Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Winner's Curse Free Robust Mendelian Randomization with Summary Data

Zhongming Xie, Wanheng Zhang, Jingshen Wang & Chong Wu

To cite this article: Zhongming Xie, Wanheng Zhang, Jingshen Wang & Chong Wu (14 Nov 2025): Winner's Curse Free Robust Mendelian Randomization with Summary Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2587321

To link to this article: https://doi.org/10.1080/01621459.2025.2587321

Winner's Curse Free Robust Mendelian Randomization with Summary Data

Zhongming Xie[a], Wanheng Zhang[b], Jingshen Wang[c,*], Chong Wu[d,*]

[a]Division of Biostatistics, University of California Berkeley

[b]Department of Biostatistics, The University of Texas MD Anderson Cancer Center

[c]Division of Biostatistics, University of California Berkeley. Corresponding author

[d]Department of Biostatistics, The University of Texas MD Anderson Cancer Center. Corresponding author.

*Jingshen Wang, jingshenwang@berkeley.edu; Chong Wu, cwu18@mdanderson.org

Abstract

In the past decade, the increased availability of genome-wide association studies summary data has popularized Mendelian Randomization (MR) for conducting causal inference. MR analyses, incorporating genetic variants as instrumental variables, are known for their robustness against reverse causation bias and unmeasured confounders. Nevertheless, classical MR analyses utilizing summary data may still produce biased causal effect estimates due to the winner's curse and pleiotropy issues. To address these two issues and establish valid causal conclusions, we propose a unified robust Mendelian Randomization framework with summary data, which systematically removes the winner's curse and screens out invalid genetic instruments with pleiotropic effects. Unlike existing robust MR literature, our framework delivers valid statistical inference on the causal effect without requiring the genetic pleiotropy effects to follow any parametric distribution or relying on perfect instrument screening property. Under appropriate conditions, we demonstrate that our proposed estimator converges to a normal distribution, and its variance can be well estimated. We demonstrate the performance of our proposed estimator through Monte Carlo simulations and two case studies.

Keywords: Bootstrap aggregation; GWAS; Post-selection inference.

1 Introduction

1.1 Background and motivation

Drawing inferences about cause and effect lies at the core of uncovering essential scientific principles. In biological and biomedical sciences, causal inference deepens our understanding of underlying etiology and advances developments in disease diagnosis, treatment, and prevention. While observational data present unique opportunities for causal inference by employing large and rich datasets, causal discoveries from observational studies are often susceptible to unmeasured confounding and reverse causation bias issues (Imai et al., 2011; Flegal et al., 2011; Gelman and Imbens, 2013; Smith and Ebrahim, 2004). As a remedy, Mendelian Randomization (MR) has become a popular research design. Its popularity is not only ascribed to the fact that MR mitigates unmeasured confounding bias by using genetic variants as instrumental variables (IVs) to assess the causal relationship between exposures and outcomes but also credited to the increasing availability of large-scale genome-wide association studies (GWAS) summary data on various complex traits (Smith and Ebrahim, 2004; Didelez and Sheehan, 2007; Lawlor et al., 2008; Skrivankova et al., 2021).

However, MR with GWAS summary may still produce biased estimates of causal effects due to several sources of bias. These include measurement error in exposure GWAS, winner's curse bias resulting from using the same exposure GWAS for both IV selection and effect estimation, and most crucially, bias from including invalid IVs with pleiotropy (Sadreev et al., 2021). Firstly, the effect of IV on exposure is measured by exposure GWAS, which inherently contains measurement error. Ignoring such measurement error can produce biased causal effect estimates, especially when the strength of IVs is weak (Ye et al., 2021; Ma et al., 2023). Secondly, the practice of selecting genetic instruments based on their estimated associations with the exposure variable from GWAS, and using the same data for both instrument selection and estimation, can lead to biased causal effect estimates due to the winner's curse phenomenon (Zöllner and Pritchard, 2007; Zhong and Prentice, 2010; Gkatzionis and Burgess, 2019). Lastly, typical MR analyses inevitably involve some invalid IVs that either directly affect the outcome or through unmeasured confounding factors—a phenomenon known as pleiotropy (Hemani et al., 2018; Watanabe et al., 2019). The nature of pleiotropy is widespread and usually unknown or complex (Watanabe et al., 2019). Failure to fully account for pleiotropy will also lead to biased causal effect estimates.

A broad literature addresses the biases discussed above to improve the credibility of MR analyses, yet no single approach can simultaneously tackle all these biases. Some methods have made progress in addressing individual issues. For instance, Ye et al. (2021) formally tackled the measurement error bias in the popular inverse variance weighted estimator, while Ma et al. (2023) proposed a randomized instrument selection and Rao-Blackwellization procedure to address both measurement error bias and winner's curse bias. However, the validity of these methods relies heavily on the assumption that all IVs either have no pleiotropic effects or exhibit balanced pleiotropic effects—an assumption unlikely to hold in practice due to the unknown and complex nature of pleiotropy (Watanabe et al., 2019), potentially leading to biased causal effect estimates.

To account for widespread pleiotropy, many robust MR methods have been proposed. These methods primarily focus on addressing the issue raised by invalid IVs, but often at the expense of neglecting measurement error and winner's curse biases. They can be broadly categorized into two strategies. The first strategy imposes normal mixture model assumptions on the pleiotropic effects. By modeling the observed GWAS summary data within a joint likelihood function, these methods simultaneously estimate the unknown parameters and the desired causal effect. Such methods include RAPS (Zhao et al., 2020), ContMix (Burgess et al., 2020), MR-APSS (Hu et al., 2022), MRMix (Qi and Chatterjee, 2019). However, as demonstrated in our simulation studies, when the normal mixture model assumption is violated, these approaches tend to produce false positive findings or have low detection power. Moreover, incorporating procedures to address winner's curse bias, such as that proposed by Ma et al. (2023), is challenging within this framework as it may violate parametric modeling assumptions and result in an incorrect likelihood function. The second strategy avoids imposing parametric modeling assumptions on the pleiotropic effects. Instead, it adopts penalization methods to screen out invalid instruments with pleiotropic effects, using only the selected valid instruments for causal effect estimation. Such methods include, for example, cML (Xue et al., 2021) and MR-Lasso (Luo et al., 2008). However, these methods either lack rigorous statistical justifications or require that the selected IVs are valid and include all valid IVs (a condition we refer to as "perfect IV screening"). For example, Xue et al. (2021) prove that their procedure can screen out all invalid IVs with a probability tending to one under the asymptotic regime where the number of IVs is fixed, and the sample size tends to infinity. When this is achieved, the resulting causal effect estimate is consistent and asymptotically normal. However, the theoretical results under this asymptotic regime do not account for how the magnitudes of the pleiotropic effects impact the validity of statistical inference. In fact, perfect IV screening is often unattainable when the pleiotropic effects are small, and the differences between valid and invalid IVs in MR studies are subtle. Notably, two-sample MR is a rapidly evolving field with numerous methodological advancements, such as (Morrison et al., 2020; Liu et al., 2023; Grant and Burgess, 2024). For comprehensive reviews of statistical methods in MR, we refer readers to Sanderson et al. (2022) and Boehm and Zhou (2022).

1.2 Contribution

To bridge the aforementioned gaps in the existing literature, we propose a unified MR framework with summary data that simultaneously addresses winner's curse bias, bias from measurement error in exposure GWAS, and bias from invalid IVs with pleiotropy (Section 3). Specifically, we propose an l_0 constrained optimization framework that can simultaneously screen out invalid IVs, account for measurement error, and seamlessly integrate with the winner's removal step from Ma et al. (2023). Moreover, we demonstrate that the proposed l_0 constrained optimization framework maintains computational efficiency due to the special form of our objective function. Furthermore, to improve statistical efficiency, we adopt a bootstrap aggregation procedure and use a non-parametric delta method to perform valid inference on the final causal effect.

On the theoretical side, we provide comprehensive theoretical investigations of the proposed method in Section 4. We prove that the final estimator in our proposed method is asymptotically unbiased and converges to a normal distribution even in the presence of directional pleiotropy. Moreover, different from existing theoretical analyses in robust MR, we show that our method can deliver consistent causal effect estimates without perfect

invalid IV screening; see detailed discussion in Supplementary Material Section S.6. In brief, our theoretical investigation indicates that our proposed method can screen out IVs with large pleiotropic effects, and the resulting causal effect estimator remains consistent even if the selected IVs include some invalid ones with small pleiotropic effects. These theoretical investigations better characterize scenarios where our method performs well and demonstrate its robustness.

Benefiting from the above features in both methodological and theoretical aspects, we demonstrate that our proposed MR framework delivers robust causal effect estimates with improved statistical power in simulated Monte Carlo experiments (Section 5) and in two case studies (Section 6). From our simulated Monte Carlo experiments, we confirm that our proposed method outperforms benchmark methods in terms of type 1 error rates, power, absolute bias, mean squared error, and coverage probability in most scenarios. The results also highlight the importance of simultaneously correcting the winner's curse bias and accounting for measurement error bias and generic pleiotropic effects. From our case study of negative control outcome analyses, in which the population causal effects are believed to be zero by design, we confirm that our approach yields well-controlled Type I error rates (Section 6.1). From our case study to identify causal risk factors for COVID-19 severity, our approach identifies more causal risk factors than the existing approaches, and the identified causal exposures by our proposed method have more supporting evidence.

2 Framework and challenges

In this section, we review the classical two-sample Mendelian Randomization (MR) framework with summary data. We then revisit the pleiotropic effects, measurement error bias, and winner's curse bias within this framework.

Referring to the causal diagram in Figure 1, we let X denote the exposure, Y the outcome, and U the unmeasured confounder between the exposure and the outcome. The goal of MR analysis is to estimate the causal effect (denoted by θ) of the exposure variable X on the outcome variable Y. However, in the presence of unmeasured confounder U, it is challenging to directly estimate θ solely using the information stored in X and Y. To overcome this, two-sample MR analyses incorporate p mutually independent SNPs $G_1,\ldots,G_p$ as instrumental variables (IVs) and estimate θ using the estimated association pairs $\{(\hat{\beta}_{X_j},\hat{\beta}_{Y_j})\}_{j=1}^{p}$ collected from two independent GWAS datasets, where $\hat{\beta}_{X_j}$ and $\hat{\beta}_{Y_j}$ are the estimated effect sizes for IV j in exposure and outcome GWAS, respectively. Here, genetic variant $G_j \in \{0,1,2\}$ represents the number of effect alleles of a single-nucleotide polymorphism (SNP) j inherited by an individual. Following the two-sample summary-data MR literature Ye et al. (2021); Zhao et al. (2020), we assume the following linear structural equation model:

$$
\begin{aligned}
U &= \sum_{j=1}^{p} \phi_j G_j + E_U, \\
X &= \sum_{j=1}^{p} \gamma_j G_j + \beta_{XU} U + E_X, \qquad (1) \\
Y &= \sum_{j=1}^{p} \alpha_j G_j + \beta_{YU} U + \theta X + E_Y,
\end{aligned}
$$

where E_U, E_X, and E_Y are mutually independent random noises. E_U is independent of $(G_1,\ldots,G_p)$, and E_X and E_Y are independent of $(G_1,\ldots,G_p,U)$. To allow for the valid inference of the causal effect θ, we need G_j ($j=1,\ldots,p$) to be valid IVs in the sense that they satisfy the following three conditions: (1) $\gamma_j \neq 0$, meaning that G_j is associated with *X* (relevance assumption); (2) $\phi_j = 0$, meaning that G_j has no correlated pleiotropic effect with *Y* (effective random assignment assumption); (3) $\alpha_j = 0$, meaning that G_j has no uncorrelated pleiotropic effect with *Y* (exclusion restriction assumption).

Provided that all included genetic IVs are valid, two-sample MR analyses can deliver valid inference on θ by appropriately using information stored in two independent GWAS datasets. To provide some justifications for this claim, we follow the causal model proposed in Pearl (2009). In particular, in the structural equation models given in Eq (1), the total effect of SNP G_j on *Y* and the total effect of G_j on *X* are given by:

$$\beta_{Y_j} = \mathbb{E}[Y \mid do(G_j = g_j + 1)] - \mathbb{E}[Y \mid do(G_j = g_j)] = \alpha_j + \beta_{YU}\phi_j + \theta\cdot(\gamma_j + \beta_{XU}\phi_j),$$
$$\beta_{X_j} = \mathbb{E}[X \mid do(G_j = g_j + 1)] - \mathbb{E}[X \mid do(G_j = g_j)] = \gamma_j + \beta_{XU}\phi_j.$$

For a valid IV G_j, when G_j satisfies $\phi_j = 0$ (effective random assignment assumption) and $\alpha_j = 0$ (exclusion restriction assumption), the target causal effect θ will satisfy $\beta_{Y_j} = \theta\beta_{X_j}$, where $\beta_{X_j} = \gamma_j$ and $\beta_{Y_j} = \theta\gamma_j$. If the relevance assumption $\gamma_j \neq 0$ is also met, we are then able to use β_{Y_j} and β_{X_j} to assist valid inference on θ, as they can be well estimated through the estimated association pairs $\{(\hat{\beta}_{X_j}, \hat{\beta}_{Y_j})\}_{j=1}^{p}$ collected from two independent GWAS dataset in two-sample summary-data MR framework.

However, in practice, due to the widespread pleiotropy in human genetics (Hemani et al., 2018; Watanabe et al., 2019), the effective random assignment ($\phi_j = 0$) and exclusion restriction assumptions ($\alpha_j = 0$) are frequently violated, leading to invalid IVs. In the presence of invalid IVs, the total effect of G_j on *Y* can be expressed as:

$$\beta_{Y_j} = \underbrace{\theta\cdot\beta_{X_j}}_{\substack{\text{causal}\\ \text{effect}}} + \underbrace{\alpha_j}_{\substack{\text{uncorrelated}\\ \text{pleiotropy}}} + \underbrace{\beta_{YU}\cdot\phi_j}_{\substack{\text{correlated}\\ \text{pleiotropy}}} \equiv \theta\cdot\beta_{X_j} + r_j. \quad (2)$$

Here, α_j is the uncorrelated pleiotropic effect that captures the direct effect of G_j on *Y*, and $\beta_{YU}\cdot\phi_j$ is the correlated pleiotropic effect that captures the effect of G_j on *Y* through the pathway $G_j \to U \to Y$. Their combined effect, $r_j = \alpha_j + \beta_{YU}\cdot\phi_j$, represents the total effect of a genetic variant G_j on the outcome *Y* induced by pleiotropy. These violations make it challenging to accurately estimate θ using MR. If not appropriately accounted for, genetic pleiotropy can result in biased causal effect estimates in MR analyses (see Section 5 for our simulation results).

On top of the potential bias induced by pleiotropic effects, two additional sources of bias in MR analyses are measurement error bias and winner's curse bias. Measurement error bias arises from the fact that the true effect of an IV on the exposure, β_{X_j}, is unobserved. Instead, we rely on $\hat{\beta}_{X_j}$, an estimate derived from exposure GWAS (I), which inherently contains measurement error, to conduct MR. The winner's curse bias, on the other hand, is induced by pre-selecting IVs that are strongly associated with the exposure variable to meet the relevance assumption (that is, $\gamma_j \neq 0$). This selection exercise is often based on hard-thresholding measured SNP *z*-scores obtained from GWAS (I): SNP *j* is selected if $|\hat{\beta}_{X_j} / \sigma_{X_j}| > \lambda$, where λ is a pre-specified cut-off value, and $\hat{\beta}_{X_j}$ and σ_{X_j} are estimated effect size and its standard error from exposure GWAS dataset, respectively. The selected IVs are then used to construct downstream causal effect estimators. The selected IV-exposure associations tend to overestimate the underlying true association effects β_{X_j}, as the distribution of any $\hat{\beta}_{X_j}$ that survives the selection is a truncated Gaussian and the post-selection mean is no longer β_{X_j} when commonly used Gaussian assumption on $\hat{\beta}_{X_j}$ is adopted. Subsequently, by doubly using the data in GWAS (I) for IV selection and estimation, classical MR estimators are expected to be biased and have an intractable limiting distribution, making statistical inference problematic.

In the rest of this manuscript, we employ the following model frequently adopted in the Mendelian Randomization literature (Zhao et al., 2020; Qi and Chatterjee, 2019; Xue et al., 2021):

Assumption 1 (Measurement error model)

(i)For any $j \neq j'$, $(\hat{\beta}_{Y_j}, \hat{\beta}_{X_j})$ *and* $(\hat{\beta}_{Y_{j'}}, \hat{\beta}_{X_{j'}})$ *are mutually independent. (ii)For each j, the association pair* $(\hat{\beta}_{Y_j}, \hat{\beta}_{X_j})$ *follows*

$$\begin{bmatrix} \hat{\beta}_{X_j} \\ \hat{\beta}_{Y_j} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta_{X_j} \\ \theta\beta_{X_j} + r_j \end{bmatrix}, \begin{bmatrix} \sigma_{X_j}^2 & 0 \\ 0 & \sigma_{Y_j}^2 \end{bmatrix} \right).$$

Furthermore, there exists a positive integer $n \to \infty$ *and positive constants m and M such that* $\frac{m}{n} \leq \sigma_{X_j}^2 \leq \frac{M}{n}$, $\frac{m}{n} \leq \sigma_{Y_j}^2 \leq \frac{M}{n}$ *for* $j = 1, \ldots, p$.

The assumption of independent SNPs, while seemingly stringent, is grounded in established practice in two-sample MR analyses (Ye et al., 2021; Zhao et al., 2020; Ma et al., 2023). This approach helps ensure that each selected SNP represents a signal from a unique genetic locus, thereby mitigating potential confounding effects from LD and facilitating clearer interpretation of causal effect estimates. We acknowledge that alternative cis-MR methods such as Transcriptome-Wide Association Studies (TWAS) (Gusev et al., 2016; Wainberg et al., 2019) and Proteome-Wide Association Studies (PWAS), effectively utilize correlated SNPs, particularly for investigating relationship between omics and complex traits. However, as the reviewer suggested, when inferring causal relationships between complex

traits/diseases (such as the two case studies in Section 6), using independent IVs from the whole genome is typically efficient enough and simple to implement. This strategy is also widely adopted in the literature. Therefore, in line with this common practice, we adopt the independence assumption. To ensure independent IVs, we apply a sigma-based LD pruning method (Ma et al., 2023).

3 Methodology

3.1 Measurement error correction and invalid IV screening

To estimate the causal effect θ, a straightforward approach is to replace the population association effects with their empirical estimates from GWAS in the causal structure equation in (2). Given that all population associations are measured with error in GWAS, the sample analogue of the structure equations can be represented as the following two-stage regression model with measurement errors:

$$\underset{\text{response}}{\hat{\beta}_{Y_j}} = \underset{\substack{\text{target}\\ \text{parameter}}}{\theta} \cdot \underset{\substack{\text{true}\\ \text{covariate}}}{\beta_{X_j}} + \underset{\substack{\text{unknown}\\ \text{parameter}}}{r_j} + \underset{\text{noise}}{\nu_j}, \qquad \underbrace{\hat{\beta}_{X_j} = \beta_{X_j} + u_j}_{\substack{\text{covariates are}\\ \text{measured with error}}},$$

where ν_j and u_j are centered noises.

To operationalize an accurate estimate of θ using the above two-stage least squares model, we first consider a situation where a set of IVs with $\beta_{X_j} \neq 0$ (denoted as $\mathcal{S}$) is known. Our method does not require $\mathcal{S}$ to be known, and we will discuss the selection of $\mathcal{S}$ and the practical implementation of our algorithm in the next subsection. With a known $\mathcal{S}$, we propose estimating θ by solving the following constrained optimization problem:

$$\begin{aligned} \min_{\theta, r_j} \quad & l(\theta, \{r_j\}_{j\in\mathcal{S}}) = \sum_{j\in\mathcal{S}} l_j(\theta, r_j) \triangleq \sum_{j\in\mathcal{S}} \frac{(\hat{\beta}_{Y_j} - \theta\cdot\hat{\beta}_{X_j} - r_j)^2}{\sigma_{Y_j}^2} - \sum_{j\in\mathcal{S}} \frac{\theta^2\cdot\sigma_{X_j}^2}{\sigma_{Y_j}^2} 1_{(r_j=0)}, \\ s.t. \quad & \sum_{j\in\mathcal{S}} 1_{(r_j=0)} = v. \end{aligned} \tag{3}$$

Intuitively, the objective function above is a bias-corrected least squares function designed to account for measurement error, subject to the constraint that the adopted IVs for estimating θ are valid. In the following, we will show that the optimization problem above not only accounts for the measurement errors in $\hat{\beta}_{X_j}$ but also accurately identifies invalid IVs with $r_j \neq 0$. This is achieved with computational efficiency, even when an l_0-type constraint is adopted. As a result, the solution of this optimization problem provides an accurate estimate of θ.

To start with, when the set of IVs with $r_j = 0$ is known, the solution of the above optimization problem provides an unbiased estimate of θ. As in this case, we have

$$L(\theta) \triangleq \min_{r_j} l(\theta, \{r_j\}_{j\in\mathcal{S}}) = \frac{1}{2}\sum_{j\in\mathcal{V}} \frac{(\hat{\beta}_{Y_j} - \theta\cdot\hat{\beta}_{X_j})^2}{\sigma^2_{Y_j}} - \frac{1}{2}\sum_{j\in\mathcal{V}} \frac{\theta^2\cdot\sigma^2_{X_j}}{\sigma^2_{Y_j}}.$$

We can verify that $L(\theta)$ is unbiased for the weighted least squares loss function in the sense that $\mathbb{E}[L(\theta)] = \mathbb{E}[\sum_{j\in\mathcal{V}} (\hat{\beta}_{Y_j} - \theta\cdot\beta_{X_j})^2 / (2\sigma^2_{Y_j})]$. This suggests that its minimizer is unbiased for the causal effect θ.

Next, as the set of IVs with $r_j = 0$ is unknown, Problem (3) incorporates an l_0-type constraint to screen out invalid IVs. While classical l_0-type optimization problems are solved by their convex relaxations, this technique does not apply to our problem due to the inclusion of a measurement error bias correction term in our objective function (that is, the term $\sum_{j\in\mathcal{S}_\lambda} \theta^2\cdot\sigma^2_{X_j} / \sigma^2_{Y_j} 1_{(r_j=0)}$). To address this issue, we propose an iterative algorithm that mimics block coordinate descent and guarantees the decay of our objective function in Algorithm 1; see justification in the Supplementary Material Section S.1.

Lastly, the number of valid IVs *v* is unknown and requires tuning. To choose the final set of valid IVs, we propose a generalized Bayesian Information Criteria (GBIC), that is:

$$\text{GBIC}(v) = -2\hat{l}(\hat{\theta}(v), \{\hat{r}_j(v)\}_{j\in\mathcal{V}}) + \kappa_n\cdot(s - v), \quad s = |\mathcal{S}|,$$

where $\kappa_n = \log(n)$, and choose the final set of valid IVs by minimizing the GBIC. The proposed GBIC with $\kappa_n = \log(n)$ is different from the classical BIC criteria that adopts $\kappa_n = \log(s_\lambda)$. The reason for this choice is that the classical model selection consistency result of the BIC is established in the asymptotic regime with fixed s_λ. As we are in an asymptotic regime with $s_\lambda \to \infty$, our proposed GBIC criteria adjusts κ_n accordingly to ensure invalid IV screening consistency. In particular, in Section S.6 of the Supplemental Material, we demonstrate that our procedure provides a consistent causal effect estimator without requiring the perfect IV screening property under a simplified scenario and Conditions 1-2 and 8-9. One of these conditions imposes a constraint on the penalization coefficient κ_n: $\kappa_n \gg \log(s_\lambda)$. We argue that $\kappa_n = \log(n)$ is a feasible choice to satisfy this condition, as the order of the sample size is typically larger than the order of the number of selected relevant IVs in a two-sample MR study.

3.2 Unknown $\mathcal{S}$ and practical implementation

We now consider the realistic scenario where the set $\mathcal{S}$ is unknown. Because the collection of relevant IVs is not known, practitioners typically perform a pre-selection procedure to identify IVs strongly associated with the exposure. These selected IVs are then used to estimate the causal effect. As discussed in Section 2, selecting genetic instruments based on their estimated associations with the exposure variable from GWAS and using the same data for both instrument selection and estimation can lead to biased causal effect estimates due to the winner's curse phenomenon. To address the issue of winner's curse bias when $\mathcal{S}$ is

unknown, we integrate the proposed method from the previous section with the approach described in Ma et al. (2023) to perform Rao-Blackwellized randomized instrument selection.

For each SNP $j = 1, 2, \ldots, p$, we generate a pseudo SNP-exposure association effect $Z_j \sim \mathcal{N}(0, \eta^2)$, and select SNP j if $|\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j| > \lambda.$ Define the set of selected SNPs as $\mathcal{S}_\lambda = \{j : |\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j| > \lambda,\ j = 1, 2, \ldots, p\}$ and its cardinality $|S_\lambda| = s_\lambda$. For each selected SNP $j \in \mathcal{S}_\lambda$, we construct an unbiased estimator of β_{X_j} as

$$\hat{\beta}_{X_j,\mathrm{RB}} = \hat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta} \frac{\phi(A_{j,+}) - \phi(A_{j,-})}{1 - \Phi(A_{j,+}) + \Phi(A_{j,-})}, \text{ where } A_{j,\pm} = -\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta} \pm \frac{\lambda}{\eta},$$

Algorithm 1: Algorithm to solve the optimization problem in (4)

Input: Data inputs and initial parameters

Output: Estimated parameters $\hat{\theta}$ and $\hat{r}_j$

Initialization Set $k = 0$, generate $\theta^{(0)} \sim \mathrm{Uniform}\left(\min_{1 \le j \le s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}}, \max_{1 \le j \le s_\lambda} \frac{\hat{\beta}_{Y_j}}{\hat{\beta}_{X_j}} \right)$;

Block Coordinate Descent

repeat

Fix $\theta^{(k)}$, update $r_j^{(k+1)}$;

Order $\frac{\left(\hat{\beta}_{Y_j} - \theta^{(k)} \cdot \hat{\beta}_{X_j,\mathrm{RB}}\right)^2}{\sigma_{Y_j}^2} - \frac{\theta^2 \cdot \hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma_{Y_j}^2}$, $j = 1, 2, \ldots, s_\lambda - v$ in decreasing order;

Set $r_j^{(k+1)} = \hat{\beta}_{Y_j} - \theta^{(k)} \hat{\beta}_{X_j,\mathrm{RB}}$ for the largest $s_\lambda - v$ components, $j = 1, \ldots, s_\lambda - v$, and $r_j^{(k+1)} = 0$ for $j = s_\lambda - v + 1, \ldots, s_\lambda$;

Fix $r_j^{(k+1)}$, update $\theta^{(k)}$ by minimizing the following objective function:

$$\theta^{(k+1)} = \arg\min_{\theta \in \mathbb{R}} \sum_{j \in \mathcal{S}_\lambda} \frac{\left(\hat{\beta}_{Y_j} - \theta \cdot \hat{\beta}_{X_j,\mathrm{RB}} - r_j^{(k+1)}\right)^2 - \theta^2 \cdot \hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma_{Y_j}^2} 1_{(r_j^{(k+1)} = 0)}.$$

If $\left|\frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}}\right| < 10^{-7}$ **then** Stop and output $\hat{\theta}(v) = \theta^{(k+1)}$ and $\hat{r}_j(v) = r_j^{(k+1)}$;

else Set $k = k+1$;

until $\left|\frac{\theta^{(k+1)} - \theta^{(k)}}{\theta^{(k)}}\right| < 10^{-7}$;

end

Valid IV Selection via GBIC

for $v = 2, \ldots, s_\lambda$ **do**

Calculate

$$\mathrm{GBIC}(v) = -2\hat{l}\left(\hat{\theta}(v), \{\hat{r}_j(v)\}_{j\in\mathcal{V}}\right) + \log(n)\cdot(s_\lambda - v);$$

end for

Select $\mathcal{V}$ with the smallest $\mathrm{GBIC}(v)$;

end

$\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative distribution functions. Here, η is a pre-specified constant that reflects the noise level of the pseudo SNPs. We recommend using $\eta = 0.5$ as a default value (Ma et al., 2023). This choice balances the need for sufficient randomization to address the winner's curse bias while maintaining the stability of the selection process. The above procedure only randomizes the IV selection near the cut-off value λ, which implies that the strong IVs with large β_{X_j} are invariably selected. Here, the choice of the significance cutoff (λ) for selecting IVs presents a trade-off between including a sufficient number of informative IVs and maintaining the overall strength of the selected IV set. While lowering the cutoff may improve statistical power by incorporating more IVs with moderate effects, setting it too low can introduce weak or null IVs that potentially violate the relevance assumption and compromise the validity of the MR analysis. In our proposed method, we provide a sufficient condition to ensure the asymptotic normality of the estimator, which depends on the average strength of the selected IVs relative to the cutoff value. Specifically, we choose a cutoff of 5×10^{-5}, commonly used as a threshold for suggestive significance in GWAS, to strike a balance between including informative IVs and maintaining the validity of the selected IV set. We note that Rao-Blackwellization has also been applied in Bowden and Dudbridge (2009) to efficiently combine information from an initial GWAS and a replication study to obtain unbiased estimates of SNP effect sizes. Our approach differs as we do not require a replication study to construct an unbiased estimation for β_{X_j} (see Supplement Materials Section S5 for details). Benefiting from such randomized IV selection, $\hat{\beta}_{X_j,\mathrm{RB}}$ is free of winner's curse bias, implying that $\mathbb{E}[\hat{\beta}_{X_j,\mathrm{RB}} \mid j \in \mathcal{S}_\lambda] = \beta_{X_j}$. Therefore, our proposed bias-corrected least squares objective function and l_0 constraint optimization framework in the previous section can be applied:

$$\min_{\theta\in\mathbb{R}, r_j\in\mathbb{R}} \hat{l}(\theta, \{r_j\}_{j\in\mathcal{S}_\lambda}), \text{ s.t. } \sum_{j\in\mathcal{S}_\lambda} 1_{(r_j=0)} = v. \quad (4)$$

We also implemented two l_1-type methods and make comparison with our l_0 based method through simulations. Our results demonstrate that while both approaches maintain comparable Type I error control, absolute bias, mean squared error (MSE), and coverage probability across various scenarios, the l_0-based CARE method achieves higher statistical power. We have added relevant descriptions, methods, and results in Supplemental Material Section S.2-S.3 and Section S.8.12. where the loss function is defined as

$$\hat{l}(\theta,\{r_j\}_{j\in\mathcal{S}_\lambda}) = \sum_{j\in\mathcal{S}_\lambda} \hat{l}_j(\theta,r_j) = \sum_{j\in\mathcal{S}_\lambda} \frac{(\hat{\beta}_{Y_j} - \theta\cdot\hat{\beta}_{X_j,\mathrm{RB}} - r_j)^2}{\sigma^2_{Y_j}} - \frac{\theta^2\cdot\hat{\sigma}^2_{X_j,\mathrm{RB}}}{\sigma^2_{Y_j}} 1_{(r_j=0)},$$

$$\hat{\sigma}^2_{X_j,\mathrm{RB}} = \sigma^2_{X_j}\left(1 - \frac{1}{\eta^2}\frac{A_{j,+}\phi(A_{j,+}) - A_{j,-}\phi(A_{j,-})}{1-\Phi(A_{j,+})+\Phi(A_{j,-})} + \frac{1}{\eta^2}\left(\frac{\phi(A_{j,+})-\phi(A_{j,-})}{1-\Phi(A_{j,+})+\Phi(A_{j,-})}\right)^2\right).$$

3.3 Bootstrap aggregation and statistical inference

Since the IV screening step can be rather noisy and we do not expect to perfectly screen out all invalid IVs, we next incorporate bagging (or bootstrap aggregation) (Breiman, 1996) to reduce IV screening variability and to further improve statistical efficiency. Then, we adopt the non-parametric delta method Efron (1982) to construct a confidence interval for our bagged estimator.

To be specific, we draw bootstrap sample *B* times from $\mathcal{S}_\lambda$. For the *b*-th bootstrap sample (Denoted by $\mathcal{S}^*_{\lambda,b}$), we adjust the loss function as $\hat{l}^*_b(\theta,\{r_j\}_{j\in\mathcal{S}_\lambda}) = \sum_{j\in\mathcal{S}_\lambda} w^*_{jb}\hat{l}_j(\theta,r_j)$, where w^*_{jb} is the number of occurrences in $\mathcal{S}^*_{\lambda,b}$ for j-th IVs in $\mathcal{S}_\lambda$. Then, we conduct the invalid IV screening step for each bootstrap sample $\mathcal{S}^*_{\lambda,b}$ and select $\hat{\mathcal{V}}_b = \left\{ j : \hat{r}_{jb} = 0 \text{ and } j \in \mathcal{S}^*_{\lambda,b} \right\}$. The downstream causal estimator is derived by aggregating the estimated effects from all bootstrap samples, that is:

$$\theta_b = \frac{\sum_{j\in\mathcal{V}_b} \hat{\beta}_{Y_j}\hat{\beta}_{X_j,\mathrm{RB}}/\sigma^2_{Y_j}}{\sum_{j\in\mathcal{V}_b}(\hat{\beta}^2_{X_j,\mathrm{RB}} - \hat{\sigma}^2_{X_j,\mathrm{RB}})/\sigma^2_{Y_j}}, \quad \tilde{\theta} = \frac{1}{B}\sum_{b=1}^{B}\theta_b, \quad (5)$$

where $\hat{\theta}_b$ is obtained by refitting the loss function $\hat{l}(\theta,\{r_j\}_{j\in\mathcal{V}_b})$.

To provide valid statistical inference on the true causal effect θ, we use the non-parametric delta method (Efron, 2014) to estimate the variance of the bagged estimator with $\hat{\sigma}^2_n = \sum_{j\in\mathcal{S}_\lambda}\hat{S}^2_j$, where $\hat{S}_j = B^{-1}\sum_{b=1}^{B}(w^*_{ib} - B^{-1}\sum_{k=1}^{B} w^*_{ik})(\hat{\theta}_b - \tilde{\theta})$. Then we construct a $(1-\alpha)$-level confidence interval for θ with $[\tilde{\theta} - z_{\alpha/2}\cdot\hat{\sigma}_n, \tilde{\theta} + z_{\alpha/2}\cdot\hat{\sigma}_n]$. Here α is the upper $\alpha/2$-quantile of the standard normal distribution.

In the remainder of this manuscript, we refer to the proposed method as Causal Analysis with Randomized Estimators (CARE). The formalization of our proposed algorithm can be found

in Algorithm 2. We also provide the discussion on the time complexity of this algorithm in Section S.1 in Supplemental Material.

4 Theoretical investigations

To discuss our theoretical investigations in detail, we begin by revisiting and introducing notations and assumptions. Recall that the set of selected IVs after rerandomization is defined as $\mathcal{S}_\lambda = \{j : |\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j| > \lambda, j = 1, \ldots, p\}$ and its cardinality is denoted as $|\mathcal{S}_\lambda| = s_\lambda$. We next define κ_λ as the average of squared standardized IV effects to measure the selected IV strength in $\mathcal{S}_\lambda$, that is $\kappa_\lambda = \frac{1}{s_\lambda}\sum_{j\in\mathcal{S}_\lambda}\frac{\beta_{X_j}^2}{\sigma_{Y_j}^2}$. Among the selected IVs after rerandomization, we denote $\mathcal{V}_\lambda = \{j : j \in \mathcal{S}_\lambda \text{ and } r_j = 0\}$ as the set of valid IVs in $\mathcal{S}_\lambda$ and denote its cardinality as $|\mathcal{V}_\lambda| = v_\lambda$.

Considering the dual sources of randomness in our proposed estimator (one from the original GWAS sample, and the other from the bootstrap resampling), we separate these two sources of randomness by denoting the conditional expectation taken with respect to bootstrap resampling as $\mathbb{E}^*[\cdot] = \mathbb{E}[\cdot | S_\lambda, \{(\hat{\beta}_{Y_j}, \hat{\beta}_{X_{j,\mathrm{RB}}})\}_{j\in S_\lambda}]$. Next, we introduce three additional assumptions for our theoretical investigations:

Assumption 2 (Variance stabilization)

There exists a variance stabilizing quantity a_λ *and a vector* $\boldsymbol{\tau} \in \mathbb{R}^{s_\lambda}$ *in which each component is independent of* $\{(u_j, v_j)\}_{j\in S_\lambda}$ *and uniformly bounded away from infinity in probability in the sense that*

$$\sup_{j\in S_\lambda} |a_\lambda \cdot \mathbb{E}^*[A_b^{-1}\cdot\hat{w}_{jb}] - \tau_j| = o_p(1),$$

where $A_b = \sum_{k\in\mathcal{S}_\lambda}\hat{w}_{kb}\cdot(\hat{\beta}^2_{X_k,\mathrm{RB}} - \hat{\sigma}^2_{X_k,\mathrm{RB}})/\sigma^2_{Y_k}$, *and* $\hat{w}_{jb} = w^*_{jb}\cdot \mathrm{I}(\hat{r}_{jb} = 0)\cdot\mathrm{I}(w^*_{jb} \geq 1)$. *In addition, there is no dominating IV in the sense that* $\frac{\max_{j\in\mathcal{S}_\lambda}\beta^2_{X_j}}{\sum_{j\in\mathcal{S}_\lambda}\beta^2_{X_j}} \xrightarrow{\mathrm{p}} 0$.

The first part of the above assumption, intuitively, ensures that our estimator $\tilde{\theta}$ converges to a non-degenerative distribution asymptotically when appropriately scaled by $a_\lambda / \sqrt{s_\lambda \cdot \kappa_\lambda}$. This scaling factor accounts for the number of selected instruments and their average strength, enabling valid statistical inference. The second part of the condition requires that, after selection, no single IV exerts a "dominating effect" on exposure, which aligns with the biological understanding that complex traits are influenced by many genetic variants with small effects (i.e., the omnigenic model (Boyle et al., 2017)). To cast more insight into Assumption 2, in Section S.4.3 of the Supplemental Material, we consider a special case

where perfect IV screening is achieved. We show that in this case, Assumption 2 holds for both valid and invalid IVs in $\mathcal{S}_\lambda$.

Assumption 3 (Negligible invalid IV induced bias)

There is negligible bias induced by potential imperfect screening of invalid IVs after bootstrap aggregation in the sense that

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \mathbb{E}^*[A_b^{-1} \sum_{j \in S_\lambda} \hat{\beta}_{X_{j,\mathrm{RB}}} \cdot r_j \cdot \hat{w}_{jb} / \sigma_{Y_j}^2] = o_p(1).$$

Our theoretical investigations reveal two sets of sufficient conditions under which Assumption 3 holds (See Section S.5 and S.6 in the Supplemental Material). The first set of sufficient conditions ensures that the selected IVs are "nearly perfect," meaning they are valid but do not include all possible valid IVs. We show that this nearly perfect IV screening property can be satisfied when there is strong prior knowledge about the trait's genetic architecture or where valid and invalid IVs are easily distinguishable. The second set of sufficient conditions ensures Assumption 3 holds even if our proposed IV screening procedure does not screen all invalid IVs. In particular, our analysis indicates that when IVs with large r_j values (strong pleiotropic effects) are effectively screened out, our estimator maintains consistency even if the selected set includes some invalid IVs with small r_j values (weak pleiotropic effects). Together, these theoretical investigations suggest that perfect IV screening is not a prerequisite for valid inference in our proposed method.

Assumption 4 (Instrument Selection)

Define $\underline{\eta} = \min_{1 \le j \le p} \eta_j$ *and* $\bar{\eta} = \max_{1 \le j \le p} \eta_j$, *then both* $\underline{\eta}$ *and* $\bar{\eta}$ *are bounded and bounded away from zero.*

The above assumption requires that the parameter η should not be too small or too large, as it impacts the concentration behavior and asymptotic normality of our estimator. This assumption can be satisfied by design in our method. We recommend using a default value of $\eta_j = 0.5$ for all j (where $1 \le j \le p$), which ensures that both $\underline{\eta}$ and $\bar{\eta}$ are bounded and bounded away from zero. This choice simplifies the implementation while maintaining the theoretical guarantees of our method. Our simulation study also suggests that our method is not sensitive to the choice of η.

We are now in a position to describe the asymptotic behavior of our bootstrap aggregated estimator. Without loss of generality, we consider a particular form of our estimator in an ideal case where $\tilde{\theta} = \mathbb{E}^*[\hat{\theta}_b]$.

Theorem 1

Under Assumptions 1-4, as $s_\lambda \xrightarrow{\mathrm{p}} \infty$ *and* $\frac{\kappa_\lambda}{\lambda^2} \xrightarrow{\mathrm{p}} \infty$, *our proposed estimator satisfies the following representation*

$$\frac{a_\lambda}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \cdot (\tilde{\theta} - \theta) = \frac{1}{\sqrt{s_\lambda \cdot \kappa_\lambda}} \sum_{j \in S_\lambda} \tau_j \cdot \tilde{u}_j + o_p(1).$$

where $\tilde{u}_j = \hat{\beta}_{X_{j,\mathrm{RB}}}(\theta \cdot \beta_{X_j} + \nu_j) - \theta(\hat{\beta}^2_{X_{j,\mathrm{RB}}} - \hat{\sigma}^2_{X_{j,\mathrm{RB}}})$. *Therefore, conditional on the selection event* $\mathcal{S}_\lambda$*, our estimator converges to a Gaussian distribution, that is*

$$\tilde{\sigma}^{-1}(\tilde{\theta} - \theta); \quad N(0,1), \text{ where } \tilde{\sigma}^2 = \frac{\sum_{j \in S_\lambda} \tau_j^2 \mathbb{V}[\tilde{u}_j \mid \mathcal{S}_\lambda]}{a_\lambda^2}.$$

In the theorem above, we consider the asymptotic regime in which both $s_\lambda \xrightarrow{\mathrm{p}} \infty$ and $\frac{\kappa_\lambda}{\lambda^2} \xrightarrow{\mathrm{p}} \infty$ tend towards infinity. This asymptotic regime is quite natural in the context of MR. On the one hand, $s_\lambda \xrightarrow{\mathrm{p}} \infty$ requires the number of IVs selected through re-randomization to be large enough, so that our inverse variance weighting-based estimator exhibits concentrated behavior. On the other hand, the condition $\frac{\kappa_\lambda}{\lambda^2} \xrightarrow{\mathrm{p}} \infty$ does not involve the bootstrapping procedure; instead, it pertains to the strength of the selected IVs relative to the threshold λ used in the re-randomization step (Step 1). This assumption ensures that, on average, the selected IVs are sufficiently strong compared to the threshold, thereby satisfying the relevance assumption. It is also likely to hold, as it is of the same order as the GWAS sample size *n* after IV selection through re-randomization. From a theoretical standpoint, both conditions have been rigorously verified in Ma et al. (2023) under appropriate conditions.

5 Simulations studies

We generate different simulation settings to evaluate the methods performance. To save space, the simulation settings are put into Supplementary Section S.8.1. Figure 2 summarizes the performance of various MR methods under the setting of 50% of the IVs are invalid, which we discuss below.

First, both cML (Type 1 error rate: 0.136) and MR-Lasso (0.112) produce inflated Type 1 error rates. This is because cML and MR-Lasso ignore the randomness in the valid IV selection procedure and assume all invalid IVs have been screened out, which is not the case under this simulation setting. In contrast, cML-DP (0.042) and CARE (0.042), which explicitly consider the randomness in valid IV selection, yield well-calibrated Type 1 error rates. Furthermore, other benchmark methods, including (random effects) IVW (0.056), MR-Egger (0.050), MRmix (0.020), MR-Median (0.032), MR-mode (0.004), MR-APSS (0.054) and RAPS (0.038) also yield well-controlled Type 1 error rates, though MRmix, MR-Median, MR-mode, and RAPS yield slightly conservative Type 1 error rates. Notably, the winner's curse bias itself does not cause an inflated Type 1 error rate issue (Ma et al., 2023), partially explaining the robust performance of many MR methods under the null.

Second, CARE achieves considerably higher statistical power than benchmark methods (Figure 2a). Notably, CARE corrects the winner's curse bias and measurement error bias, which allows for a more liberal threshold (say, $p < 5 \times 10^{-5}$) for instrument selection,

resulting in higher power than other methods that typically use the genome-wide significance level ($p < 5\times10^{-8}$) as the threshold. Even though MR-APSS, like CARE, allows a liberal threshold ($p < 5\times10^{-5}$) due to its direct winner's curse bias correction without theoretical guarantee, CARE outperforms MR-APSS, because of its full correction of the winner's curse bias and meticulous consideration of measurement errors and invalid IVs. To assess the influence of the IV selection threshold, we also compared all methods using the same liberal threshold of $p < 5\times10^{-5}$. While some competing methods showed increased power, this often came at the cost of inflated Type I error rates and poor confidence interval coverage. CARE maintained its advantages in terms of bias, mean squared error, and valid inference (see Supplementary Figure S28 for details).

Algorithm 2: CARE

for $j \leftarrow 1$ *to p* do

Generate a pseudo SNP-exposure association effect $Z_j \sim \mathcal{N}(0, \eta^2)$,

If $\left| \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j \right| > \lambda$, **Then** select SNP *j* .

end

Define the set of selected SNPs as $\mathcal{S}_\lambda = \{ j : \left| \frac{\hat{\beta}_{X_j}}{\sigma_{X_j}} + Z_j \right| > \lambda,\ j = 1, 2, \ldots, p \}$ and $|\mathcal{S}_\lambda| = s_\lambda$,

for $j \in \mathcal{S}_\lambda$ **do**

Construct an unbiased estimator of $\hat{\beta}_{X_j,\mathrm{RB}}$ as

$$\hat{\beta}_{j,\mathrm{RB}} = \hat{\beta}_{X_j} - \frac{\sigma_{X_j}}{\eta} \frac{\phi(A_{j,+}) - \phi(A_{j,-})}{1 - \Phi(A_{j,+}) + \Phi(A_{j,-})}, \text{ where } A_{j,\pm} = -\frac{\hat{\beta}_{X_j}}{\sigma_{X_j}\eta} \pm \frac{\lambda}{\eta}$$

and $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative distribution functions.

end

for $b = 1$ *to B* **do**

Draw bootstrap sample $\mathcal{S}^*_{\lambda,b}$ from $\mathcal{S}_\lambda$,

Conduct the invalid IV screening procedure for $\mathcal{S}^*_{\lambda,b}$

$$\min_{\theta\in\mathbb{R}, r_j\in\mathbb{R}} \left\{ \hat{l}^*_b\left(\theta, \{r_j\}_{j\in\mathcal{S}_\lambda}\right) : \sum_{j\in\mathcal{S}^*_{\lambda,b}} 1_{r_j=0} = v \right\} \quad \Rightarrow \quad \mathcal{V}^*_b(v) = \left\{ j : \hat{r}_j = 0,\ j \in \mathcal{S}^*_{\lambda,b} \right\},$$

where $\hat{l}_b^*\left(\theta,\{r_j\}_{j\in\mathcal{S}_\lambda}\right)=\sum_{j\in\mathcal{S}_\lambda} w_{jb}^*\hat{l}_j\left(\theta,\{r_j\}_{j\in\mathcal{S}_\lambda}\right)$.

Select the final estimated set of Valid IVs $\hat{\mathcal{V}}_b^*$ by GBIC,

Derive the causal estimator for the *b* -th bootstrap

$$\hat{\theta}_b = A_b^{-1}\sum_{j\in\mathcal{V}_b}\frac{\hat{\beta}_{Y_j}\hat{\beta}_{X_j,\mathrm{RB}}}{\sigma_{Y_j}^2},\quad A_b=\sum_{j\in\mathcal{V}_b}\frac{\hat{\beta}_{X_j,\mathrm{RB}}^2-\hat{\sigma}_{X_j,\mathrm{RB}}^2}{\sigma_{Y_j}^2}.$$

end

Obtain the final estimator by bootstrap aggregation $\tilde{\theta}=\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$,

Adopt the non-parametric delta method to estimate the variance of the bagged estimator with

$$\hat{\sigma}_n^2=\sum_{j\in\mathcal{S}_\lambda}\hat{S}_j^2,\quad \hat{S}_j=\frac{1}{B}\sum_{b=1}^{B}\left(w_{ib}^*-\frac{1}{B}\sum_{k=1}^{B}w_{ik}^*\right)(\hat{\theta}_b-\tilde{\theta}),$$

Construct a $(1-\alpha)$ -level confidence interval for θ with $\left[\tilde{\theta}-z_{\frac{\alpha}{2}}\cdot\hat{\sigma}_n,\tilde{\theta}+z_{\frac{\alpha}{2}}\cdot\hat{\sigma}_n\right]$, here $z_{\frac{\alpha}{2}}$ is the upper $\alpha/2$ -quantile of the standard normal distribution. CARE

Third, CARE yields smaller absolute bias compared to benchmark methods, attributable to its comprehensive approach to simultaneously addressing multiple sources of bias (measurement error bias, pleiotropic effects, and winner's curse bias). In comparison, benchmark methods focus on addressing some biases specifically, leading to biased results. For instance, while MR-APSS directly corrects for the winner's curse bias and considers potential invalid IVs, it still presents a larger absolute bias compared to CARE, possibly due to its more limited scope in bias correction and incomplete correction of the winner's curse bias. However, while CARE significantly reduces bias, its estimates are not entirely bias-free. This residual bias likely stems from the subtle differences between valid and invalid IVs. Consequently, the estimates are inevitably influenced by some invalid IVs, albeit to a lesser extent than in other methods. Furthermore, we confirm that ignoring the winner's curse bias and directly applying the measurement error model with $\hat{\beta}_{X_j}$ in CARE generally results in worse performance, particularly concerning the absolute bias (Supplementary Figure S1). As expected, CARE yields much smaller MSE compared to benchmark methods as CARE has higher power and smaller absolute bias than any benchmark methods.

Fourth, the confidence intervals provided by CARE have coverage probabilities close to the nominal 95% level. When the absolute causal effect $|\theta|$ is large (say, 0.1), the absolute bias is relatively large, resulting in slight undercoverage of the true causal effect.

Furthermore, to validate the results are not sensitive to the specific value of η within a reasonable range, we conducted sensitivity analyses using different values of η (0.1, 0.3, 0.5, 0.7, 0.9) in our main setting. The results demonstrate that the performance of our method remains stable and consistent for η values between 0.3 and 0.9 (Section S.8.6 in Supplementary Material). As expected, a very small η (0.1) led to worse results, likely due to

insufficient rerandomization to fully account for the winner's curse bias. Based on these findings, we recommend that practitioners use the default value of $\eta = 0.5$ in most cases without the need for dataset-specific fine-tuning.

While CARE demonstrates robust performance across various scenarios, it is important to note its limitations. As one reviewer suggested, we consider a simulation scenario that the parameter assumptions of other methods are true (where a three-sample MR design is used and the first GWAS is reserved solely for IV selection based on association strength so that the normality of $\hat{\beta}_{X_j}$ is not distorted). In this case, some alternative robust MR methods may outperform CARE, indicating that other robust MR methods may outperform CARE in a three-sample MR desgin (Supplementary Section S.8.13). Further simulations revealed two situations CARE is suboptimal. Firstly, in settings with non-linear relationships between genetic variants and exposures, CARE showed slightly inflated Type 1 error rates, larger bias, and worse coverage (Section S.8.8 in Supplementary Material). This limitation stems from the method's underlying assumption of linear relationships, which is common in MR studies and often justified by the predominantly linear or additive nature of genetic effects on complex traits (Wainschtein et al., 2022). Unlike our current approach, which exclusively utilizes GWAS summary data to estimate causal effects, recent advancements have addressed the non-linearity issue through methods like DeepMR Malina et al. (2022), a deep learning-based approach applicable when individual-level DNA sequence data are available. Secondly, CARE's performance may be compromised when the sample size of the exposure GWAS is small, resulting in a limited number of selected candidate IVs (Section S.8.9 in Supplementary Material). This issue may also arise due to a relatively small number of independent IVs (Section S.8.10 in Supplementary Material). Such scenarios can lead to increased sensitivity to violations of IV assumptions and challenge our asymptotic normality results, which require the number of candidate IVs to approach infinity. Users should exercise caution when applying CARE and other MR methods in these scenarios and consider alternative methods or larger sample sizes when possible.

In the end, it is worth mentioning that the core algorithm in CARE is written in C++ using the R package RcppArmadillo, and each step within the algorithm has a closed-form solution. Consequently, CARE has similar computational efficiency to many other methods, such as cML-DP and MRmix (Supplementary Figure S4), despite utilizing a larger number of IVs and a relatively high number of bootstrap iterations (2,000). Under the main simulation setting (12,000 simulations across 30%, 50%, and 70% invalid IVs), the average computational time of CARE is 12.6 seconds. Notably, the computational time for all methods is less than a minute in most situations when using one single core in a server. Thus, computational time should not be the primary consideration when deciding the method to be used.

6 Case studies

In this section, we investigate the performance of proposed CARE in two case studies. We put the data harmonization details in Supplementary Section S.9.1.

6.1 Negative control outcomes

To evaluate the Type 1 error rates in real data, we employ negative control outcome analyses, applying CARE and benchmark methods to investigate the causal effect of exposures on outcomes known a priori to have no causal relationship with the exposures. Briefly, in these negative control outcome analyses, the causal effect size is expected to be $\theta = 0$ (Sanderson et al., 2021) because negative control outcomes are determined prior to the exposures. However, unmeasured confounding factors may affect the estimates of θ. In particular, following others (Sanderson et al., 2021), we use ease of skin tanning to sun exposures and natural hair color before greying (six outcomes: Ease of skin tanning, Hair color black, Hair color red, Hair color blonde, Hair color light brown, and Hair color dark brown) as negative control outcomes. These data were downloaded from the IEU OpenGWAS Project (Lyon et al., 2021) with GWAS ID: ukb-b-533 and ukb-d-1747. Notably, both tanning ability and natural hair color before greying are primarily determined at birth (thus, prior to considered exposures) but could be affected by unmeasured confounders (Sanderson et al., 2021). In this setting, the inclusion of invalid IVs due to widespread pleiotropic effects or unmeasured confounding factors (e.g., population stratification) may result in incorrect rejections of the null hypothesis ($\theta = 0$) for MR analyses, leading to inflated Type 1 error rates.

We consider 45 exposures, which include HDL cholesterol, body mass index (BMI), height, Alzheimer's disease, Lung cancer, Type 2 diabetes, stroke, asthma, and many others. All GWAS data are downloaded from the IEU OpenGWAS Project (Lyon et al., 2021), and details of each exposure are relegated to the Supplementary Table 1. These exposures were selected based on their prevalence in existing literature and relevance to public health. Specifically, traits such as BMI, height, and HDL cholesterol have been extensively studied in genetic epidemiology and are known to be associated with various health outcomes. Disease outcomes like Alzheimer's disease, Type 2 diabetes, and cardiovascular diseases represent major public health concerns and have been the focus of numerous Mendelian randomization studies. This diverse set of exposures covers a wide range of physiological and pathological processes, allowing us to evaluate CARE's performance across various scenarios commonly encountered in Mendelian randomization studies. We apply CARE and benchmark methods to infer causal effects between these 45 exposures and six negative control outcomes (tanning ability and natural hair color before greying), resulting in 270 trait pairs. The corresponding *p*-values should follow a standard uniform distribution, given that the causal effect size $\theta = 0$ under the negative control outcomes analysis.

Figure 3 summarizes the QQ-plots of $-\log_{10}(p)$ values for different methods. First, CARE yields well-calibrated *p*-values, indicating its reliability in controlling type 1 error rates under this negative control outcome analysis (Figure 3A). Similarly, IVW, cML-DP and MR-APSS also achieve good performance (Figure 3B). In contrast, MR-mix, MR-Egger, RAPS, ContMix, cML, Weighted-Median, Weighted-Mode, and MR-Lasso yield inflated p-values (Figures 3C and 3D). One may be surprised that widely used IVW achieves good performance. This is because we make every effort to make a fair comparison between different methods and use the (random effects) IVW to consider pleiotropic effects (i.e.,

invalid IVs) by allowing over-dispersion in the regression model. As expected, the fixed effects IVW that assumes all used IVs are valid leads to inflated p-values (Supplementary Figure S34A).

To understand why CARE performs well, we highlight two aspects. First, selecting valid IVs can be noisy in real data applications. That explains why cML and MR-Lasso, methods that ignore the screening variability in IV selection, produce inflated p-values (Figure 3D). Applying bagging reduces the screening variability and thus helps achieve well-calibrated p-values in CARE. Similarly, as cML-DP uses a data perturbation method to account for the screening variability, it also achieves relatively good performance. Second, CARE adopts a rerandomization step to select candidate IVs, accounting for the impact of the winner's curse bias. Breaking the winner's curse bias helps CARE achieve well-calibrated p-values as CARE uses a measurement error model and relies on the unbiasedness estimation of exposure-SNP effect β_{X_j}. This rerandomization step is crucial for CARE, and we confirm that applying CARE without the rerandomization step leads to inflated p-values (Supplementary Figure S34B).

6.2 Risk factors identification for COVID-19 severity

To better understand the underlying causal risk factors for COVID-19 severity and demonstrate the performance of our proposed method CARE, we apply CARE and competing MR methods to systematically identify causal risk factors for COVID-19 severity. Specifically, we investigate the same 45 exposures used in the negative control outcome analysis and use COVID-19 severity (B2) from the covid-19hg (B2, version v7, European ancestry only; (Initiative, 2021)) as our outcome data. The dataset includes data from 32,519 hospitalized COVID-19 patients and 2,062,805 population controls.

First, we compare the number of significant causal exposures identified by CARE and competing methods under the Bonferroni correction ($< 0.05/45 \simeq 10^{-3}$) (Figure 4A). CARE identifies 6 causal exposures. In comparison, the competing methods RAPS, cML-DP, IVW, MR-Lasso, MR-APSS, MR-mix, ContMix, Weighted-Median, Weighted-Mode, MR-Egger identify 7, 5, 5, 5, 4, 4, 3, 0, 0 and 0 causal exposures, respectively. In terms of statistical power, CARE ranks second among all MR methods considered. RAPS achieves the highest power but also yields inflated p-values in our negative control outcome analysis and simulations, primarily due to neglecting variability in valid IV selection step.

Second, we compared the risk factors identified by different MR methods to known factors that meet two criteria: (1) they have been reported by the CDC or in peer-reviewed literature, and (2) they overlap with the 45 exposures used in our negative control outcome analyses. Through a comprehensive manual review by two researchers, we identified 24 well-established risk factors for COVID-19 severity (Supplementary Table 1). Notably, our new method, CARE, demonstrated superior performance by correctly identifying six of these 24 known risk factors: BMI, extreme BMI, HDL cholesterol, obesity class 1, obesity class 2, and overweight. In comparison, benchmark methods showed lower detection rates: MR-LASSO identified 5 risk factors, while cML-DP, IVW, MR-APSS, MR-Mix, and RAPS each identified 4. ContMix detected 3, and Median identified 2. Both Weighted-Mode and MR-Egger failed to identify any risk factors (Figure 4B). Importantly, CARE also avoided false positives, i.e., it did not incorrectly identify any factors lacking strong supporting evidence in the literature. In contrast, several benchmark methods produced potential false positives. For

example, cML-DP incorrectly identified childhood obesity as a risk factor, while IVW erroneously identified both celiac disease and childhood obesity. Finally, when we focus on four methods with relatively good performance under our negative control outcome analysis, the result patterns are similar (Supplementary Section S.9.2).

In summary, CARE achieves high power in identifying likely causal risk factors for COVID-19 severity, and the identified risk factors can be largely validated by complementary analyses and literature.

7 Conclusion

We introduced a unified two-sample Mendelian randomization within the summary data framework, referred to as Causal Analysis with Randomized Estimators (CARE), that accounts for winner's curse, measurement error bias, and genetic pleiotropy simultaneously. Through simulations and biomedical applications, we demonstrate that CARE delivers robust causal effect estimates with improved statistical power. More importantly, the CARE estimator enjoys rigorous theoretical guarantees under mild assumptions, which is often lacking for competing methods.

Funding and Acknowledgments

The research was supported by NIH R01AG089512, NSF DMS-2239047, NIH R01CA263494 and NIH U01CA293883. The authors thank the editor, the associate editor, and referees for their constructive feedback, which led to significant improvements in the article.

Disclosure Statement

The authors report there are no competing interests to declare.

References

Boehm, F. J. and Zhou, X. (2022). Statistical methods for mendelian randomization in genome-wide association studies: A review. *Computational and Structural Biotechnology Journal*, 20:2338–2351.

Bowden, J. and Dudbridge, F. (2009). Unbiased estimation of odds ratios: combining genomewide association scans with replication studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(5):406–418.

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust and efficient method for mendelian randomization with hundreds of genetic variants. *Nature Communications*, 11(1):1–11.

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

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007.

Flegal, K. M., Graubard, B. I., Williamson, D. F., and Cooper, R. S. (2011). Reverse causation and illness-related weight loss in observational studies of body weight and mortality. *American Journal of Epidemiology*, 173(1):1–9.

Gelman, A. and Imbens, G. (2013). Why ask why? Forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research.

Gkatzionis, A. and Burgess, S. (2019). Contextualizing selection bias in mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*, 48(3):691–701.

Grant, A. J. and Burgess, S. (2024). A bayesian approach to mendelian randomization using summary statistics in the univariable and multivariable settings with correlated pleiotropy. *The American Journal of Human Genetics*, 111(1):165–180.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252.

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

Hu, X., Zhao, J., Lin, Z., Wang, Y., Peng, H., Zhao, H., Wan, X., and Yang, C. (2022). Mendelian randomization for causal inference accounting for pleiotropy and sample structure using genome-wide summary statistics. *Proceedings of the National Academy of Sciences*, 119(28):e2106858119.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789.

Initiative, C.-. H. G. (2021). Mapping the human genetic architecture of covid-19. *Nature*, 600(7889):472–477.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163.

Liu, Z., Qin, Y., Wu, T., Tubbs, J. D., Baum, L., Mak, T. S. H., Li, M., Zhang, Y. D., and Sham, P. C. (2023). Reciprocal causation mixture model for robust mendelian randomization analysis using genome-scale summary data. *Nature Communications*, 14(1):1131.

Luo, R., Wang, H., and Tsai, C.-L. (2008). On mixture regression shrinkage and selection via the mr-lasso. *International Journal of Pure and Applied Mathematics*, 46(3):403–414.

Lyon, M. S., Andrews, S. J., Elsworth, B., Gaunt, T. R., Hemani, G., and Marcora, E. (2021). The variant call format provides efficient and robust storage of gwas summary statistics. *Genome Biology*, 22(1):1–10.

Ma, X., Wang, J., and Wu, C. (2023). Breaking the winner's curse in mendelian randomization: Rerandomized inverse variance weighted estimator. *The Annals of Statistics*, 51(1):211–232.

Malina, S., Cizin, D., and Knowles, D. A. (2022). Deep mendelian randomization: Investigating the causal knowledge of genomic deep learning models. *PLOS Computational Biology*, 18(10):e1009880.

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

Pearl, J. (2009). *Causality*. Cambridge University Press.

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

Sadreev, I. I., Elsworth, B. L., Mitchell, R. E., Paternoster, L., Sanderson, E., Davies, N. M., Millard, L. A., Smith, G. D., Haycock, P. C., Bowden, J., et al. (2021). Navigating sample overlap, winner's curse and weak instrument bias in Mendelian randomization studies using the UK biobank. *medRxiv*.

Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., et al. (2022). Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):6.

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

Skrivankova, V. W., Richmond, R. C., Woolf, B. A., Davies, N. M., Swanson, S. A., VanderWeele, T. J., Timpson, N. J., Higgins, J. P., Dimou, N., Langenberg, C., et al. (2021). Strengthening the reporting of observational studies in epidemiology using Mendelian randomisation (STROBE-MR): Explanation and elaboration. *BMJ*, 375.

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

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599.

Wainschtein, P., Jain, D., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., Mitchell, B. D., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics*, 54(3):263–273.

Watanabe, K., Stringer, S., Frei, O., Mirkov, M. U., de Leeuw, C., Polderman, T. J., van der Sluis, S., Andreassen, O. A., Neale, B. M., and Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348.

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

Ye, T., Shao, J., and Kang, H. (2021). Debiased inverse-variance weighted estimator in two-sample summary-data mendelian randomization. *The Annals of Statistics*, 49(4):2079–2100.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, 48(3):1742–1769.

Zhong, H. and Prentice, R. L. (2010). Correcting "winner's curse" in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology*, 34(1):78–91.

Zöllner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, 80(4):605–615.

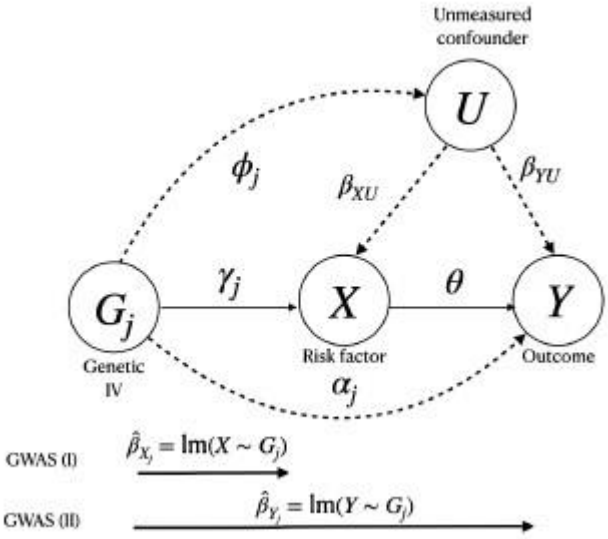

Figure 1: The causal diagram and GWAS (I) and (II) summary data adopted in the two-sample MR. The corresponding causal effect for each pathway is labeled near the directed edge.

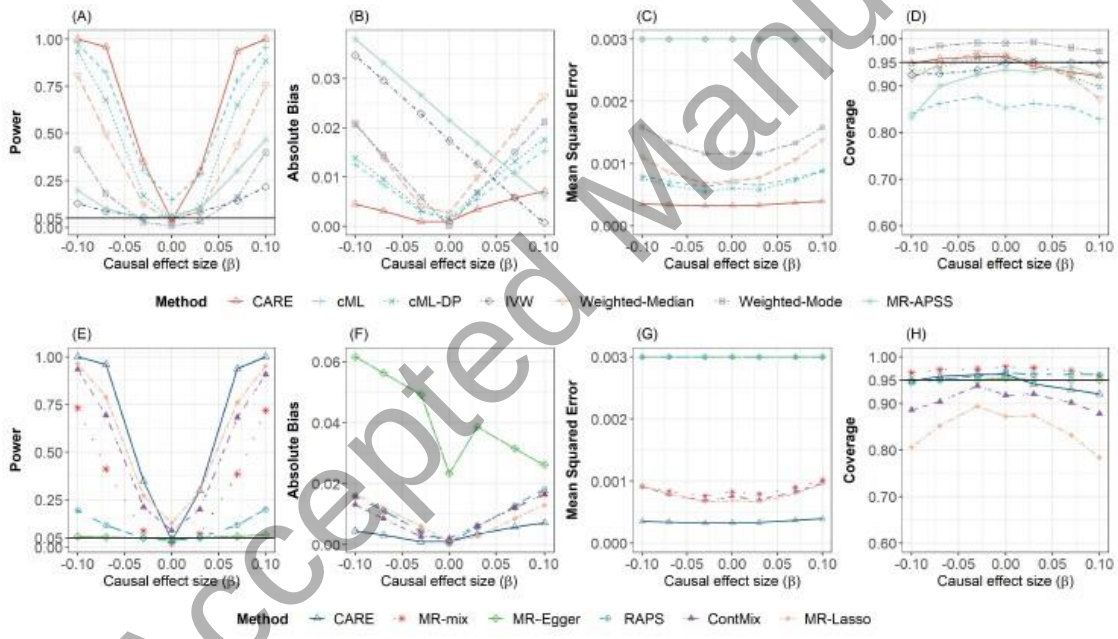

Figure 2: Power, absolute bias, mean squared error, and coverage of the CARE estimator and several robust MR methods under the main setting with 50% invalid IVs. Power is the empirical power estimated by the proportion of p-values less than the significance threshold of 0.05. Coverage is the empirical coverage probability of the 95% confidence interval.

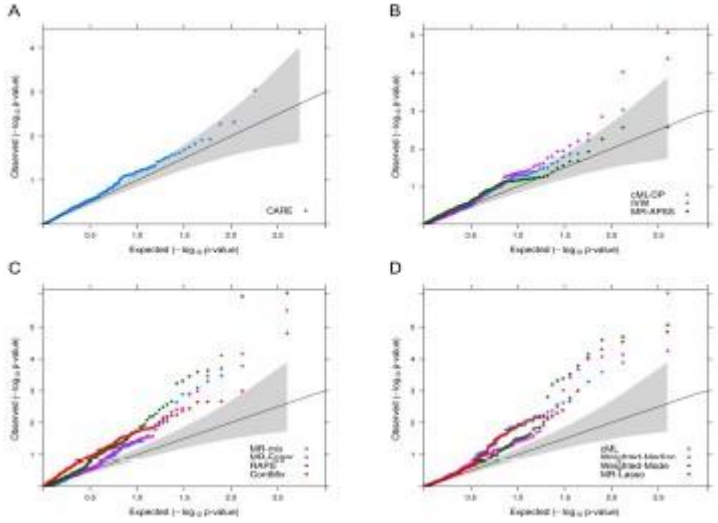

Figure 3: QQ plots of p-values in negative control outcome analysis. The gray-shaded part is 95% confidence interval.

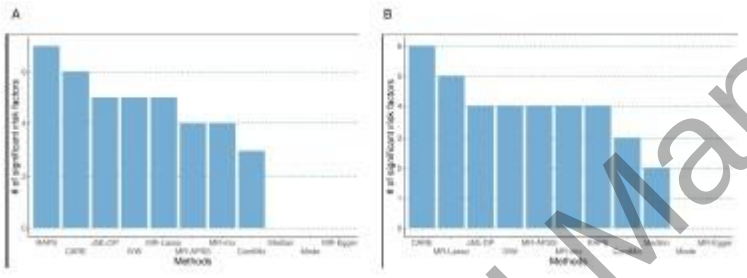

Figure 4: Number of significant causal pairs identified by different methods under Bonferroni-correction threshold $< 0.05/45 \simeq 10^{-3}$ using (A) 45 exposures used in negative control analysis and (B) 24 exposures that are reported by CDC and existing literature.