL-6) levels which is a potential biomarker of COVID-19 progression[38], indicating that SNP rs616154 might exhibit horizontal pleiotropy in the relationship of cardioembolic stroke on SARS-CoV-2 infection and thus is a potentially 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 the left panel of Figure 4(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 the left panel of

Figure 4(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 the left panel of Figure 4(a), respectively) might be biased. In contrast, the right panel of Figure 4(a) illustrates the causal effect estimates for 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 gives a causal effect estimate that is similar to IVW and MR-RAPS.

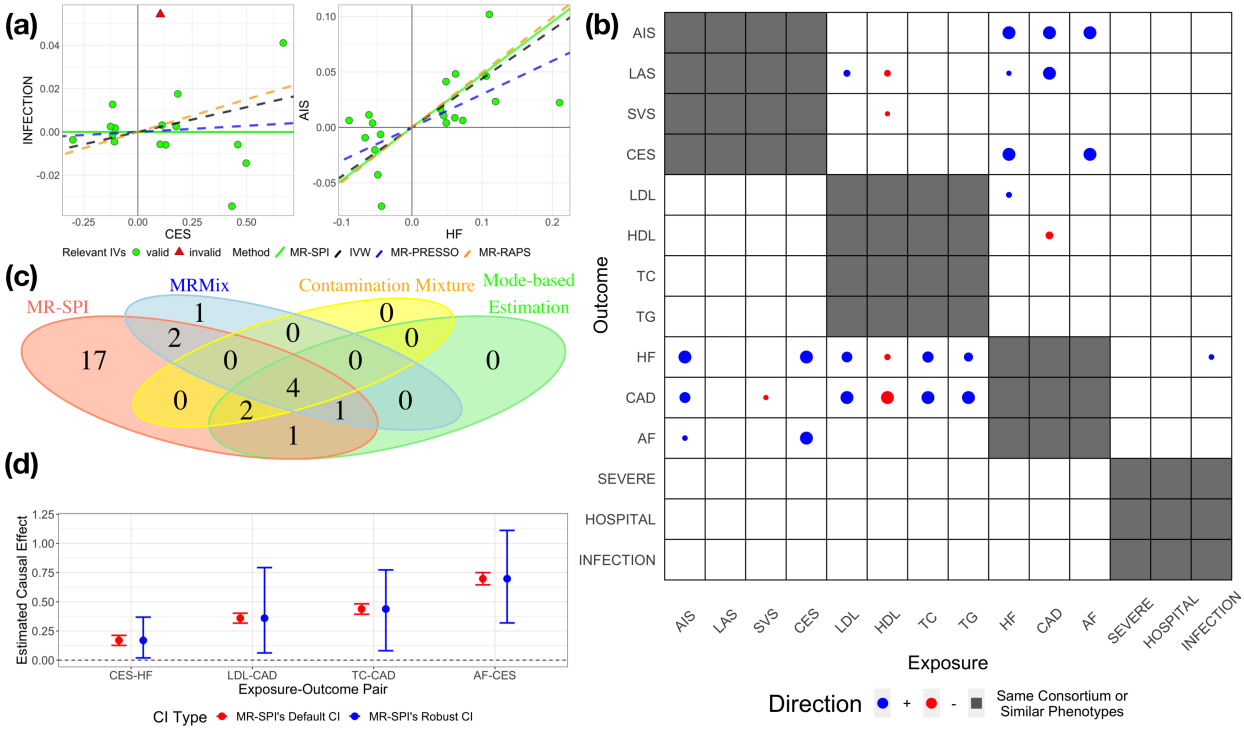


Figure 4: **(a)** Scatter plot of cardioembolic stroke on SARS-CoV-2 infection (left panel), and heart failure on any ischemic stroke (right panel). The slope of the green solid line represents the causal effect estimate of MR-SPI. The slopes of the black, blue and orange dashed line represent the causal effect estimates of IVW, MR-PRESSO and MR-RAPS, respectively. Green circles represent valid IVs and red triangles represent invalid IVs detected by MR-SPI. **(b)** Direction of causal associations detected by 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. **(c)** Venn diagram of significant associations detected by MR-SPI, the mode-based estimation, MRMix and the contamination mixture method after Bonferroni correction. **(d)** 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 represent the robust confidence interval constructed by the searching and sampling method. 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-CoV-2 infection.

Figure 4(a) illustrates that the inclusion of invalid IVs might lead to misleading scientific findings, 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[39], which are summarized in Figure 4(b). 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 S2-S9. 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 risks of CAD and HF[40,41]. 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 ($\widehat{\beta} = 0.14, p\text{-value} = 1.43 \times 10^{-4}$), 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[42].

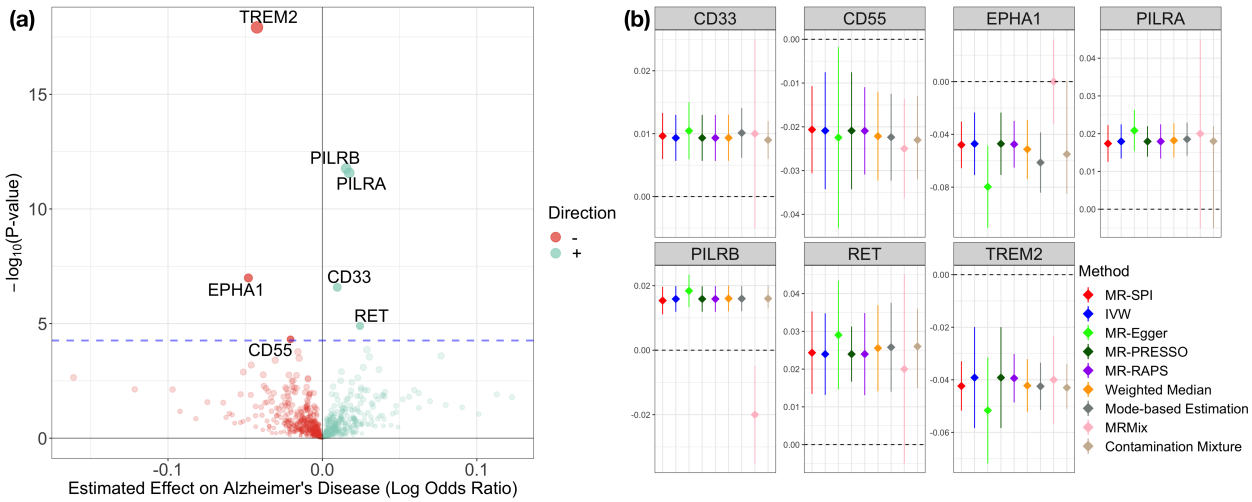
To demonstrate the similarities and differences in the results of MR-SPI and other MR methods, we plot the Venn diagrams to show the number of significant associations that are either shared or uniquely detected by these methods. We present the Venn diagram of the significant pairs using MR-SPI, the mode-based estimation, MRMix and the contamination mixture method in Figure 4(c), 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 S10 and S11. From Figure 4(c), MR-SPI 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 ($\widehat{\beta} = -0.18, p\text{-value} = 3.73 \times 10^{-17}$) and has been supported with evidence by previous epidemiological studies[43,44]. Supplementary Figure S10 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., cholesterol level on the risk of stroke, has been reported in previous epidemiological studies[40,45]. In Supplementary Figure S11, we compare MR-SPI with MR-PRESSO and the weighted median

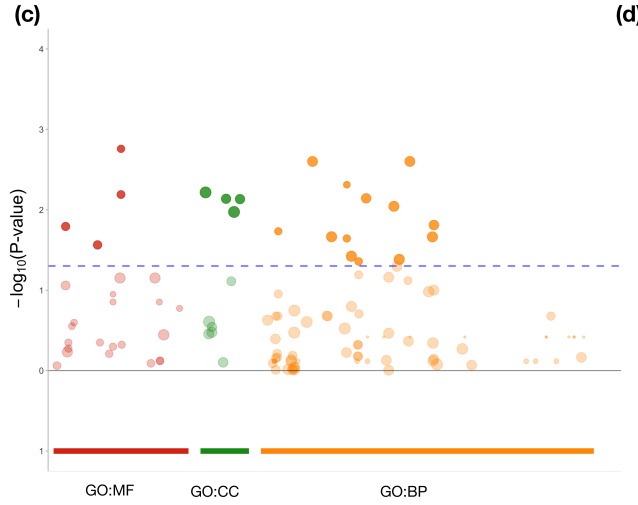
method that both require the majority rule condition. MR-PRESSO and the weighted median detect 14 and 10 significant associations, respectively, all of which are also identified by MR-SPI. Besides, MR-SPI identifies 11 more significant associations, most of which are in line with previous epidemiological studies, for example, HF might be a risk factor for ischemic stroke[46,47].

To deal with the issue of potential finite-sample IV selection error, we also construct robust confidence intervals of these exposure-outcome pairs by MR-SPI according to Algorithm 2 in Online Methods. 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 (8) in Online Methods (represented by red bars) in Figure 4(d). As shown in Figure 4(d), 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 as valid IVs in these datasets. Therefore, we suggest using the robust confidence intervals for these four relationships to provide more reliable causal findings.

2.5 Identifying proteins associated with Alzheimer's disease using MR-SPI

Omics MR (xMR) aims to identify omics biomarkers (e.g., proteins) causally associated with complex traits and diseases. In particular, xMR with proteomics data enables the identification of disease-associated proteins, facilitating crucial advancements in novel drug target discovery or drug repurposing, targeted prevention, and better treatment strategies. In this section, we apply MR-SPI to identify protein biomarkers putatively causally associated with the risk of Alzheimer's disease (AD). The proteomics data used in our analysis comprises 54,306 participants from the UK Biobank Pharma Proteomics Project (UKB-PPP)[48]. Following the guidelines proposed by Sun et al.[48], significant (p-value $< 3.40 \times 10^{-11}$, accounting for Bonferroni correction) and independent ($r^2 < 0.01$) SNPs are extracted from the proteomics data as candidate genetic instruments, and thus all of these candidates SNPs are strongly associated with the exposures (proteins). Summary statistics for AD are obtained from a meta-analysis of GWAS studies for clinically diagnosed AD and AD-by-proxy, comprising 455,258 samples in total[49]. For MR method comparison, we analyze 912 plasma proteins that share four or more candidate SNPs within the summary statistics for AD.





(d)

ID	Description	Source
GO:0042288	MHC class I protein binding	GO:MF
GO:0042287	MHC protein binding	GO:MF
GO:0004714	transmembrane receptor protein tyrosine kinase activity	GO:MF
GO:0019199	transmembrane receptor protein kinase activity	GO:MF
GO:0005886	plasma membrane	GO:CC
GO:0045121	membrane raft	GO:CC
GO:0098857	membrane microdomain	GO:CC
GO:0071944	cell periphery	GO:CC
GO:0010562	positive regulation of phosphorus metabolic process	GO:BP
GO:0045937	positive regulation of phosphate metabolic process	GO:BP
GO:0030450	regulation of complement activation	GO:BP
GO:0033674	positive regulation of kinase activity	GO:BP
GO:0043085	positive regulation of catalytic activity	GO:BP
GO:0051347	positive regulation of transferase activity	GO:BP
GO:0002923	regulation of humoral immune response mediated by circulating immunoglobulin	GO:BP
GO:0019220	regulation of phosphate metabolic process	GO:BP
GO:0051174	regulation of phosphorus metabolic process	GO:BP
GO:0030449	regulation of complement activation	GO:BP
GO:0031399	regulation of protein modification process	GO:BP
GO:0044093	positive regulation of molecular function	GO:BP
GO:0032691	negative regulation of interleukin-1 beta production	GO:BP

Figure 5: **(a)** Volcano plot of associations of proteins with Alzheimer's disease using MR-SPI. The x-axis represents the estimated effect size (on the log odds ratio scale), and the y-axis represents the $-\log_{10}(p$-value). Positive and negative associations are represented by green and red points, respectively. The size of a point is proportional to the $-\log_{10}(p$-value). The blue dashed line represents the significance threshold using Bonferroni correction (p-value$< 5.48 \times 10^{-5}$). **(b)** Forest plot of significant associations of proteins with Alzheimer's disease identified by MR-SPI. Point estimates and 95% confidence intervals for the associations using the other competing MR methods are presented in different colors. Confidence intervals are clipped to y-axis limits. **(c)** Bubble plot of GO analysis results using the 7 significant proteins detected by MR-SPI. The x-axis represents the z-score of the enriched GO term, and the y-axis represents the $-\log_{10}(p$-value) after Bonferroni correction. Each point represents one enriched GO term. The blue dashed line represents the significance threshold (adjusted p-value< 0.05). **(d)** Table of the GO ID, description and source of the significant GO terms using the 7 significant proteins detected by MR-SPI. BP: biological process; CC: cellular component; MF: molecular function.

As presented in Figure 5(a), MR-SPI identifies 7 proteins that are significantly associated with AD after Bonferroni correction, including CD33, CD55, EPHA1, PILRA, PILRB, PRSS8, RET, and TREM2. Among them, 4 proteins contribute to an increased risk of AD (CD33, PILRA,

PILRB, and RET), while the other 3 proteins contribute to a decreased risk of AD (CD55, EPHA1, and TREM2). Previous studies have revealed that these proteins and the corresponding protein-coding genes might contribute to the pathogenesis of AD[50,51,52,53,54]. For example, it has been found that CD33 plays a key role in modulating microglial pathology in AD, with TREM2 acting downstream in this regulatory pathway[52]. Additionally, RET at mitochondrial complex I is activated during ageing, which might contribute to an increased risk of ageing-related diseases including AD[54]. These findings highlight the potential therapeutic opportunities that target these proteins for the treatment of AD.

In Figure 5(b), we present the point estimates and 95% confidence intervals of the effects (on the log odds ratio scale) of these 7 proteins on AD using the other competing MR methods. From Figure 5(b), these proteins are identified by most of the competing MR methods, confirming the robustness of our findings. Notably, in the relationship of TREM2 on AD, MR-SPI detects one possibly invalid IV, SNP rs10919543, which is associated with red blood cell count according to PhenoScanner. Red blood cell count is a known risk factor for AD[55,56], and thus SNP rs10919543 might exhibit pleiotropy in the relationship of TREM2 on AD. After excluding this potentially invalid IV, MR-SPI suggests that TREM2 is negatively associated with the risk of AD ($\widehat{\beta} = -0.04, p\text{-value} = 1.20 \times 10^{-18}$). Additionally, we perform the gene ontology (GO) enrichment analysis using the g:Profiler web server[57](`https://biit.cs.ut.ee/gprofiler/gost`) to gain biological insights for the set of proteins identified by MR-SPI, and the results are presented in Figure 5(c) and 5(d). After Bonferroni correction, the GO analysis indicates that these proteins are significantly enriched in 21 GO terms, such as the metabolic process, MHC protein binding, and transmembrane receptor protein kinase activity.

3 Discussion

In this paper, we develop a novel two-sample MR method and algorithm, named MR-SPI, to automatically select valid genetic instruments 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[31], which is less vulnerable to finite-sample IV selection error. We show with extensive simulation studies that MR-SPI can be helpful to select valid genetic instruments among candidate SNPs for a specific exposure-outcome pair and provide robust confidence interval for the causal effect when locally invalid IVs exist. Through real data analyses, we demonstrate that MR-SPI can provide reliable causal findings by automatically selecting valid genetic instruments. We apply MR-SPI to infer the causal relationships among 146 trait pairs and detect significant associations. Furthermore, we employ MR-SPI to conduct xMR analysis with 912 plasma proteins using the proteomics data from UK Biobank in 54,306 UK Biobank participants and identify 7 proteins significantly associated with the risk of Alzheimer's disease. These findings highlight the potential of MR-SPI as a powerful tool in the identification of new therapeutic targets for disease prevention and treatment.

We emphasize two main advantages of MR-SPI. First, MR-SPI can incorporate both exposure and outcome data to automatically select a set of valid genetic instruments in 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 selecting valid instruments for a specific exposure-outcome pair from GWAS studies for MR analyses, which is especially advantageous in the presence 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 finite-sample 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 using the robust confidence interval.

MR-SPI also has some limitations. First, 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 linkage disequilibrium (LD) structure from summary statistics of two possibly overlapping samples. Second, the robust confidence interval is slightly more conservative than the confidence interval calculated from the limiting distribution of the causal effect estimate, which is the price to pay for the gained robustness to finite-sample IV selection error. Further studies are needed to construct less conservative confidence intervals that are robust to finite-sample IV selection error.

In conclusion, MR-SPI provides an automatic approach to selecting valid instruments among candidate SNPs and perform reliable causal inference using two-sample GWAS summary statistics.

Simulation studies and real data analyses have shown that MR-SPI can provide reliable inference for the causal relationships even in the presence of invalid IVs. Our developed software is user-friendly and computationally efficient. Therefore, MR-SPI can detect more trustworthy causal relationships with increasingly rich and publicly available GWAS and multi-omics datasets.

Software availability

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

Data availability

All the GWAS data analyzed are publicly available with the following URLs:

- CARDIoGRAMplusC4D consortium: http://www.cardiogramplusc4d.org/data-downloads/;
- GIANT consortium: https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files;
- MEGASTROKE consortium: http://megastroke.org/download.html;
- Global Lipids Genetics Consortium (GLGC): http://csg.sph.umich.edu/willer/public/lipids2013/;
- GWAS for heart failure: https://www.ebi.ac.uk/gwas/publications/31919418;
- GWAS for atrial fibrillation: https://www.ebi.ac.uk/gwas/publications/30061737;
- The COVID-19 Host Genetics Initiative: https://www.covid19hg.org/;
- GWAS for Alzheimer's disease: https://ctg.cncr.nl/software/summary_statistics;
- UK Biobank proteomics data: https://europepmc.org/article/ppr/ppr508031.

References

[1] Kenneth J Rothman and Sander Greenland. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95(S1):S144–S150, 2005.

[2] Jan P Vandenbroucke, Alex Broadbent, and Neil Pearce. Causality and causal inference in

epidemiology: the need for a pluralistic approach. *International Journal of Epidemiology*, 45 (6):1776–1786, 2016.

[3] George Davey Smith and Shah Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

[4] Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008.

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

[6] George Davey Smith and Shah Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42, 2004.

[7] Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7): 658–665, 2013.

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

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

[10] Eric AW Slob and Stephen Burgess. A comparison of robust Mendelian randomization methods using summary data. *Genetic Epidemiology*, 44(4):313–329, 2020.

[11] Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, and Dylan S Small. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48(3):1742–1769, 2020.

[12] Jean Morrison, Nicholas Knoblauch, Joseph H Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*, 52(7):740–747, 2020.

[13] Qing Cheng, Xiao Zhang, Lin S Chen, and Jin Liu. Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. *Nature Communications*, 13(1):1–13, 2022.

[14] Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, 2007.

[15] Eleanor Sanderson, Tom G Richardson, Gibran Hemani, and George Davey Smith. The use of negative control outcomes in Mendelian randomization to detect potential population stratification. *International Journal of Epidemiology*, 50(4):1350–1361, 2021.

[16] Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7): 483–495, 2013.

[17] Shanya Sivakumaran, Felix Agakov, Evropi Theodoratou, James G Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, James F Wilson, and Harry Campbell. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, 89(5):607–618, 2011.

[18] Miles Parkes, Adrian Cortes, David A Van Heel, and Matthew A Brown. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nature Reviews Genetics*, 14(9):661–673, 2013.

[19] Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala Sheehan, and John Thompson. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*, 36(11):1783–1802, 2017.

[20] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.

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

[22] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

[23] Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, 50(5):693–698, 2018.

[24] Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. 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, 2018.

[25] Stephen Burgess, Christopher N Foley, Elias Allara, James R Staley, and Joanna MM Howson. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1):1–11, 2020.

[26] Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1998, 2017.

[27] Guanghao Qi and Nilanjan Chatterjee. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications*, 10(1): 1–10, 2019.

[28] Daniel I Swerdlow, Karoline B Kuchenbaecker, Sonia Shah, Reecha Sofat, Michael V Holmes, Jon White, Jennifer S Mindell, Mika Kivimaki, Eric J Brunner, John C Whittaker, et al. Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology*, 45(5):1600–1616, 2016.

[29] Qi Ouyang, Peter D Kaplan, Shumao Liu, and Albert Libchaber. DNA solution of the maximal clique problem. *Science*, 278(5337):446–449, 1997.

[30] Yoav Benjamini and Yosef Hochberg. 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, 1995.

[31] Zijian Guo. Post-selection problems for causal inference with invalid instruments: A solution using searching and sampling. *arXiv preprint arXiv:2104.06911*, 2021.

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

[33] Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre FR Stewart, Maja Barbalic, Christian Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature Genetics*, 43(4):333–338, 2011.

[34] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[35] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[36] James R Staley, James Blackshaw, Mihir A Kamat, Steve Ellis, Praveen Surendran, Benjamin B Sun, Dirk S Paul, Daniel Freitag, Stephen Burgess, John Danesh, et al. PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics*, 32(20):3207–3209, 2016.

[37] Mihir A Kamat, James A Blackshaw, Robin Young, Praveen Surendran, Stephen Burgess, John Danesh, Adam S Butterworth, and James R Staley. PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics*, 35(22):4851–4853, 2019.

[38] Zulvikar Syambani Ulhaq and Gita Vita Soraya. Interleukin-6 as a potential biomarker of covid-19 progression. *Medecine et maladies infectieuses*, 50(4):382, 2020.

[39] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[40] Cholesterol Treatment Trialists' (CTT) Collaboration. 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, 2012.

[41] Baris Gencer, Nicholas A Marston, KyungAh Im, Christopher P Cannon, Peter Sever, Anthony Keech, Eugene Braunwald, Robert P Giugliano, and Marc S Sabatine. 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, 2020.

[42] Juan R Rey, Juan Caro-Codón, Sandra O Rosillo, Ángel M Iniesta, Sergio Castrejón-Castrejón, Irene Marco-Clement, Lorena Martín-Polo, Carlos Merino-Argos, Laura Rodríguez-Sotelo, Jose M García-Veas, et al. Heart failure in COVID-19 patients: prevalence, incidence and prognostic implications. *European Journal of Heart Failure*, 22(12):2205–2215, 2020.

[43] RonaldM Krauss, PaulT Williams, John Brensike, KatherineM Detre, FrankT Lindgren, SherylF Kelsey, Karen Vranizan, and RobertI Levy. Intermediate-density lipoproteins and progression of coronary artery disease in hypercholesterolaemic men. *The Lancet*, 330(8550): 62–66, 1987.

[44] Daniel J Rader and G Kees Hovingh. HDL and cardiovascular disease. *The Lancet*, 384(9943): 618–625, 2014.

[45] Prospective Studies Collaboration. 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, 2007.

[46] Brandi J Witt, Robert D Brown Jr, Steven J Jacobsen, Susan A Weston, Karla V Ballman, Ryan A Meverden, and Véronique L Roger. Ischemic stroke after heart failure: a community-based study. *American Heart Journal*, 152(1):102–109, 2006.

[47] Karl Georg Haeusler, Ulrich Laufs, and Matthias Endres. Chronic heart failure and ischemic stroke. *Stroke*, 42(10):2977–2982, 2011.

[48] Benjamin B Sun, Joshua Chiou, Matthew Traylor, Christian Benner, Yi-Hsiang Hsu, Tom G Richardson, Praveen Surendran, Anubha Mahajan, Chloe Robins, Steven G Vasquez-Grinnell, et al. Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *BioRxiv*, pages 2022–06, 2022.

[49] Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*, 51(3):404–413, 2019.

[50] Adam C Naj, Gyungah Jun, Gary W Beecham, Li-San Wang, Badri Narayan Vardarajan, Jacqueline Buros, Paul J Gallins, Joseph D Buxbaum, Gail P Jarvik, Paul K Crane, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature genetics*, 43(5):436–441, 2011.

[51] Nisha Rathore, Sree Ranjani Ramani, Homer Pantua, Jian Payandeh, Tushar Bhangale, Arthur Wuster, Manav Kapoor, Yonglian Sun, Sharookh B Kapadia, Lino Gonzalez, et al. Paired Immunoglobulin-like Type 2 Receptor Alpha G78R variant alters ligand binding and confers protection to Alzheimer's disease. *PLoS genetics*, 14(11):e1007427, 2018.

[52] Ana Griciuc, Shaun Patel, Anthony N Federico, Se Hoon Choi, Brendan J Innes, Mary K Oram, Gea Cereghetti, Danielle McGinty, Anthony Anselmo, Ruslan I Sadreyev, et al. TREM2 acts downstream of CD33 in modulating microglial pathology in Alzheimer's disease. *Neuron*, 103(5):820–835, 2019.

[53] Hafdis T Helgadottir, Pär Lundin, Emelie Wallén Arzt, Anna-Karin Lindström, Caroline Graff, and Maria Eriksson. Somatic mutation that affects transcription factor binding upstream of CD55 in the temporal cortex of a late-onset Alzheimer disease patient. *Human Molecular Genetics*, 28(16):2675–2685, 2019.

[54] Suman Rimal, Ishaq Tantray, Yu Li, Tejinder Pal Khaket, Yanping Li, Sunil Bhurtel, Wen Li,

Cici Zeng, and Bingwei Lu. Reverse electron transfer is activated during aging and contributes to aging and age-related disease. *EMBO reports*, 24(4):e55548, 2023.

[55] Noel G Faux, Alan Rembach, James Wiley, Kathryn A Ellis, David Ames, Christopher J Fowler, Ralph N Martins, Kelly K Pertile, Rebecca L Rumble, B Trounson, et al. An anemia of Alzheimer's disease. *Molecular Psychiatry*, 19(11):1227–1234, 2014.

[56] Laura M Winchester, John Powell, Simon Lovestone, and Alejo J Nevado-Holgado. Red blood cell indices and anaemia as causative factors for cognitive function deficits and for Alzheimer's disease. *Genome Medicine*, 10(1):1–12, 2018.

[57] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 2019.

[58] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3):291–295, 2015.

[59] Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Online Methods

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 retains one representative SNP per LD region[34]. We also assume that the SNPs are standardized[58] such that $\mathbb{E}Z_j = 0$ and $\mathrm{Var}(Z_j) = 1$ for $1 \leq j \leq p$. Let D denote the exposure and Y denote the outcome. We assume that D and Y follow the exposure model $D = \boldsymbol{Z}^\intercal\boldsymbol{\gamma} + \delta$ and the outcome model $Y = D\beta + \boldsymbol{Z}^\intercal\boldsymbol{\pi} + e$, respectively, 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 assumptions (A2) and

(A3)[24,59]. 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 respective variances σ_δ^2 and σ_e^2 are possibly correlated due to unmeasured confounding factors. By plugging the exposure model into the outcome model, 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 $\boldsymbol{\Gamma} = \beta\boldsymbol{\gamma} + \boldsymbol{\pi}$. 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 $|\mathcal{V}| > \frac{1}{2}|\mathcal{S}|$[59], and the plurality rule condition can be expressed as $|\mathcal{V}| > \max_{c\neq 0} |\{j \in \mathcal{S} : \pi_j/\gamma_j = c\}|$[24]. 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 observation, our proposed MR-SPI selects the largest group of SNPs as valid IVs with similar ratio estimates of the causal effect using a voting procedure described in detail in the next subsection.

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 estimated standard errors respectively. Let $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \cdots, \widehat{\gamma}_p)^\intercal$ and $\widehat{\boldsymbol{\Gamma}} = (\widehat{\Gamma}_1, \cdots, \widehat{\Gamma}_p)^\intercal$ denote the vector of estimated SNP-exposure and SNP-outcome associations, respectively. 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

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

where the diagonal entries of $\mathbf{V}^\gamma$ and $\mathbf{V}^\Gamma$ are $\mathbf{V}^\gamma_{jj} = \mathrm{Var}(Z_{ij}^2)\gamma_j^2 + \sum_{l\neq j}\gamma_l^2 + \sigma_\delta^2$ and $\mathbf{V}^\Gamma_{jj} = \mathrm{Var}(Z_{ij}^2)\Gamma_j^2 + \sum_{l\neq j}\Gamma_l^2 + \sigma_\epsilon^2$, respectively, and the off-diagonal entries of $\mathbf{V}^\gamma$ and $\mathbf{V}^\Gamma$ are $\mathbf{V}^\gamma_{j_1j_2} = \gamma_{j_1}\gamma_{j_2}$ and $\mathbf{V}^\Gamma_{j_1j_2} = \Gamma_{j_1}\Gamma_{j_2}$ $(j_1 \neq j_2)$, respectively. The derivation of the limit distribution can be found in Supplementary Section S2. Therefore, with the summary statistics of the exposure and the outcome, we estimate $\frac{1}{n_1}\mathbf{V}^\gamma$ and $\frac{1}{n_2}\mathbf{V}^\Gamma$ as:

$$\frac{1}{n_1}\widehat{\mathbf{V}}^\gamma_{j_1j_2} = \begin{cases} \widehat{\sigma}^2_{\gamma_{j_1}} & \text{if } j_1 = j_2, \\ \frac{1}{n_1}\widehat{\gamma}_{j_1}\widehat{\gamma}_{j_2} & \text{if } j_1 \neq j_2. \end{cases} \quad \text{and} \quad \frac{1}{n_2}\widehat{\mathbf{V}}^\Gamma_{j_1j_2} = \begin{cases} \widehat{\sigma}^2_{\Gamma_{j_1}} & \text{if } j_1 = j_2, \\ \frac{1}{n_2}\widehat{\Gamma}_{j_1}\widehat{\Gamma}_{j_2} & \text{if } j_1 \neq j_2. \end{cases} \tag{1}$$

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 in the main text.

Selecting valid instruments by voting

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:

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

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 observation, 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 assumptions (A2) and (A3) is smaller than a threshold as in equation (4);

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

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

$$\widehat{\pi}_k^{[j]} = \widehat{\Gamma}_k - \widehat{\beta}^{[j]}\widehat{\gamma}_k = (\widehat{\beta}^{[k]} - \widehat{\beta}^{[j]})\widehat{\gamma}_k, \tag{3}$$

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 the kth SNP. From equation (3), $\widehat{\pi}_k^{[j]}$ 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 assumptions (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:

$$\frac{|\widehat{\pi}_k^{[j]}|}{\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]})} \leq \sqrt{\log \min(n_1, n_2)}, \tag{4}$$

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

$$\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]}) = \sqrt{\frac{1}{n_2}\left(\widehat{\mathbf{V}}_{kk}^{\Gamma} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)^2 \widehat{\mathbf{V}}_{jj}^{\Gamma} - 2\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\widehat{\mathbf{V}}_{jk}^{\Gamma}\right) + \frac{1}{n_1}(\widehat{\beta}^{[j]})^2\left(\widehat{\mathbf{V}}_{kk}^{\gamma} + \left(\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\right)^2 \widehat{\mathbf{V}}_{jj}^{\gamma} - 2\frac{\widehat{\gamma}_k}{\widehat{\gamma}_j}\widehat{\mathbf{V}}_{jk}^{\gamma}\right)}, \tag{5}$$

and the term $\sqrt{\log \min(n_1, n_2)}$ in equation (4) ensures that the violation of (A2) and (A3) can be correctly detected with probability one as the sample sizes go 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 (4). 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:

$$\widehat{\Pi}_{k,j} = I\left(\max\left\{\frac{|\widehat{\pi}_k^{[j]}|}{\widehat{\mathrm{SE}}(\widehat{\pi}_k^{[j]})}, \frac{|\widehat{\pi}_j^{[k]}|}{\widehat{\mathrm{SE}}(\widehat{\pi}_j^{[k]})}\right\} \leq \sqrt{\log \min(n_1, n_2)}\right), \tag{6}$$

where $I(\cdot)$ is the indicator function such that $I(A) = 1$ if event A happens and $I(A) = 0$ otherwise. From equation (6), 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 SNPs 1 and 2 are similar, while $\widehat{\Pi}_{1,4} = 0$ because the ratio estimates of SNPs 1 and 4 differ substantially, as SNPs 1 and 4 mutually "vote against" each other to be valid according to equation (4).

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[29]. 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. 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 of $\mathcal{V}$ in practice. Alternatively, we can also find the maximum clique in the voting matrix as an estimate of $\mathcal{V}$. A clique in the voting matrix is a group of IVs who mutually vote for each other to be valid, and the maximum clique is the clique with the largest possible number of IVs.

Estimation and inference of the causal effect

After selecting the set of valid genetic instruments $\widehat{\mathcal{V}}$, the causal effect β is estimated by

$$\widehat{\beta}_{\mathrm{SPI}} = \frac{\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_{\widehat{\mathcal{V}}}}{\widehat{\boldsymbol{\gamma}}_{\widehat{\mathcal{V}}}^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_{\widehat{\mathcal{V}}}}, \tag{7}$$

where $\widehat{\boldsymbol{\gamma}}_{\widehat{\mathcal{V}}}$ and $\widehat{\mathbf{\Gamma}}_{\widehat{\mathcal{V}}}$ are the estimates of SNP-exposure associations and SNP-outcome associations of the selected valid IVs in $\widehat{\mathcal{V}}$, respectively. 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:

$$\mathrm{CI} = \left(\widehat{\beta}_{\mathrm{SPI}} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{SPI}})}, \widehat{\beta}_{\mathrm{SPI}} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{SPI}})}\right), \tag{8}$$

where $\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{SPI}})$ is the estimated variance of $\widehat{\beta}_{\mathrm{SPI}}$, which can be found in Supplementary Section S4. As $\min\{n_1, n_2\} \to \infty$, we have $\mathbb{P}\left\{\beta \in \left(\widehat{\beta}_{\mathrm{SPI}} - z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{SPI}})}, \widehat{\beta}_{\mathrm{SPI}} + z_{1-\frac{\alpha}{2}}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{SPI}})}\right)\right\} \to 1-\alpha$ under the plurality rule condition, as shown 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 proposed 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 (1);

Select the set of relevant IVs $\widehat{\mathcal{S}}$ as in equation (2);

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 (6);

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 as in equation (7), and construct the corresponding confidence interval as in equation (8) using the selected valid IVs in $\widehat{\mathcal{V}}$.

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" [31]. When locally invalid IVs exist and are incorrectly selected into $\widehat{\mathcal{V}}$, the confidence interval in equation (8) becomes unreliable, since its validity (i.e., the coverage probability attains the nominal level) requires that the invalid IVs are correctly filtered out. In practice, we can multiply the threshold $\sqrt{\log\min(n_1,n_2)}$ in the right-hand side of equation (4) by a scaling factor η to examine whether the confidence interval calculated by equation (8) is sensitive to the choice of the threshold. If the confidence interval varies substantially to the choice of the scaling factor η, then there might exist finite-sample IV selection error especially with locally invalid IVs. We demonstrate this issue with two numerical examples presented in Supplementary Figure S12. Supplementary Figure S12(a) shows an example in which MR-SPI provides robust inference across difference values of the scaling factor, while Supplementary Figure S12(b) shows an example that MR-SPI might suffer

from finite-sample 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.

To construct a confidence interval that is robust to finite-sample IV selection error, we borrow the idea of searching and sampling[31], with main steps described in Figure 6. The key idea is to sample the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ repeatedly from the following distribution:

$$\begin{pmatrix} \widehat{\boldsymbol{\gamma}}^{(m)} \\ \widehat{\boldsymbol{\Gamma}}^{(m)} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left[\begin{pmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\boldsymbol{\Gamma}} \end{pmatrix}, \begin{pmatrix} \frac{1}{n_1}\widehat{\mathbf{V}}^{\gamma} & \mathbf{0} \\ \mathbf{0} & \frac{1}{n_2}\widehat{\mathbf{V}}^{\Gamma} \end{pmatrix} \right], \quad m = 1, \cdots, M, \tag{9}$$

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 selected IVs 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:

$$\left(\widehat{\beta}^{[j]} - \sqrt{\log \min(n_1, n_2) \widehat{\mathrm{Var}}(\widehat{\beta}^{[j]})}, \widehat{\beta}^{[j]} + \sqrt{\log \min(n_1, n_2) \widehat{\mathrm{Var}}(\widehat{\beta}^{[j]})} \right) \quad \text{for} \quad j \in \widehat{\mathcal{V}}, \tag{10}$$

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 (4). 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 π_j by $\widehat{\pi}_j^{(m)}(b) = \left(\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)} \right) \cdot \mathbf{1}\left(|\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)}| \geq \lambda \widehat{\rho}_j(b, \alpha) \right)$ for $j \in \widehat{\mathcal{V}}$, 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

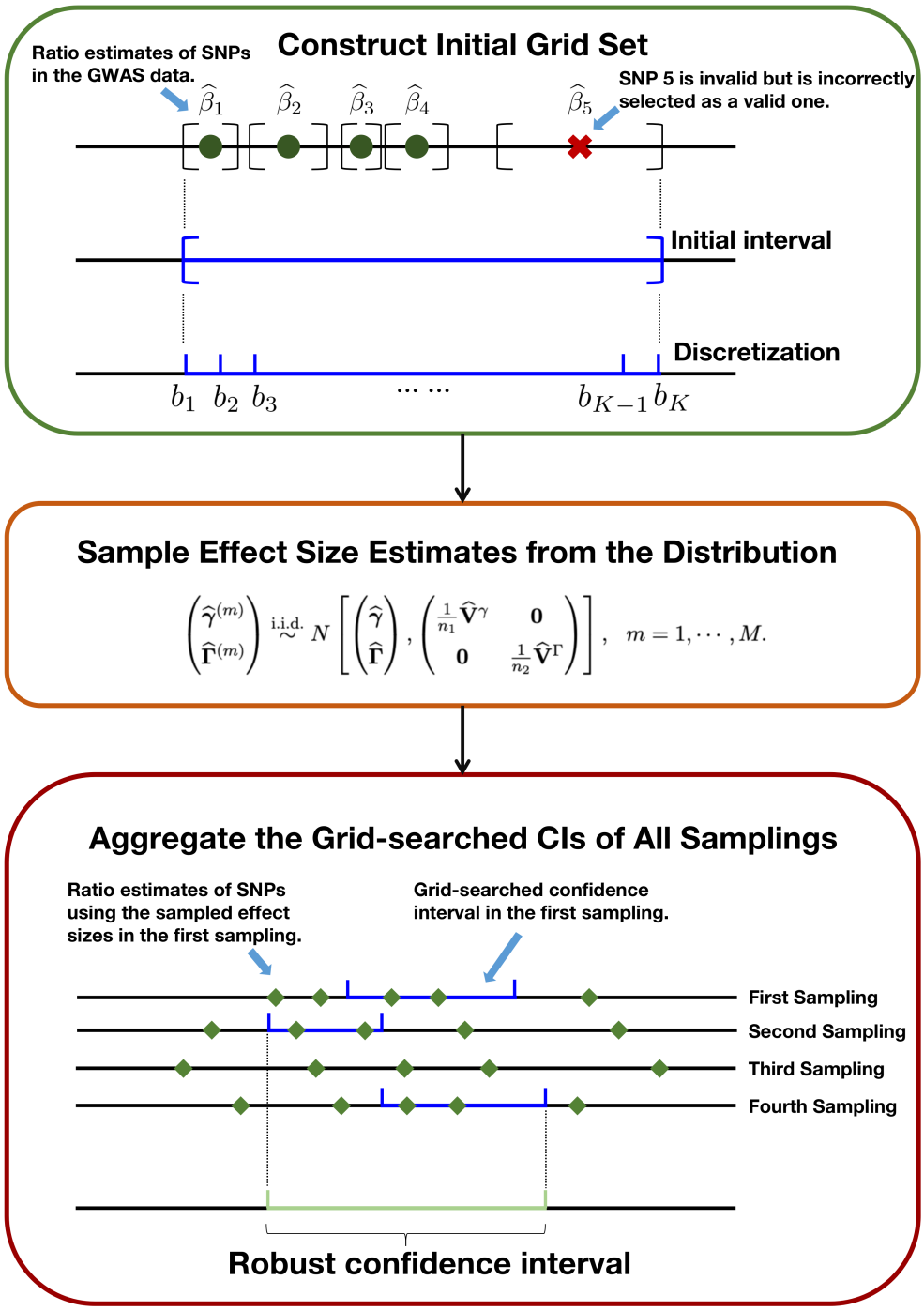


Figure 6: The procedure of constructing the robust confidence interval by MR-SPI. When locally invalid IVs exist in finite samples, MR-SPI might incorrectly select invalid IVs as valid ones (marked by the red cross). In such cases, a robust confidence interval can be constructed to improve the coverage 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 green line segment). Note that the confidence interval in the third sampling is empty since the majority rule is violated.

factor to make the thresholding more stringent so that the confidence interval in each sampling is shorter, as we will show shortly. 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 $\widehat{\boldsymbol{\pi}}_{\widehat{\mathcal{V}}}^{(m)}(b) = (\widehat{\pi}_j^{(m)}(b))_{j\in\widehat{\mathcal{V}}}$, 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. Define $\beta_{\min}^{(m)} = \min\{b \in \mathcal{B} : \|\widehat{\boldsymbol{\pi}}_{\widehat{\mathcal{V}}}^{(m)}(b)\|_0 < |\widehat{\mathcal{V}}|/2\}$ and $\beta_{\max}^{(m)} = \max\{b \in \mathcal{B} : \|\widehat{\boldsymbol{\pi}}_{\widehat{\mathcal{V}}}^{(m)}(b)\|_0 < |\widehat{\mathcal{V}}|/2\}$, then the mth sampling's confidence interval is constructed as $\mathrm{CI}^{(m)} = \left(\beta_{\min}^{(m)}, \beta_{\max}^{(m)}\right)$.

From the definitions of $\widehat{\pi}_j^{(m)}(b)$ and $\mathrm{CI}^{(m)}$, 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 fewer $b \in \mathcal{B}$ satisfying $\|\widehat{\boldsymbol{\pi}}_{\widehat{\mathcal{V}}}^{(m)}(b)\|_0 < |\widehat{\mathcal{V}}|/2$, 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 $\mathcal{M} = \{1 \le m \le M : \mathrm{CI}^{(m)} \neq \varnothing\}$ denote the set of all sampling indexes corresponding to non-empty searching confidence intervals, then the proposed robust confidence interval is given by:

$$\mathrm{CI}^{\mathrm{robust}} = \left(\min_{m\in\mathcal{M}} \beta_{\min}^{(m)}, \max_{m\in\mathcal{M}} \beta_{\max}^{(m)}\right). \tag{11}$$

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 $\widehat{\mathcal{V}}$ as in Algorithm 1;

Construct the initial interval $[L, U]$ as in equation (10) and obtain the corresponding grid set $\mathcal{B}$;

for $m \leftarrow 1$ **to** M **do**

  Sample $\widehat{\boldsymbol{\gamma}}^{(m)}$ and $\widehat{\boldsymbol{\Gamma}}^{(m)}$ from the distribution in equation (9);

  Calculate $\{\widehat{\boldsymbol{\pi}}_{\widehat{\mathcal{V}}}^{(m)}(b)\}_{b\in\mathcal{B}}$ by $\widehat{\pi}_j^{(m)}(b) = \left(\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)}\right) \cdot \mathbf{1}\left(|\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)}| \ge \lambda\widehat{\rho}_j(b,\alpha)\right), j \in \widehat{\mathcal{V}}$;

  Construct $\mathrm{CI}^{(m)}$ by grid search over $\mathcal{B}$;

end

Construct the robust confidence interval $\mathrm{CI}^{\mathrm{robust}}$ as in equation (11);

Simulation settings

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 $\text{Bin}(2, f_j)$, 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 the exposure model and the outcome model, respectively. Finally, we calculate the genetic associations and their corresponding standard errors for the exposure and the outcome, respectively. As for the parameters, we fix the causal effect $\beta = 1$, and we consider 4 settings 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 $\boldsymbol{\pi} = 0.2 \cdot (\mathbf{0}_4, \mathbf{1}_2, 0.25, \mathbf{1}_2, -0.25)^\intercal$.

Settings (**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 assumptions (A2) and (A3) down to 0.25 times in these two settings. In total, we run 1000 replications in each setting.

Implementation of existing MR methods

We compare the performance of MR-SPI with eight other MR methods in simulation studies and real data analyses. These methods are implemented as follows:

- Random-effects IVW, MR-Egger, the weighted median method, the mode-based estimation and the contamination mixture method are implemented in the R package "MendelianRandomization" (https://github.com/cran/MendelianRandomization). The mode-based estimation is run with iteration=1000. All other methods are run with the default parameters.

- MR-PRESSO is implemented in the R package "MR-PRESSO" (https://github.com/rondolab/MR-PRESSO) with outlier test and distortion test.

- MR-RAPS is performed using the R package "mr.raps" (https://github.com/qingyuanzhao/mr.raps) with the default options.
- MRMix is run with the R package "MRMix" (https://github.com/gqi/MRMix) using the default options.