Robust Mendelian Randomization Analysis by Automatically Selecting Valid Genetic Instruments with Applications to Identify Plasma Protein Biomarkers for Alzheimer's Disease

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 infer the causal effect of a modifiable exposure on the outcome of interest by removing unmeasured confounding bias. However, some genetic variants might be invalid IVs due to violations of core IV assumptions. MR analysis with invalid IVs 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. We analyze 912 plasma proteins using the large-scale UK Biobank proteomics data in 54,306 participants and identify 7 proteins (TREM2, PILRB, PILRA, EPHA1, CD33, RET, CD55) significantly associated with the risk of Alzheimer's disease. We employ AlphaFold2 to predict the 3D structural alterations of these 7 proteins due to missense genetic variations, providing new insights into their biological functions in disease etiology.

1 Introduction

In biomedical 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/diseases, 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 IV 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 significantly 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. For example, violation of (A2) may occur due to the presence of population stratification[4,15]; and 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 those methods, MRMix and the contamination mixture method require additional distributional assumptions on the genetic associations, or the ratio estimates to provide reliable causal inference. Despite many efforts, most 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 such traditional way of selecting IVs only requires the exposure data, hence the same set of selected 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 underlying genetic architecture may vary across outcomes. For example, the pattern of horizontal pleiotropy might vary across different outcomes. Therefore, it is desirable to develop an automatic algorithm to select a set of valid genetic 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 genetic IVs for a specific exposure-outcome pair and then performs honest Post-selection Inference (MR-SPI) for the causal effect of interest. The key idea of MR-SPI is based on the Anna Karenina Principle which states that all valid instruments are alike, while each invalid instrument is invalid in its own way – paralleling Leo Tolstoy's dictum that "all happy families are alike; each unhappy family is unhappy in its own way" [29]. In other words, valid instruments will form a group and should provide similar ratio estimates of the causal effect, while the ratio estimates of invalid instruments are more likely to be different from each other. Operationally, MR-SPI consists of the following four steps: (1) select relevant genetic IVs that are associated with the exposure; (2) 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; (3) 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 (4) perform post-selection inference to construct an honest confidence interval for the causal effect that is robust to any potential 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 set of valid genetic IVs for a specific exposure-outcome pair. 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 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 causal relationships among 146 exposure-outcome pairs involving COVID-19 (Coronavirus disease 2019) related traits, ischemic stroke, cholesterol levels and heart diseases, and detect significant associations among them. Furthermore, we employ MR-SPI to perform omics MR (xMR) with 912 plasma proteins using the large-scale UK Biobank proteomics data in 54,306 UK Biobank participants [30] and discover 7 proteins significantly associated with the risk of Alzheimer's disease. We also use AlphaFold2 [31,32,33] to predict the 3D structural changes of these 7 proteins due to missense genetic variations, and then illustrate the structural changes graphically using the PyMOL software (`https://pymol.org`).

2 Results

2.1 MR-SPI selects valid genetic instruments by a 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 consists of the following four steps, as illustrated in Figure 1:

(i) select relevant SNPs that are strongly associated with the exposure;

(ii) each relevant SNP provides a ratio estimate of the causal effect, and then 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 data-dependent threshold as in equation (4);

(iii) select valid SNP IVs by majority/plurality voting or by finding the maximum clique of the voting matrix that encodes whether two relevant SNP IVs mutually vote for each other to be valid (the voting matrix is defined in equation (6) in Online Methods);

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

Most 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 and should give "similar" ratio estimates according to the Anna Karenina Principle (see Online Methods). More 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 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. We propose two ways to select valid genetic instruments based on the voting matrix (see Online Methods): (1) select relevant SNPs who receive majority voting or plurality voting as valid IVs; and (2) use SNPs in the maximum clique of the voting matrix as valid IVs[34]. Our simulation studies show that the maximum clique method can empirically offer lower false discovery rate (FDR)[35] and higher true positive proportion (TPP) as shown in Table S4 and Supplementary Section S6.

In step (iv), we estimate the causal effect by fitting a zero-intercept ordinary least squares (OLS) regression of SNP-outcome associations on SNP-exposure associations using the set of selected valid genetic instruments, and then construct a standard confidence interval for the causal effect using standard linear regression theory. 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, commonly referred to as "locally invalid IVs"[36]. To address this possible issue, we propose to construct a robust confidence interval with a guaranteed nominal coverage even in the presence of IV selection error in finite-sample settings using a searching and sampling method[36], as described in Figure 5 and Online Methods.

2.2 Comparing MR-SPI to other competing MR 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 MR-SPI to the following competing MR methods: (i) the random-effects IVW method[19], (ii) MR-RAPS[11], (iii) MR-PRESSO[23], (iv) the weighted median method[22], (v) the mode-based estimation[26], (vi) MRMix[27], and (vii) the contamination mixture method[25]. We exclude MR-Egger in this simulation since it is heavily biased under our simulation settings.

For simplicity, we shall use IVW to represent the random-effects IVW method hereafter.

In Figure 2, we present the percent bias, empirical coverage, and average lengths of 95% confidence intervals of the aforementioned MR methods in simulated data with a sample size of 5,000 for both the exposure and the outcome. Additional simulation results under a range of sample sizes (n = 5,000, 10,000, 20,000, 40,000, 80,000) can be found in Supplementary Figure S1 and Tables S1-S3. When the plurality rule condition holds and no locally invalid IVs exist, MR-SPI has small bias and short confidence interval, and the empirical coverage can attain the nominal level, suggesting the superior performance of MR-SPI. When locally invalid IVs exist, the standard confidence interval might suffer from finite-sample IV selection error, and thus the empirical coverage is lower than 95% if the sample sizes are not large (e.g., 5,000). In practice, we can perform sensitivity analysis of the causal effect estimate by changing the threshold in the voting step (see Online Methods and Supplementary Figure S13). If the causal effect estimate is sensitive to the choice of the threshold, then there might exist finite-sample IV selection error. In such cases, the proposed robust confidence interval of MR-SPI can still attain the 95% coverage level and thus is recommended for use.

2.3 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 diseases, and coronavirus disease 2019 (COVID-19) related traits. We exclude the trait pairs whose exposure and outcome are in the same consortium or are two similar phenotypes (for example, heart failure and coronary artery disease), and we finally obtain 146 pairwise exposure-outcome combinations. Among the 146 exposure-outcome pairs, MR-SPI detects invalid IVs for 16 exposure-outcome pairs. For example, MR-SPI detects one invalid SNP (rs616154, marked by red triangle) in the causal relationship from cardioembolic stroke (CES) to SARS-CoV-2 infection, as illustrated in Figure 3(a). Using the PhenoScanner tool[37,38], we find that SNP rs616154 is also associated with the Interleukin-6 (IL-6) levels which is a potential biomarker of COVID-19 progression[39], indicating that this SNP might exhibit horizontal pleiotropy in the relationship of cardioembolic stroke on SARS-CoV-2 infection and might be an invalid IV.

After excluding those potentially invalid IVs, MR-SPI identifies 27 significant associations after Bonferroni correction for multiple comparison[40], with results summarized in Figure 3(c). MR-SPI

detects some significant associations that cannot be discovered by other competing MR methods considered in this paper. For example, MR-SPI suggests that SARS-CoV-2 infection might be a risk factor for HF, which 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[41]. We also present the Venn diagram of the significant pairs using MR-SPI, the mode-based estimation, MRMix and the contamination mixture method in Figure 3(b), as these four methods are all based on the plurality rule condition. Using the robust confidence intervals constructed by Algorithm 2 in Online Methods, MR-SPI also discovers four significant associations that are immune to finite-sample IV selection error (CES on heart failure (HF), low-density lipoprotein (LDL) on coronary artery disease (CAD), total cholesterol (TC) on CAD, and atrial fibrillation (AF) on CES), as shown in Figure 3(d). More detailed results can be found in Supplementary Section S10.

2.4 Identifying plasma proteins associated with the risk of Alzheimer's disease

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 disease diagnosis, monitoring, and novel drug target discovery. In this section, we apply MR-SPI to identify plasma 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)[30]. Following the guidelines[30], 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[42]. For MR method comparison, we analyze 912 plasma proteins that share four or more candidate SNPs within the summary statistics for AD, because the implementation of MR-PRESSO requires a minimum of four SNPs as candidate IVs[23].

As presented in Figure 4(a), MR-SPI identifies 7 proteins that are significantly associated with AD after Bonferroni correction, including CD33, CD55, EPHA1, PILRA, PILRB, RET, and TREM2. The detailed information of the selected SNP IVs for these 7 proteins can be found in Supplementary Table S6. Among them, four proteins (CD33, PILRA, PILRB, and RET)

are positively associated with the risk of AD while the other three proteins (CD55, EPHA1, and TREM2) are negatively associated with the risk of AD. Previous studies have revealed that some of those 7 proteins and the corresponding protein-coding genes might contribute to the pathogenesis of AD[43,44,45,46,47,48]. 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[45]. Besides, a recent study has shown that a higher level of soluble TREM2 is associated with protection against the progression of AD pathology[49]. Additionally, RET at mitochondrial complex I is activated during ageing, which might contribute to an increased risk of ageing-related diseases including AD[47]. Using the UniProt database[50], we also find that genes encoding these 7 proteins are overexpressed in tissues including hemopoietic tissues and brain, as well as cell types including microglial, macrophages and dendritic cells. These findings highlight the potential therapeutic opportunities that target these proteins for the treatment of AD. Furthermore, in the Therapeutic Target Database (TTD)[51] and DrugBank database[52], we find existing US Food and Drug Administration (FDA)-approved drugs that target these proteins identified by MR-SPI. For example, gemtuzumab ozogamicin is a drug that targets CD33 and has been approved by FDA for acute myeloid leukemia therapy[53,54]. Besides, pralsetinib and selpercatinib are two RET inhibitors that have been FDA-approved for the treatment of non-small-cell lung cancers[55,56]. Therefore, these drugs might be potential drug repurposing candidates for the treatment of AD.

In Figure 4(b), we present the 3D structural alterations of CD33 due to missense genetic variation of SNP rs2455069, as predicted by AlphaFold2[31,32]. The 3D structures are shown in blue when the allele is A, and in red when the allele is G at SNP rs2455069 A/G, which is a cis-SNP located on chromosome 19 (19q13.41) and is selected as a valid IV by MR-SPI. The presence of the G allele at SNP rs2455069 results in the substitution of the 69th amino acid of CD33, changing it from Arginine (colored in green if the allele is A) to Glycine (colored in yellow if the allele is G), consequently causing a local change in the structure of CD33. Previous studies have found that CD33 is overexpressed in microglial cells in the brain[57], and the substitution of Arginine to Glycine in the 69th amino acid of CD33 might lead to the accumulation of amyloid plaques in the brain[58], thus the presence of the G allele at SNP rs2455069 might contribute to an increased risk of AD. We also apply AlphaFold2 to predict the 3D structures of the other 6 proteins that are detected to be significantly associated with AD by MR-SPI, which are presented in Supplementary Figure S15.

In Figure 4(c), we present the point estimates and 95% confidence intervals of the causal effects (on the log odds ratio scale) of these 7 proteins on AD using the other competing MR methods. In Figure 4(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[59,60], 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[61] (https://biit.cs.ut.ee/gprofiler/gost) to gain more biological insights for the 7 proteins identified by MR-SPI, and the results are presented in Figure 4(d) and Supplementary Table S7. After Bonferroni correction, the GO analysis indicates that these 7 proteins are significantly enriched in 20 GO terms, notably, the positive regulation of phosphorus metabolic process and major histocompatibility complex (MHC) class I protein binding. It has been found that increased phosphorus metabolites (e.g., phosphocreatine) are associated with aging, and that defects in metabolic processes for phospholipid membrane function is involved in the pathological progression of Alzheimer's disease[62,63]. In addition, MHC class I proteins may play a crucial role in preserving brain integrity during post-developmental stages, and modulation of the stability of MHC class I proteins emerges as a potential therapeutic target for restoring synaptic function in AD[64,65,66].

3 Discussion

In this paper, we develop a novel two-sample MR method called MR-SPI, to automatically select valid genetic instruments for a specific exposure-outcome pair from GWAS studies and perform valid post-selection inference. MR-SPI first selects relevant IVs with strong SNP-exposure associations, and then applies the proposed voting procedure to select valid IVs whose ratio estimates are similar to each other as valid IVs. In the possible presence of locally invalid IVs in finite-sample settings, MR-SPI provides a robust confidence interval constructed by the searching and sampling method[36], which is immune to finite-sample IV selection error. We employ MR-SPI to conduct xMR analysis with 912 plasma proteins using the proteomics data in 54,306 UK Biobank partici-

pants and identify 7 proteins significantly associated with the risk of Alzheimer's disease. The 3D structural changes in these proteins, as predicted by AlphaFold2 in response to missense genetic variations of selected SNP IVs, shed new insights to their biological functions in the etiology of Alzheimer's disease. We also find existing FDA-approved drugs that target some of our identified proteins, which provide opportunities for potential existing drug repurposing for the treatment of Alzheimer's disease. These findings highlight the great potential of MR-SPI as a powerful tool for identifying protein biomarkers as new therapeutic targets and drug repurposing for disease prevention and treatment.

We emphasize two main advantages of MR-SPI. First, MR-SPI incorporates 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 any additional distributional assumptions on the genetic effects. Therefore, MR-SPI is the first to offer such a practically robust approach to selecting valid genetic instruments for a specific exposure-outcome pair from GWAS studies for more reliable 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 immune to finite-sample IV selection error. Therefore, when locally invalid IVs are incorrectly selected and the causal effect estimate is biased in finite samples, MR-SPI can still provide reliable inference for the causal effect using the proposed robust confidence interval.

MR-SPI also has some limitations. First, MR-SPI uses independent SNPs from two non-overlapping samples. For future work, we plan to extend MR-SPI to include SNPs with arbitrary linkage disequilibrium (LD) structure from GWAS summary statistics of two possibly overlapping samples. Second, the proposed 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. Future work is 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 genetic instruments among candidate SNPs and performs reliable causal inference using two-sample GWAS summary statistics. Our software is user-friendly and computationally efficient. Therefore, MR-SPI can help 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://www.biorxiv.org/content/10.1101/2022.06.17.496443v1.supplementary-material.

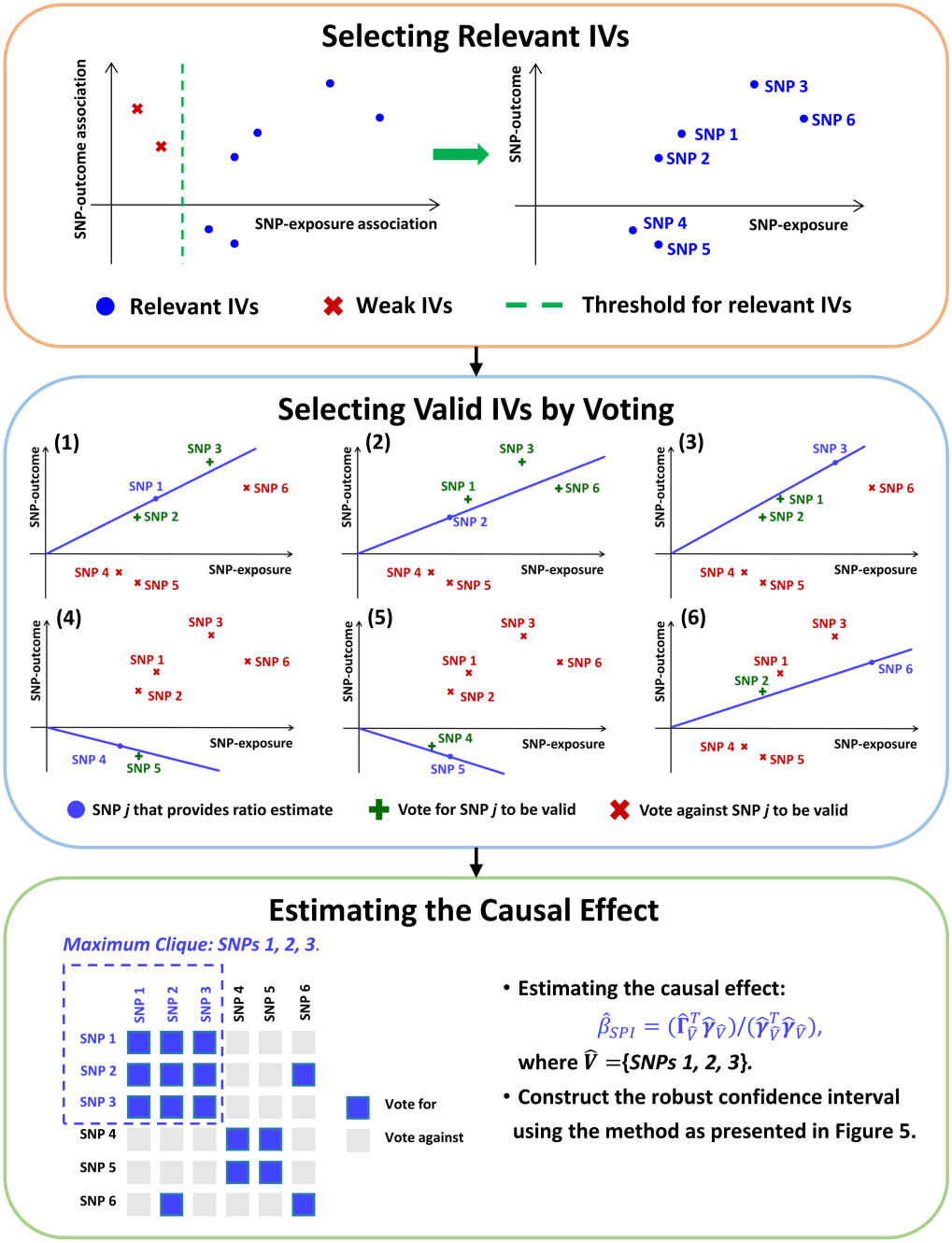

Figure 1: The MR-SPI framework. First, MR-SPI selects relevant IVs with strong SNP-exposure associations. Second, each relevant IV provides a ratio estimate of the causal effect and then receives votes on itself to be valid from the other relevant IVs whose degrees of violation of (A2) and (A3) are small under this ratio estimate of causal effect. For example, by assuming SNP 1 is valid, the slope of the line connecting SNP 1 and the origin represents the ratio estimate of SNP 1, and 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. Third, MR-SPI estimates the causal effect by fitting a zero-intercept OLS regression of SNP-outcome associations on SNP-exposure associations and construct the robust confidence interval using selected valid SNP IVs in the maximum clique of the voting matrix, which encodes whether two SNPs mutually vote for each other to be valid IVs.

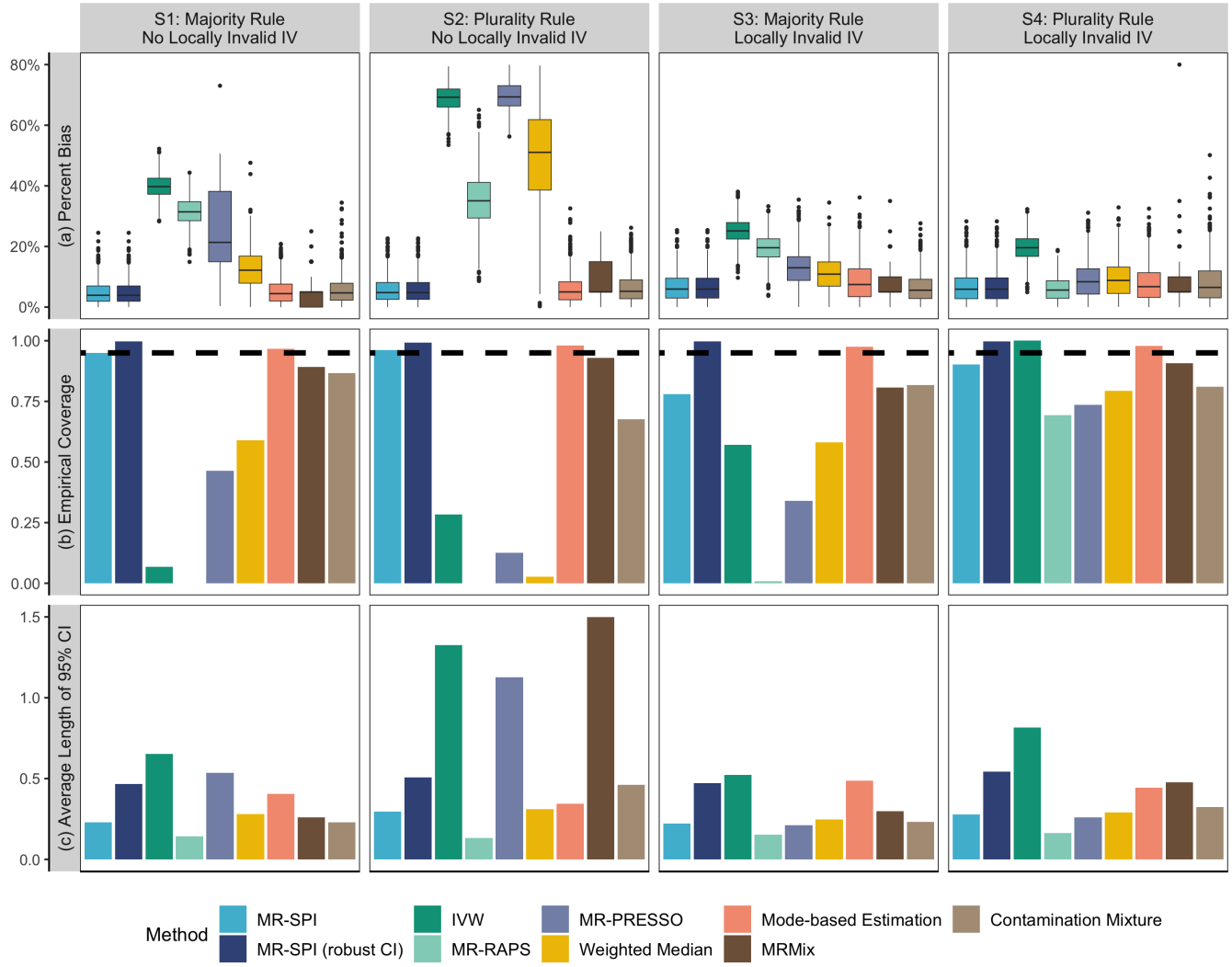

Figure 2: Empirical performance of MR-SPI and the other competing MR methods in simulated data with sample sizes of 5,000. (a) Boxplot of the percent bias in causal effect estimates. (b) Empirical coverage of 95% confidence intervals. The black dashed line in (b) represents the nominal level (95%). (c) Average lengths of 95% confidence intervals.

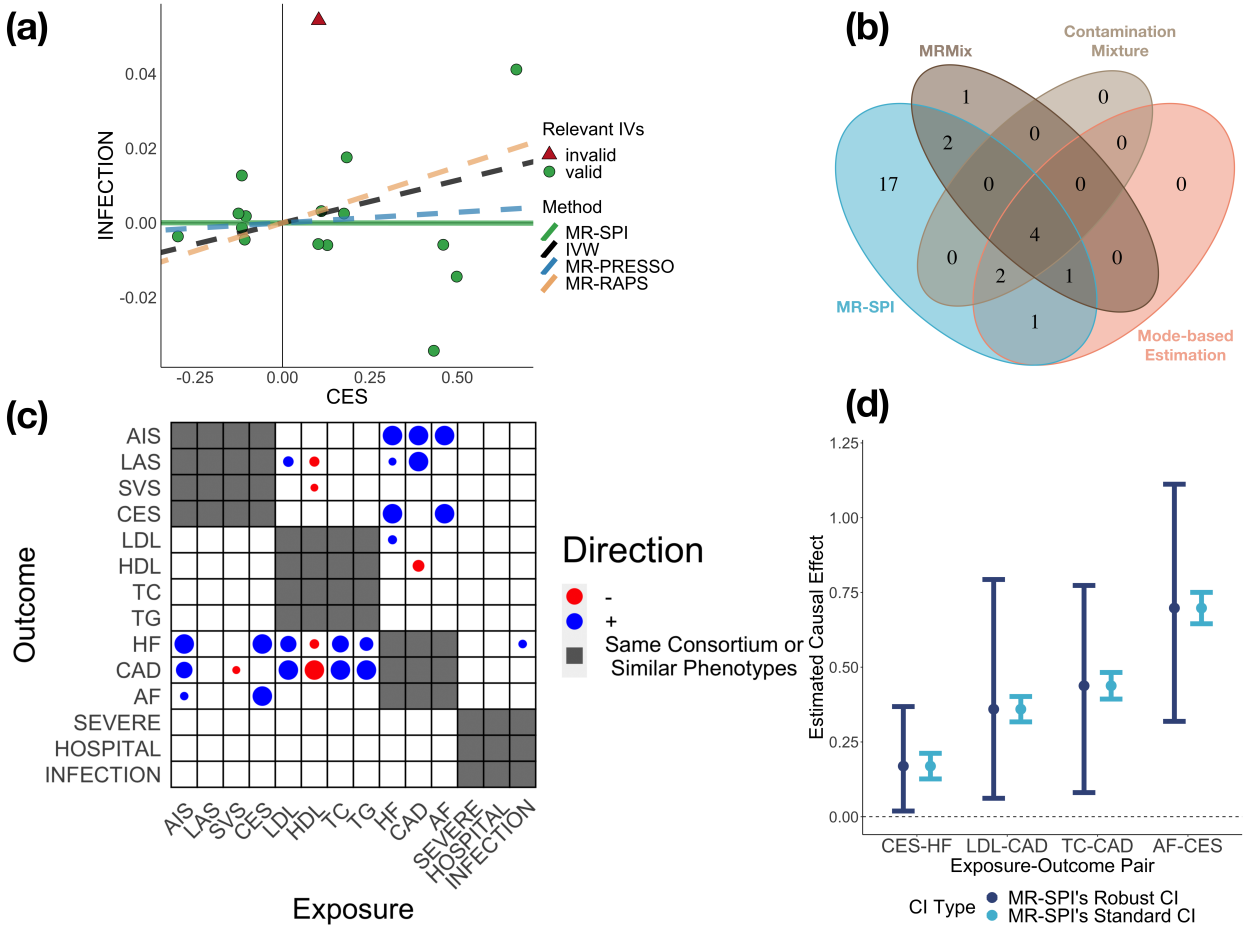

Figure 3: **(a)** Scatter plot of cardioembolic stroke on SARS-CoV-2 infection. The horizontal and vertical axes represent the SNP-exposure and SNP outcome associations, respectively. The slope of the green solid line represents the causal effect estimate of MR-SPI. The slopes of the black, blue and 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)** Venn diagram of significant associations detected by MR-SPI, the mode-based estimation, MRMix and the contamination mixture method after Bonferroni correction. **(c)** 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. **(d)** Significant associations detected by MR-SPI using the robust confidence interval. The light blue bars represent the standard confidence interval calculated using the causal effect estimates and the corresponding standard errors using standard linear regression theory. The dark blue bars represent the robust confidence interval constructed by the searching and sampling method, which allows for finite-sample IV selection error.
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.

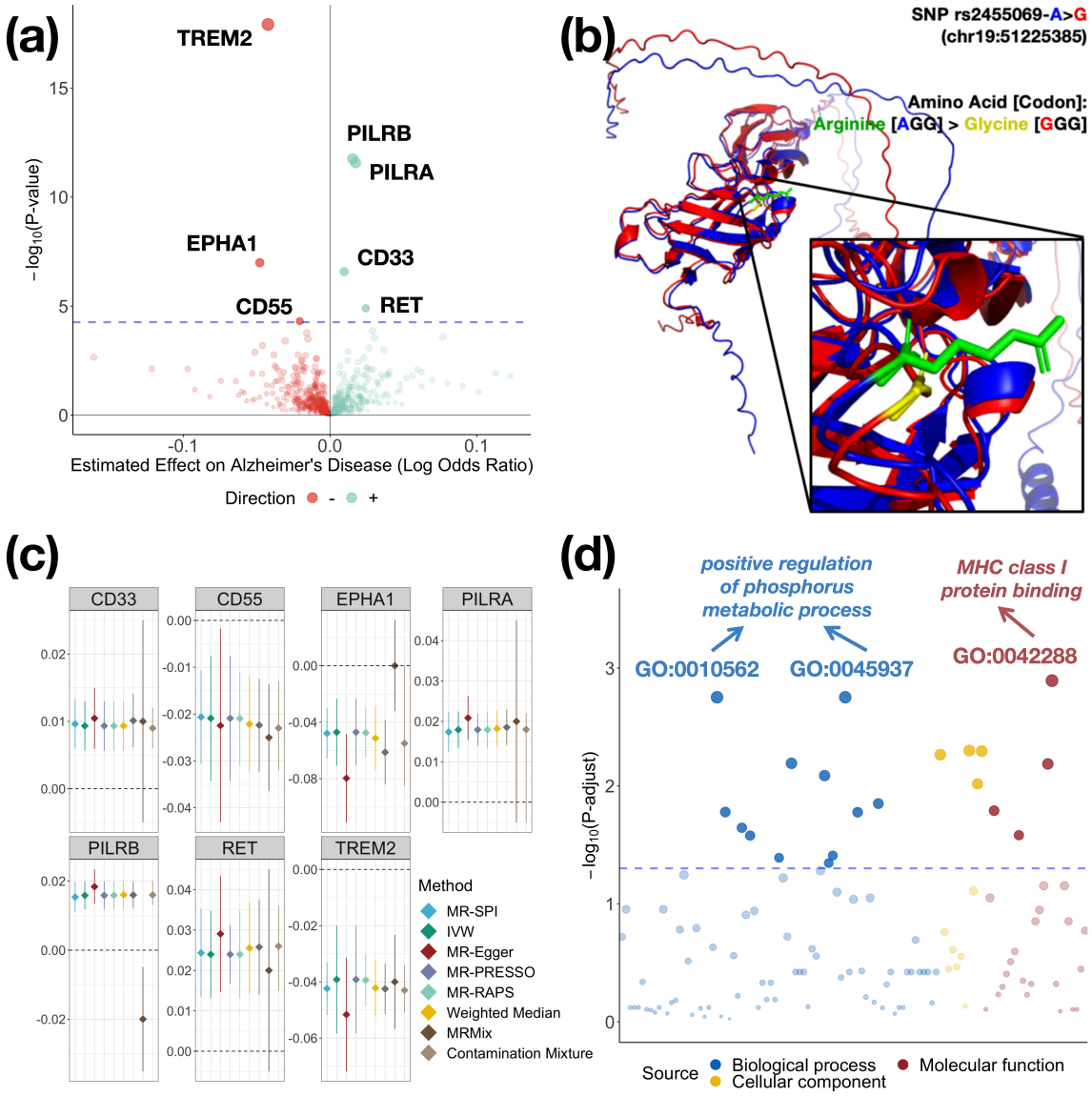

Figure 4: **(a)** Volcano plot of associations of plasma proteins with Alzheimer's disease using MR-SPI. The horizontal axis represents the estimated effect size (on the log odds ratio scale), and the vertical axis represents the $-\log_{10}(p\text{-value})$. Positive and negative associations are represented by green and red points, respectively. The size of a point is proportional to the $-\log_{10}(p\text{-value})$. The blue dashed line represents the significance threshold using Bonferroni correction (p-value $< 5.48\times10^{-5}$). **(b)** 3D Structural alterations of CD33 predicted by AlphaFold2 due to missense genetic variation of SNP rs2455069. The ribbon representation of 3D structures of CD33 with Arginine and Glycine at position 69 are colored in blue and red, respectively. The amino acids at position 69 are displayed in stick representation, with Arginine and Glycine colored in green and yellow, respectively. The predicted local-distance difference test (pLDDT) yields a value of 77.1% for both structures, which suggests that AlphaFold2 generally provides good backbone predictions for these two structures. **(c)** 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 vertical axis limits. **(d)** Bubble plot of GO analysis results using the 7 significant proteins detected by MR-SPI. The horizontal axis represents the z-score of the enriched GO term, and the vertical axis represents the $-\log_{10}(p\text{-value})$ after Bonferroni correction. Each point represents one enriched GO term. The blue dashed line represents the significance threshold (adjusted p-value < 0.05 after Bonferroni correction).

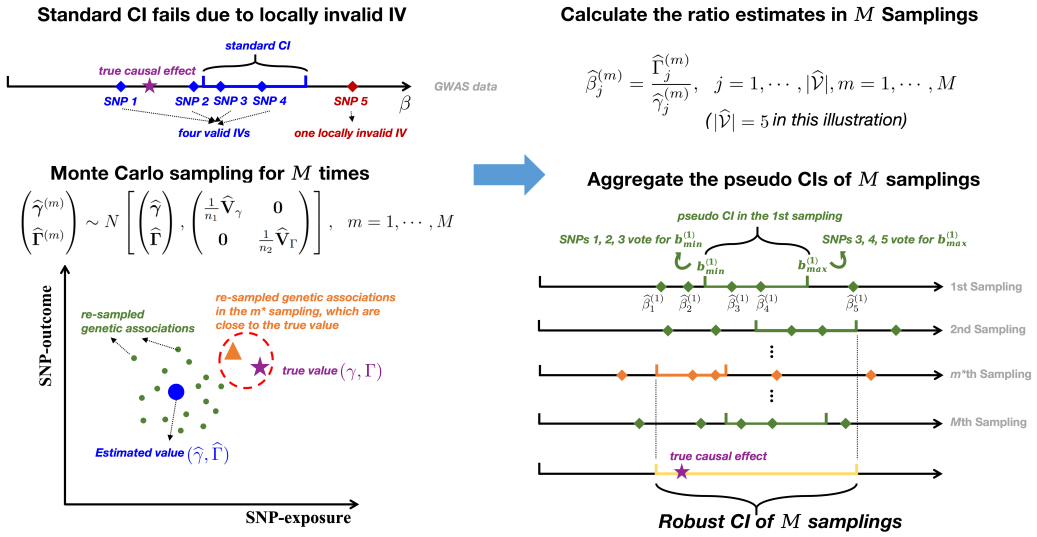

Figure 5: The proposed procedure for constructing the robust confidence interval by MR-SPI that allows for finite-sample IV selection error. When Llocally invalid IVs may be exist in finite samples, MR-SPI might incorrectly selected invalid IVs as valid ones (marked by the red cross) in finite-sample settings, and thus the standard CI might fail to cover the true causal effect. First, we construct an initial interval using SNPs in and discretize it to a grid set. SecondTo deal with this issue, we repeatedly sample the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ for M times (by default, we set $M = 1{,}000$) from the sampling distribution. When M is sufficiently large, there exists m^*th sampling such that the re-sampled genetic associations (marked by orange triangle) are close enough to the true values $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$. In each sampling, we calculate the ratio estimates using the re-sampled genetic associations, and then construct a pseudo CI for the causal effect by line searching. Specifically, for any value b in the pseudo CI, more than half of the IVs selected by MR-SPI (in this illustration, at least three IVs) should vote for b to be the true causal effect. We then aggregate all the pseudo CIs of M samplings by taking the minimum of the lower bounds and the maximum of the upper bounds to construct the robust CI (marked by yellow segment).

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] Jesse R Zaneveld, Ryan McMinds, and Rebecca Vega Thurber. Stress and stability: applying the anna karenina principle to animal microbiomes. *Nature microbiology*, 2(9):1–8, 2017.

[30] 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.

[31] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[32] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19 (6):679–682, 2022.

[33] Hannah K Wayment-Steele, Adedolapo Ojoawo, Renee Otten, Julia M Apitz, Warintra Pitsawong, Marc Hömberger, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, pages 1–3, 2023.

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

[35] 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.

[36] Zijian Guo. Causal inference with invalid instruments: post-selection problems and a solution using searching and sampling. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):959–985, 2023.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] Rebecca L Winfree, Mabel Seto, Logan Dumitrescu, Vilas Menon, Philip De Jager, Yanling Wang, Julie Schneider, David A Bennett, Angela L Jefferson, and Timothy J Hohman. Trem2 gene expression associations with alzheimer's disease neuropathology are region-specific: implications for cortical versus subcortical microglia. *Acta Neuropathologica*, 145(6):733–747, 2023.

[49] Xiaoyu Yang, Jia Wen, Han Yang, Ian R Jones, Xiaodong Zhu, Weifang Liu, Bingkun Li, Claire D Clelland, Wenjie Luo, Man Ying Wong, et al. Functional characterization of alzheimer's disease genetic variants in microglia. *Nature Genetics*, pages 1–10, 2023.

[50] Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023.

[51] Ying Zhou, Yintao Zhang, Xichen Lian, Fengcheng Li, Chaoxin Wang, Feng Zhu, Yunqing Qiu, and Yuzong Chen. Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. *Nucleic Acids Research*, 50(D1):D1398–D1407, 2022.

[52] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

[53] Peter F Bross, Julie Beitz, Gang Chen, Xiao Hong Chen, Eric Duffy, Lydia Kieffer, Sandip Roy, Rajeshwari Sridhara, Atiqur Rahman, Grant Williams, et al. Approval summary: gemtuzumab ozogamicin in relapsed acute myeloid leukemia. *Clinical cancer research*, 7(6):1490–1496, 2001.

[54] Kelly J Norsworthy, Chia-Wen Ko, Jee Eun Lee, Jiang Liu, Christy S John, Donna Przepiorka, Ann T Farrell, and Richard Pazdur. Fda approval summary: mylotarg for treatment of patients with relapsed or refractory cd33-positive acute myeloid leukemia. *The oncologist*, 23 (9):1103–1108, 2018.

[55] Janice Kim, Diana Bradford, Erin Larkins, Lee H Pai-Scherf, Somak Chatterjee, Pallavi S Mishra-Kalyani, Emily Wearne, Whitney S Helms, Amal Ayyoub, Youwei Bi, et al. Fda

approval summary: pralsetinib for the treatment of lung and thyroid cancers with ret gene mutations or fusions. *Clinical Cancer Research*, 27(20):5452–5456, 2021.

[56] Diana Bradford, Erin Larkins, Sirisha L Mushti, Lisa Rodriguez, Amy M Skinner, Whitney S Helms, Lauren SL Price, Jeanne Fourie Zirkelbach, Yangbing Li, Jiang Liu, et al. Fda approval summary: selpercatinib for the treatment of lung and thyroid cancers with ret gene mutations or fusions. *Clinical Cancer Research*, 27(8):2130–2135, 2021.

[57] Ana Griciuc, Alberto Serrano-Pozo, Antonio R Parrado, Andrea N Lesinski, Caroline N Asselin, Kristina Mullin, Basavaraj Hooli, Se Hoon Choi, Bradley T Hyman, and Rudolph E Tanzi. Alzheimer's disease risk gene cd33 inhibits microglial uptake of amyloid beta. *Neuron*, 78(4):631–643, 2013.

[58] Fabiana Tortora, Antonella Rendina, Antonella Angiolillo, Alfonso Di Costanzo, Francesco Aniello, Aldo Donizetti, Ferdinando Febbraio, and Emilia Vitale. Cd33 rs2455069 snp: correlation with alzheimer's disease and hypothesis of functional role. *International Journal of Molecular Sciences*, 23(7):3629, 2022.

[59] 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.

[60] 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.

[61] 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.

[62] Anne Rijpma, Marinette van der Graaf, Olga Meulenbroek, Marcel GM Olde Rikkert, and Arend Heerschap. Altered brain high-energy phosphate metabolism in mild alzheimer's disease: A 3-dimensional 31p mr spectroscopic imaging study. *NeuroImage: Clinical*, 18:254–261, 2018.

[63] Prodromos Parasoglou, Ricardo S Osorio, Oleksandr Khegai, Zanetta Kovbasyuk, Margo Miller, Amanda Ho, Seena Dehkharghani, Thomas Wisniewski, Antonio Convit, Lisa Mosconi, et al. Phosphorus metabolism in the brain of cognitively normal midlife individuals at risk for alzheimer's disease. *Neuroimage: Reports*, 2(4):100121, 2022.

[64] Maciej J Lazarczyk, Julia E Kemmler, Brett A Eyford, Jennifer A Short, Merina Varghese, Allison Sowa, Daniel R Dickstein, Frank J Yuk, Rishi Puri, Kaan E Biron, et al. Major histocompatibility complex class i proteins are critical for maintaining neuronal structural complexity in the aging brain. *Scientific reports*, 6(1):26199, 2016.

[65] Min-Seok Kim, Kwangmin Cho, Mi-Hyang Cho, Na-Young Kim, Kyunggon Kim, Dong-Hou Kim, and Seung-Yong Yoon. Neuronal mhc-i complex is destabilized by amyloid-β and its implications in alzheimer's disease. *Cell & Bioscience*, 13(1):181, 2023.

[66] Yann Le Guen, Guo Luo, Aditya Ambati, Vincent Damotte, Iris Jansen, Eric Yu, Aude Nicolas, Itziar de Rojas, Thiago Peixoto Leal, Akinori Miyashita, et al. Multiancestry analysis of the hla locus in alzheimer's and parkinson's diseases uncovers a shared adaptive immune response mediated by hla-drb1* 04 subtypes. *Proceedings of the National Academy of Sciences*, 120(36):e2302720120, 2023.

[67] 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.

[68] 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.

[69] 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.

[70] Stephen Burgess, Frank Dudbridge, and Simon G Thompson. Combining information on multiple instrumental variables in mendelian randomization: comparison of allele score and summarized data methods. *Statistics in medicine*, 35(11):1880–1906, 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[67]. We also assume that the SNPs are standardized[68] 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)[24,69]. 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 $|\mathcal{V}| > \frac{1}{2}|\mathcal{S}|$[69], 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 the covariance matrices $\frac{1}{n_1}\mathbf{V}_\gamma$ 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 then 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:

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

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[34]. 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 three 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[34].

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

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

where $\widehat{\boldsymbol{\gamma}}_{\widehat{\mathcal{V}}}$ and $\widehat{\boldsymbol{\Gamma}}_{\widehat{\mathcal{V}}}$ are the estimates of SNP-exposure associations and SNP-outcome associations of the selected valid IVs in $\widehat{\mathcal{V}}$, respectively. The MR-SPI estimator in equation (7) is the regression coefficient obtained by fitting a zero-intercept ordinary least squares regression of $\widehat{\boldsymbol{\Gamma}}_{\widehat{\mathcal{V}}}$ on $\widehat{\boldsymbol{\gamma}}_{\widehat{\mathcal{V}}}$. Since the SNPs are standardized, the genetic associations $\widehat{\gamma}_j$ and $\widehat{\Gamma}_j$ are scaled by $\sqrt{2f_j(1-f_j)}$ (compared to the genetic associations calculated using the unstandardized SNPs, denoted by $\check{\gamma}_j$ and $\check{\Gamma}_j$), where f_j is the minor allele frequency of SNP j. As $f_j(1-f_j)$ is approximately proportional to the inverse variance of $\check{\Gamma}_j$ when each SNP IV explains only a small proportion of variance in the outcome[70], the MR-SPI estimator of the causal effect in equation (7) is approximately equal to the inverse-variance weighted estimator[19] calculated with $\{\check{\gamma}_j, \check{\Gamma}_j\}_{j\in\widehat{\mathcal{V}}}$.

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: Selecting Valid Instruments and Performing Inference of the Causal Effect 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 $\widehat{\beta}_{\text{SPI}}$ 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 $|\widehat{\pi}_k^{[j]}|/\widehat{\text{SE}}(\widehat{\pi}_k^{[j]}) \le \sqrt{\log \min(n_1, n_2)}$;

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" [36]. 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 S13. Supplementary Figure S13(a) shows an example in which MR-SPI provides robust inference across different values of the scaling factor, while Supplementary Figure S13(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[36], with main steps described in Figure 5. The key idea is to sample the estimators of $\boldsymbol{\gamma}$ and $\boldsymbol{\Gamma}$ repeatedly from the following distribution:

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

where M is the number of sampling times (by default, we set $M = 1,000$). 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 values $\boldsymbol{\gamma}$ 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 $|b_{k+1} - b_k| = n_{\min}^{-0.6}$ for $1 \le k \le K-2$, where $n_{\min} = \min(n_1, n_2)$. We set the grid size $n_{\min}^{-0.6}$ so that the error caused by discretization is smaller than the parametric rate $n_{\min}^{-1/2}$.

For each grid value $b \in \mathcal{B}$ and sampling index $1 \le m \le M$, we propose an estimate of π_j by $\widehat{\pi}_j^{(m)}(b) = \left(\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)} \right) \cdot \mathbf{1}\left(|\widehat{\Gamma}_j^{(m)} - b\widehat{\gamma}_j^{(m)}| \ge \lambda \widehat{\rho}_j(b, \alpha) \right)$ 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 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 associations and b as the true causal effect. Let $\widehat{\boldsymbol{\pi}}^{(m)}_{\widehat{\mathcal{V}}}(b) = (\widehat{\pi}^{(m)}_j(b))_{j\in\widehat{\mathcal{V}}}$, then we construct the mth sampling's pseudo confidence interval $\text{pCI}^{(m)}$ 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 $\beta^{(m)}_{\min} = \min\{b \in \mathcal{B} : \|\widehat{\boldsymbol{\pi}}^{(m)}_{\widehat{\mathcal{V}}}(b)\|_0 < |\widehat{\mathcal{V}}|/2\}$ and $\beta^{(m)}_{\max} = \max\{b \in \mathcal{B} : \|\widehat{\boldsymbol{\pi}}^{(m)}_{\widehat{\mathcal{V}}}(b)\|_0 < |\widehat{\mathcal{V}}|/2\}$, then the mth sampling's pseudo confidence interval is constructed as $\text{pCI}^{(m)} = \left(\beta^{(m)}_{\min}, \beta^{(m)}_{\max}\right)$.

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

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

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 $\text{CI}^{\text{robust}}$.

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

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

end

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

Simulation settings

We set the number of candidate IVs $p = 10$, as the average number of candidate SNP IVs for the plasma proteins in the UK Biobank proteomics data is around 7.4. We set the sample sizes $n_1 = n_2 \in \{5{,}000, 10{,}000, 20{,}000, 40{,}000, 80{,}000\}$. 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 1,000 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.