MRBEE: A novel bias-corrected multivariable Mendelian Randomization method

Noah Lorincz-Comi[1], Yihe Yang[1], Gen Li[1] and Xiaofeng Zhu[1*]

[1*]Department of Population and Quantitative Health Sciences, Case Western Reserve University.

*Corresponding author(s). E-mail(s): xxz10@case.edu;

Abstract

Mendelian randomization (MR) is an instrumental variable approach used to infer causal relationships between exposures and outcomes using summary data from genome-wide association studies (GWAS). Since GWAS summary statistics are subject to estimation errors, most existing MR approaches suffer from the error-in-variable bias, whose scale and direction are influenced by weak instrumental variables and GWAS sample overlap, respectively. We introduce MRBEE (MR using Bias-corrected Estimating Equation), a novel multivariable MR method capable of simultaneously removing error-in-variable bias and identifying horizontal pleiotropy. In simulations, we showed that MRBEE is the only method capable of effectively removing error-in-variable bias with weak instrumental variables and arbitrary sample overlap. In two independent real data analyses, we discovered that the causal effect of BMI on coronary artery disease risk is entirely mediated by blood pressure, and that existing MR methods may underestimate the causal effect of cannabis use disorder on schizophrenia risk compared to MRBEE. MRBEE holds significant potential for advancing genetic research by providing a valuable tool to study causality between multiple risk factors and disease outcomes, particularly as a large number of GWAS summary statistics become publicly available.

Keywords: mendelian randomization, genetic epidemiology, statistical genetics, complex disease

1 Introduction

Mendelian randomization (MR) is a powerful epidemiological method that leverages genetic variants as instrumental variables (IVs) to infer causal relationships between exposures and outcomes, reducing confounding and reverse causation, while providing a cost-effective, ethical, and generalizable alternative to randomized controlled trials [1–3]. Originally developed for application in individual-level data [2], many MR methods can also be applied to summary-level statistics obtained from genome-wide association studies (GWAS) and have therefore become increasingly popular to infer causality of disease risk factors [3], identify biological drug targets [4], and causal effects of genes on phenotypes [5]. Inverse-variance weighting (IVW) [62] is the most fundamental approach to perform MR with GWAS summary data, the validity of which relies heavily on three so-called valid IVs assumptions: the genetic IVs are (i) strongly associated with the exposures, (ii) not directly associated with the outcome conditional on the exposures, and (iii) not associated with any confounders of the exposure-outcome relationships. Violations of the (i) - (iii) assumptions will respectively introduce weak instrument [11], uncorrelated horizontal pleiotropy (UHP) [3], and correlated horizontal pleiotropy (CHP) [8] biases into the casual effect estimation of IVW.

From a statistical perspective, UHP in an MR model exhibits similar characteristics to outliers or heavy-tail errors in traditional regression, which can be addressed using robust tools. Methods such as MR pleiotropy residual sum and outlier (MR-PRESSO) [23] and iterative MR pleiotropy (IMRP) [21] detect and remove potential UHP through pleiotropy tests, while MR-Median [24], MR-Robust [64], and MR-Lasso [7] mitigate the impact of UHP by implementing robust loss functions. Conversely, CHP typically leads to a misspecified MR model that can result in biased causal effect estimates. Techniques such as MRMix [16], MR contamination mixture (MR-Conmix) [63], causal analysis using summary effect (CAUSE) [8], MR constrained maximum likelihood (MR-CML) [28], and MR with correlated horizontal pleiotropy unraveling shared etiology and confounding (MR-CUE) [17] employ Gaussian mixture models to remove CHP effects when estimating causal effects.

While the aforementioned univariable MR methods allow for partial IVs to exhibit horizontally pleiotropic effects, they typically assume that the majority of IVs impact the outcome solely through the mediation of the exposure. Nevertheless, considerable evidence suggests that common human traits share a significant amount of causal variants, such as low-density lipoprotein cholesterol and total cholesterol [14], or systolic blood pressure and diastolic blood pressure [18], making it difficult to satisfy this assumption in reality. A more robust, straightforward, and computationally efficient way to reduce or even eliminate the effect of horizontal pleiotropy is to employ multivariable MR, which can automatically account for horizontally pleiotropic variants shared by multiple exposures. To date, multivariable versions of IVW [29], MR-Egger [30], MR-Median [24], MR-Robust [31], and MR-CML [32] have been developed. As demonstrated in an examination by [12], multivariable MR is a

reliable tool for estimating the direct causal effects of one or more exposures, using either individual-level or summary-level data.

As the sample sizes in GWAS expand, increasing numbers of causal variants with moderate or small effects are being identified, making the weak instrument problem, or the violation of assumption (i), more significant and difficult to disregard. For instance, [67] detected 12111 independent variants in a height GWAS with 5.4 million participants, [65] found nearly 3952 independent variants in an educational attainment GWAS with 3.0 million participants, and [66] uncovered around 2286 independent variants in a lipid GWAS involving 1.6 million participants. Since the heritability of a trait is fixed, the variance explained by each causal variant should be very small if there is thousands of them, which thus causes a weak instrument problem. The traditional solution to mitigate the weak instrument problem is discarding IVs with small effect sizes such that the F- or conditional F-statistic of the remained instruments exceeds 10, which approximately guarantees that the relative bias in causal effect estimation remains within 10% [11, 26]. However, excluding instruments with weaker effects can result in a "winner's curse", which alternatively inflates the bias in causal estimation [68]. Additionally, the statistical principle underlying how weak IVs lead to biased causal effect estimation has not been well understood.

Error-in-variable bias occurs when explanatory variables are measured with random error, leading to biased estimation of model parameters. Since GWAS summary statistics are subject to estimation errors, IVW and most existing MR approaches are presumed to suffer from error-in-variable bias even in large samples. Weak IVs can worsen this bias since MR essentially performs regression on both the estimation errors and true IV effect sizes. This is the primary reason why violating assumption (i) introduces bias into causal effect estimates in IVW and other MR approaches. Furthermore, unlike traditional error-in-variable models that require uncorrelated estimation errors in exposures and outcomes, overlapping samples in exposure and outcome GWAS can result in correlated errors, making the direction of error-in-variable bias not always uniform. This is the key reason why, in empirical studies (e.g., Figure 1 in [10]), IVW estimates exhibit negative bias with small numbers of overlapping samples and positive bias with large numbers of overlapping samples.

In this paper, we propose a multivariable MR method, MR using Bias-corrected Estimating Equations (MRBEE), to eliminate error-in-variable bias while simultaneously accounting for horizontal pleiotropy. In contrast to existing methods that only address error-in-variable bias in specific cases such as no sample overlap (debiased IVW [27]) or no horizontal pleiotropy (MRlap [59]), MRBEE offers a comprehensive solution to error-in-variable bias, uses clear statistical principles, accommodates arbitrary sample overlap, and adapts to both univariable and multivariable MR models. Through numerical simulations, we demonstrate that MRBEE is the only method capable of estimating causal effects without bias across various real-world conditions compared to its competitors. To exhibit its practical significance, we apply MRBEE in

two independent real-data analyses, first estimating the causal effects of cardiometabolic risk factors on coronary artery disease risk in two populations, and secondly estimating the causal effects of modifiable and non-modifiable risk factors for schizophrenia and bipolar disorder.

2 Method

2.1 Measurement error and MRBEE

Let $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{im})^\top$ be a vector of m independent genetic variants where each variant is standardized with mean zero and variance one, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ be a vector of p exposures, and y_i be an outcome. Consider the following linear structural model:

$$\boldsymbol{x}_i = \mathbf{B}^\top \boldsymbol{g}_i + \boldsymbol{u}_i, \tag{1}$$

$$y_i = \boldsymbol{\theta}^\top \boldsymbol{x}_i + \boldsymbol{\gamma}^\top \boldsymbol{g}_i + v_i, \tag{2}$$

where $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)^\top$ is an $(m \times p)$ matrix of genetic effects on exposures with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^\top$ being a vector of length p, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is a vector of length p representing the causal effects of the p exposures on the outcome, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^\top$ is a vector of length m representing pleiotropy, which violates the (IV2) and (IV3) conditions, and $\boldsymbol{u}_i$ and v_i are noise terms. Substituting for $\boldsymbol{x}_i$ in (2), we obtain the reduced-form model:

$$y_i = \boldsymbol{g}_i^\top \boldsymbol{\alpha} + \boldsymbol{u}_i^\top \boldsymbol{\theta} + v_i, \tag{3}$$

where $\boldsymbol{\alpha} = \mathbf{B}\boldsymbol{\theta} + \boldsymbol{\gamma}$. In practice, the random errors $\boldsymbol{u}_i$ and v_i are usually correlated, and hence traditional linear regression between $\boldsymbol{x}_i$ and y_i cannot obtain a consistent estimate of the causal effect vector $\boldsymbol{\theta}$. The genetic variant vector $\boldsymbol{g}_i$ is generally independent of $\boldsymbol{u}_i$ and v_i because the genotypes of individuals are randomly inherited from their parents and do not change during their lifetime. Therefore, $\boldsymbol{g}_i$ can be used as IVs to remove the confounding effect of $\boldsymbol{u}_i$ and v_i.

However, individual-level data from GWAS are rarely publicly available. Therefore, most of the current MR analyses are performed with GWAS summary data through the following linear regression:

$$\hat{\alpha}_j = \hat{\boldsymbol{\beta}}_j^\top \boldsymbol{\theta} + \gamma_j + \varepsilon_j, \tag{4}$$

where $\hat{\alpha}_j$ and $\hat{\boldsymbol{\beta}}_j$ are estimated from the outcome and exposure GWAS for jth IV, γ_j is the horizontal pleiotropy, and ε_j represents the residual of this regression model. Multivariable IVW is a weighted least squares method that estimates $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}}_{\texttt{IVW}} = \arg\min_{\boldsymbol{\theta}} \left\{ (\hat{\boldsymbol{\alpha}} - \widehat{\mathbf{B}}\boldsymbol{\theta})^\top \mathbf{V}^{-1} (\hat{\boldsymbol{\alpha}} - \widehat{\mathbf{B}}\boldsymbol{\theta}) \right\} = (\widehat{\mathbf{B}}^\top \mathbf{V}^{-1} \widehat{\mathbf{B}})^{-1} \widehat{\mathbf{B}}^\top \mathbf{V}^{-1} \hat{\boldsymbol{\alpha}}, \tag{5}$$

where $\mathbf{V}$ is a diagonal matrix consisting of the variance of estimation errors of $\hat{\boldsymbol{\alpha}}$. In practice, it is routine to standardize $\hat{\alpha}_j$ and $\hat{\beta}_{js}$ by $\hat{\alpha}_j/\mathrm{se}(w_{\alpha_j})$ and $\hat{\beta}_{js}/\mathrm{se}(w_{\beta_{js}})$ to remove the minor allele frequency effect [18]. With this standardization, the multivariable IVW is indeed an ordinal least estimate (OLS) which estimates $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}}_{\mathtt{IVW}} = \arg\min_{\boldsymbol{\theta}} \left\{ \|\hat{\boldsymbol{\alpha}} - \widehat{\mathbf{B}}\boldsymbol{\theta}\|_2^2 \right\} = (\widehat{\mathbf{B}}^\top \widehat{\mathbf{B}})^{-1} \widehat{\mathbf{B}}^\top \hat{\boldsymbol{\alpha}}. \tag{6}$$

Mutivariable IVW is biased due to the estimation errors of $\hat{\alpha}_j$ and $\hat{\boldsymbol{\beta}}_j$:

$$\hat{\alpha}_j = \boldsymbol{\alpha}_j + w_{\alpha_j}, \tag{7}$$

$$\hat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j + \boldsymbol{w}_{\beta_j}. \tag{8}$$

To see this, we consider the estimating equation and Hessian matrix of $\hat{\boldsymbol{\theta}}_{\mathtt{IVW}}$

$$\boldsymbol{S}_{\mathtt{IVW}}(\boldsymbol{\theta}) = \widehat{\mathbf{B}}^\top(\widehat{\mathbf{B}}\boldsymbol{\theta} - \hat{\boldsymbol{\alpha}}), \quad \mathbf{H}_{\mathtt{IVW}} = \widehat{\mathbf{B}}^\top \widehat{\mathbf{B}}. \tag{9}$$

That is, $\boldsymbol{S}_{\mathtt{IVW}}(\boldsymbol{\theta})$ is the score function of (6) and $\hat{\boldsymbol{\theta}}_{\mathtt{IVW}}$ is estimated by solving $\boldsymbol{S}_{\mathtt{IVW}}(\hat{\boldsymbol{\theta}}_{\mathtt{IVW}}) = \mathbf{0}$, and $\mathbf{H}_{\mathtt{IVW}}$ is the 2nd derivative matrix of (6). Since $\hat{\boldsymbol{\theta}}_{\mathtt{IVW}} - \boldsymbol{\theta} = -\mathbf{H}_{\mathtt{IVW}}^{-1}\boldsymbol{S}_{\mathtt{IVW}}(\boldsymbol{\theta})$, the bias of $\hat{\boldsymbol{\theta}}_{\mathtt{IVW}}$ is approximately:

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\theta}}_{\mathtt{IVW}} - \boldsymbol{\theta}) \approx & \underbrace{(\mathtt{cov}(\boldsymbol{\beta}_j) + \mathtt{cov}(\boldsymbol{w}_{\beta_j}))^{-1}}_{\text{weak instrument bias}} \underbrace{(\mathtt{cov}(\boldsymbol{w}_{\beta_j})\boldsymbol{\theta} - \mathtt{cov}(\boldsymbol{w}_{\beta_j}, w_{\alpha_j}))}_{\text{error-in-variable bias}} \\ & + (\mathtt{cov}(\boldsymbol{\beta}_j) + \mathtt{cov}(\boldsymbol{w}_{\beta_j}))^{-1} \underbrace{\mathtt{cov}(\boldsymbol{\beta}_j, \gamma_j)}_{\text{pleiotropy bias}} \end{aligned} \tag{10}$$

Here, $\mathtt{cov}(\boldsymbol{\beta}_j)$ can be regarded as the averaged information carried by each IV and $\mathtt{cov}(\boldsymbol{w}_{\beta_j})$ can be regarded as the information carried by each estimation error. If $\mathtt{cov}(\boldsymbol{\beta}_j)$ is not significantly larger than $\mathtt{cov}(\boldsymbol{w}_{\beta_j})$, then the weak instrument bias $(\mathtt{cov}(\boldsymbol{\beta}_j) + \mathtt{cov}(\boldsymbol{w}_{\beta_j}))^{-1}$ will inflate the error-in-variable bias.

On the other hand, $\mathtt{cov}(\boldsymbol{w}_{\beta_j}, w_{\alpha_j}))$ can be affected by the fraction of overlapping samples between exposures GWAS and outcome GWAS. If the exposures GWAS and outcome GWAS are independent of each other, then $\mathtt{cov}(\boldsymbol{w}_{\beta_j}, w_{\alpha_j})) = 0$ and hence the error-in-variable bias always shrinks $\hat{\boldsymbol{\theta}}_{\mathtt{IVW}}$ toward to null. In contrast, if the exposures GWAS and outcome GWAS are estimated from the same cohorts, $\mathtt{cov}(\boldsymbol{w}_{\beta_j}, w_{\alpha_j}))$ is usually towards the direction of $\mathtt{cov}(\boldsymbol{u}_i, v_i)$. This is the reason why in some empirical studies [10, 68], IVW cannot thoroughly remove the confounding bias if the overlapping sample fraction is large.

If $\mathtt{cov}(\boldsymbol{\beta}_j, \gamma_j) \neq \mathbf{0}$, there is additional pleiotropy bias due to the horizontal pleiotropy. In univariable MR, it is challenging to satisfy $\gamma_j = 0$ or $\mathtt{cov}(\gamma_j, \boldsymbol{\beta}_j) = \mathbf{0}$, resulting in a biased IVW estimate. Traditional solutions assume that only a small proportion of IVs exhibit horizontally pleiotropic

effects, and robust tools or Gaussian mixture models can be employed to identify these IVs. However, for complex traits, it is plausible that the majority of IVs ($> 50\%$) possess horizontally pleiotropic effects, leading to the failure of univariable MR methods. Multivariable MR can balance these pleiotropic effects shared by specified exposures, significantly reducing the number of IVs with pleiotropy evidence when conditioned on specified exposures. Notably, these pleiotropic variants are likely to appear as outliers and can be detected by existing tools such as pleiotropy tests.

2.2 MR using bias-corrected estimating equation

In this paper, we propose MRBEE which estimates causal effects by solving a new unbiased estimating equation of causal effect. Let $\mathtt{cov}(\boldsymbol{w}_{\beta_j}) = \boldsymbol{\Sigma}_{W_\beta W_\beta}\boldsymbol{\theta}$ and $\mathtt{cov}(\boldsymbol{w}_{\beta_j}, w_{\alpha_j}) = \boldsymbol{\sigma}_{W_\beta w_\alpha}$. The unbiased estimating equation of $\boldsymbol{\theta}$ is

$$\boldsymbol{S}_{\mathtt{BEE}}(\boldsymbol{\theta}) = \boldsymbol{S}_{\mathtt{IVW}}(\boldsymbol{\theta}) - m(\boldsymbol{\Sigma}_{W_\beta W_\beta}\boldsymbol{\theta} - \boldsymbol{\sigma}_{W_\beta w_\alpha}), \tag{11}$$

where $\boldsymbol{S}_{\mathtt{IVW}}(\boldsymbol{\theta}) = -\widehat{\mathbf{B}}^\top(\hat{\boldsymbol{\alpha}} - \widehat{\mathbf{B}}\boldsymbol{\theta})$. The solution $\hat{\boldsymbol{\theta}}_{\mathtt{BEE}}$ such that $\boldsymbol{S}_{\mathtt{BEE}}(\hat{\boldsymbol{\theta}}_{\mathtt{BEE}}) = \mathbf{0}$ is

$$\hat{\boldsymbol{\theta}}_{\mathtt{BEE}} = (\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}} - m\boldsymbol{\Sigma}_{W_\beta W_\beta})^{-1}(\widehat{\mathbf{B}}^\top\hat{\boldsymbol{\alpha}} - m\boldsymbol{\sigma}_{W_\beta w_\alpha}). \tag{12}$$

In MRBEE-IVW, how to estimate the bias-correction terms $\boldsymbol{\Sigma}_{W_\beta W_\beta}$ and $\boldsymbol{\sigma}_{W_\beta w_\alpha}$ may be the most important issue. Here, we recommend estimating them from insignificant GWAS summary statistics. Let $\hat{\alpha}_j^*, \hat{\beta}_{j1}^*, \ldots, \hat{\beta}_{jp}^*$ ($j = 1, \ldots, M$) be M insignificant GWAS effect size estimates of outcome and exposures, where insignificance means that the p-value of the genetic variants are larger than 0.05 for all exposures and outcome, and the independence means that these variants are in LE. Here, $\hat{\alpha}_j^*$ and $\hat{\beta}_{js}^*$ follow the same distributions of w_{α_j} and $w_{\beta_{js}}$, and hence $\boldsymbol{\Sigma}_{W_\beta \times w_\alpha}$ can be estimated by

$$\widehat{\boldsymbol{\Sigma}}_{W_\beta \times w_\alpha} = \frac{1}{M}\sum_{j=1}^{M}(\hat{\beta}_{j1}^*, \ldots, \hat{\beta}_{jp}^*, \hat{\alpha}_j^*)^\top(\hat{\beta}_{j1}^*, \ldots, \hat{\beta}_{jp}^*, \hat{\alpha}_j^*). \tag{13}$$

Here, $\widehat{\boldsymbol{\Sigma}}_{W_\beta W_\beta}$ is the first $(p \times p)$ sub-matrix of $\widehat{\boldsymbol{\Sigma}}_{W_\beta \times w_\alpha}$ and $\boldsymbol{\sigma}_{W_\beta w_\alpha}$ consists of the first $p-1$ elements of the last column of $\widehat{\boldsymbol{\Sigma}}_{W_\beta \times w_\alpha}$. Note that $\widehat{\boldsymbol{\Sigma}}_{W_\beta \times w_\alpha}$ is theoretically a correlation matrix if each GWAS effect size estimate is standardized. An alternative is estimating the correlation matrix of $(\hat{\beta}_{j1}^*, \ldots, \hat{\beta}_{jp}^*, \hat{\alpha}_j^*)^\top$.

We yield the covariance matrix of $\hat{\boldsymbol{\theta}}_{\mathtt{BEE}}$ through the sandwich formula:

$$\boldsymbol{\Sigma}_{\mathtt{BEE}}(\boldsymbol{\theta}) = \mathbf{F}_{\mathtt{BEE}}^{-1}\mathbf{V}_{\mathtt{BEE}}(\boldsymbol{\theta})\mathbf{F}_{\mathtt{BEE}}^{-1}, \tag{14}$$

where the outer matrix $\mathbf{F}_{\mathtt{BEE}}$ is the Fisher information matrix, i.e., the expectation of the Hessian matrix of $\boldsymbol{S}_{\mathtt{BEE}}(\boldsymbol{\theta})$, and the inner matrix $\mathbf{V}_{\mathtt{BEE}}(\boldsymbol{\theta})$ is the

covariance matrix of $\boldsymbol{S}_{\texttt{BEE}}(\boldsymbol{\theta})$. A consistent estimate of $\boldsymbol{\Sigma}_{\texttt{BEE}}(\boldsymbol{\theta})$ is

$$\widehat{\boldsymbol{\Sigma}}_{\texttt{BEE}}(\hat{\boldsymbol{\theta}}_{\texttt{BEE}}) = \widehat{\mathbf{F}}_{\texttt{BEE}}^{-1}\widehat{\mathbf{V}}_{\texttt{BEE}}(\hat{\boldsymbol{\theta}}_{\texttt{BEE}})\widehat{\mathbf{F}}_{\texttt{BEE}}^{-1}, \tag{15}$$

where $\widehat{\mathbf{F}}_{\texttt{BEE}} = \widehat{\mathbf{B}}^\top\widehat{\mathbf{B}}/m - \widehat{\boldsymbol{\Sigma}}_{W_\beta W_\beta}$, $\widehat{\mathbf{V}}_{\texttt{BEE}}(\hat{\boldsymbol{\theta}}_{\texttt{BEE}}) = \sum_{j=1}^m \hat{\boldsymbol{S}}_j(\hat{\boldsymbol{\theta}}_{\texttt{BEE}})\hat{\boldsymbol{S}}_j(\hat{\boldsymbol{\theta}}_{\texttt{BEE}})^\top/m$, and $\hat{\boldsymbol{S}}_j(\hat{\boldsymbol{\theta}}_{\texttt{BEE}}) = -(\hat{\alpha}_j - \hat{\boldsymbol{\theta}}_{\texttt{BEE}}^\top\hat{\boldsymbol{\beta}}_j)\hat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\Sigma}}_{W_\beta W_\beta}\hat{\boldsymbol{\theta}}_{\texttt{BEE}} + \hat{\boldsymbol{\sigma}}_{W_\beta w_\alpha}$.

2.3 Detecting IVs with horizontal pleiotropy

In the section above, we assumed the absence of horizontal pleiotropy in the IV set $\mathcal{S}$ used in MR (i.e., $j \in \mathcal{S} = \{j : \gamma_j = 0\}$). In practice, we may initially have a larger set of $m^* \geq m$ candidate IVs, some of which may have nonzero horizontally pleiotropic effects. Generally, identification of nonzero γ for a SNP, whether it may be considered UHP or CHP, requires that $(\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top\hat{\boldsymbol{\theta}}) \in \mathcal{S}$ is separable from $(\hat{\alpha}_s - \hat{\boldsymbol{\beta}}_s^\top\hat{\boldsymbol{\theta}}) \notin \mathcal{S}$ for $j \neq s$. This can be clearly seen from Equation 4 and is assumed by existing MR methods intended to address UHP/CHP [8, 16, 17]. Since $\texttt{E}(\varepsilon_j) = 0$, it is natural to assume that any IV with $\gamma_j \neq 0$ will behave like an outlier in the multivariable regression of $\hat{\boldsymbol{\alpha}}$ on $\hat{\mathbf{B}}$. We extended the IMRP [21] horizontal pleiotropy test to accommodate multiple MR exposures and identify specific IVs with evidence of nonzero γ_j and denote the new test statistic as $S_{\texttt{pleio}}$. IVs with large $S_{\texttt{pleio}}$ values are potentially horizontally pleiotropic and are removed from causal estimation. IVs with non-large $S_{\texttt{pleio}}$ statistics are used to estimate $\boldsymbol{\theta}$ using MRBEE in Equation 12.

The null hypothesis for the jth IV not having any horizontally pleiotropic effects on the outcome is the following:

$$H_{0j} : \alpha_j - \boldsymbol{\beta}_j^\top\boldsymbol{\theta} = \gamma_j = 0 \qquad \text{vs} \qquad H_{1j} : \gamma_j \neq 0. \tag{16}$$

The statistic $S_{\texttt{pleio}}$ for the jth IV is the following:

$$S_{\texttt{pleio}}^{(j)} = (\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top\hat{\boldsymbol{\theta}})^\top\hat{\boldsymbol{\Upsilon}}_j^{-1}(\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top\hat{\boldsymbol{\theta}}) \xrightarrow{D} \chi^2(1) \tag{17}$$

under H_{0j} where $\hat{\boldsymbol{\Upsilon}}_j$ is any consistent estimator for the variance of $\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top\hat{\boldsymbol{\theta}}$ (see below). The only assumption here is that $\hat{\alpha}_j - \hat{\boldsymbol{\beta}}_j^\top\hat{\boldsymbol{\theta}}$ is asymptotically normal, which it is as proven in [33] and shown in the **Supplement**. In practice, we can estimate Υ_j using the delta method as

$$\hat{\boldsymbol{\Upsilon}}_j = \sigma^2_{\alpha_{w_j}} + \hat{\boldsymbol{\theta}}^\top\boldsymbol{\Sigma}_{\mathbf{w}_{\beta_j}\mathbf{w}_{\beta_j}}\hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\beta}}_j^\top\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}_j - 2\hat{\boldsymbol{\theta}}^\top\boldsymbol{\sigma}_{\mathbf{w}_{\beta_j}w_{\alpha_j}}, \tag{18}$$

which is shown to converge in probability to the true value $\boldsymbol{\Upsilon}_j$ in [33] as $m \to \infty$. The test of the null hypothesis in 16 is tested using Equation 17 for all M candidate IVs that the researcher initially intends to use in MR. Some of these IVs will have sufficiently large $S_{\texttt{pleio}}$ and will be removed from causal

estimation using our software. Since $S_{\texttt{pleio}}$ requires the estimates $\hat{\boldsymbol{\theta}}$, we use the IMRP algorithm [21] to arrive at stable causal estimates from MRBEE after repeatedly removing IVs with horizontal pleiotropy and updating $\hat{\boldsymbol{\theta}}$ using the algorithm in Table 1.

Similar to the MR-Egger [44] adjustment, we can also perform a test for global unbalanced horizontal pleiotropy by adding an intercept term to MRBEE causal estimation. Let θ_0 denote the intercept term in multivariable MRBEE. It is shown in the **Supplement** that testing the null hypothesis that $\theta_0 = 0$ is equivalent to testing the null hypothesis that the mean of the IV-specific horizontally pleiotropic effects is 0. More explicitly we can test

$$H_0 : \theta_0 = m^{-1} \sum_{j=1}^{m} \gamma_j = 0 \qquad \text{vs} \qquad H_1 : m^{-1} \sum_{j=1}^{m} \gamma_j \neq 0 \tag{19}$$

to identify the presence of global unbalanced horizontal pleiotropy in the final IV set of m SNPs. Let $\boldsymbol{\theta}^* = (\theta_0, \boldsymbol{\theta}^\top)^\top$ and $\boldsymbol{\Lambda}_0$ be the asymptotic variance of $\hat{\boldsymbol{\theta}}^*$. We can test H_0 in Equation 19 using the statistic

$$Q_{\texttt{pleio}} = \frac{\mathbf{s}^\top \boldsymbol{\theta}^* \boldsymbol{\theta}^{*T} \mathbf{s}}{\mathbf{s}^\top \hat{\boldsymbol{\Lambda}} \mathbf{s}} \xrightarrow{D} \chi^2(1), \tag{20}$$

where $\mathbf{s}$ is a $(p+1) \times 1$ vector with leading element 1 and remaining elements 0 that is used only to extract the intercept-relevant terms from $\hat{\boldsymbol{\theta}}^*$ and $\hat{\boldsymbol{\Lambda}}$. The result in Equation 20 is proven in the **Supplement**. In practice, if $Q_{\texttt{pleio}}$ is sufficiently large, there is evidence of global unbalanced horizontal pleiotropy that can bias causal esitmation. In practice, IVs with large $S_{\texttt{pleio}}$ statistics should generally be removed irrespective of how large $Q_{\texttt{pleio}}$ is. Examples of how $S_{\texttt{pleio}}$ identifies horizontally pleiotropic IVs and $Q_{\texttt{pleio}}$ identifies imbalance in horizontal pleiotropy are shown in Figure 4A and Section S2.1.1 of the **Supplement**.

Table 1 IMRP algorithm with MRBEE

(1) Initialize: Estimate $\boldsymbol{\theta}$ using MRBEE and all IVs
(2) Pleiotropy test: Remove all IVs with $S_{\texttt{pleio}} > \xi$
(3) Causal estimation: Estimate $\boldsymbol{\theta}$ using MRBEE
(4) Iteration: If $\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_{k-1}\|_1 > \tau$ at iteration k, repeat steps 2-3; else, stop: $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_k$

ξ is chosen by the researcher, but is generally a quantile of the $\chi^2(1)$ distribution that is corrected for the number of IVs used; p is the number of exposures; τ is a tolerance level selected by the researcher and should be relatively small (e.g., <0.001) to ensure that true convergence can be achieved. All MRBEE estimates used in this iterative procedure can include intercept terms. All steps in this algorithm can be completed automatically using the `MRBEE` software (see **Software** section).

2.4 Simulation settings

Unless stated otherwise, all simulations used simulated GWAS summary statistics that were drawn from normal distributions specified by parameters that depended on GWAS sample size, sample overlap proportions, phenotypic correlations, and true causal effects that are fully described in Section 3 of the **Supplement**. True causal effects were determined by genetic correlations, exposure and outcome SNP heritability, and any specified horizontally pleiotropic effects. All phenotypic and genotypic variances were equal to 1 so the true variance of any GWAS estimate was the inverse of the respective GWAS sample size. Pleiotropic effects were randomly drawn with specified variances by allowing the final term in Equation 4 (γ_j) to be nonzero for a defined proportion of IVs. In cases of CHP, γ_j was allowed to correlate with some elements in $\boldsymbol{\beta}_j$. See the legends of Figures 3 and 4 for additional details.

2.5 Real Data Analysis 1: Coronary Artery Disease

We performed two real data analyses, the first of which is described here and the second in Section 2.6. In Real Data Analysis I, we estimated direct causal effects of 9 exposures on coronary artery disease (CAD) risk in East Asian (EAS) and European (EUR) populations using multivariable MRBEE and existing alternatives.

2.5.1 GWAS data

EAS exposure GWAS data were provided exclusively by Biobank Japan [45]. EAS CAD GWAS data were provided by [46] (n=212k). EUR exposure GWAS data were provided by a range of consortia as described in the **Supplement**. EUR CAD GWAS data were provided by the CARDIoGRAM consortium (n=184k) [47]. Cardiometabolic risk factors used in multivariable MR for CAD included serum lipids (HDL, LDL, triglycerides), BMI, systolic blood pressure, uric acid, height, HbA1c, and hemoglobin. Hematocrit, diastolic blood pressure, and red blood cell count were initially considered but later excluded from multivariable MR because of high correlations (>0.75) in IV estimates with other exposures. Additional details of the GWAS data used are available in Section 4 of the **Supplement**.

2.5.2 MR instrumental variables

We performed univariable (single exposure) and multivariable (multiple exposure) MR and generally followed the methods of [48] in selecting IVs. Candidate IVs in univariable MR analysis were associated ($P<5\text{x}10^{-8}$) with the exposure in a within-phenotype and between-ancestry fixed-effects GWAS meta-analysis of EAS- and EUR-specific GWAS, had the same sign in the EAS and EUR GWAS, and had at least $P<0.05$ in both GWAS. We then selected only independent ($r^2<0.01$ in a 1Mb window using PLINK v1.9; [49]) SNPs from this set using ancestry-specific linkage disequilibrium (LD) reference panels from

1000 Genomes Phase 3 [50]. In [48], the authors then removed IVs associated ($P< 5x10^{-8}$) with any of 36 other phenotypes, which is a step we did not complete so as to not unnecessarily reduce the exposure variance explained by our IVs. Indeed, our IVs explained more exposure variances than those in [48] (see **Supplement**). Only ancestry-specific GWAS estimates were used in ancestry-specific MR. For multivariable MR in each ancestry group, all IVs from univariable MR were initially considered, then filtered to only include those SNPs that were independent (r^2<0.01 in a 1Mb window using ancestry-specific LD reference panels from 1000 Genomes [50]). As mentioned in the **Results**, there was no evidence of unbalanced horizontal pleiotropy in EAS or EUR using Q_{pleio} (Equation 20) so we did not remove any IVs using S_{pleio} in MRBEE but did include an intercept term in MRBEE. This resulted in 3,097 IVs used in EAS and 2,821 in EUR. Differences in IV counts differed between EAS and EUR only because of the differences in availability of certain SNPs in the EAS and EUR GWASs. Results from alternative selections of the IVs used in MR are available in the **Supplement** and are consistent with those presented in the **Results** section. All GWAS estimates were standardized following the methods in [16].

2.5.3 Genome-wide horizontal pleiotropy testing

Although the $S_{\texttt{pleio}}$ statistic was initially developed to identify SNPs with horizontal pleiotropy evidence in MR, we can also calculate $S_{\texttt{pleio}}$ for all SNPs genome-wide after estimating the causal effects of the 9 exposures on CAD risk. Results from these tests can be used to (i) find novel loci associated with CAD and (ii) draw inference about pathways of genetic association with CAD. Scenarios in which researchers may infer direct, exposure-mediated, and pleiotropic genetic associations with CAD are displayed in Figure 5B. Genome-wide pleiotropy testing with $S_{\texttt{pleio}}$ was performed separately in EAS and EUR. Lead SNPs in independent loci (r^2<0.01, 1Mb, $P<5x10^{-8}$) identified in the original CAD GWAS were compared to the horizontally pleiotropic effects of the same lead SNPs to produce the results in Table 2.

2.6 Real Data Analysis 2: Schizophrenia and Bipolar

In Real Data Analysis II, we estimated direct causal effects of 7 exposures on risk of schizophrenia (SCZ) and bipolar (I or II) disorder (BP) using GWAS data from European populations using multivariable MRBEE and existing alternatives.

2.6.1 GWAS data

We initially considered the following exposures for SCZ and BP: cannabis used disorder (CUD), left handedness (LH), insomnia, sleep duration, BMI, systolic and diastolic blood pressure, Attention-Deficit/Hyperactivity Disorder (ADHD), gestational duration, intelligence, neuroticism (sensitivity to environmental stress and adversity; SESA), cerebral volume, and glutamate and

dopmaine neurological measurements. Only CUD, LH, ADHD, sleep duration, education, intelligence, and neuroticism were considered in multivariable MR because only these had evidence of nonzero total causal effect on either BP or SCZ in univariable MR using MRBEE (see **Supplement**). All GWAS data were from studies in strictly EUR individuals, which are from a range of consortia (see **Supplement**). Exposure GWAS sample sizes ranged from 55k for ADHD [53] to 1.7M for LH [54]. SCZ GWAS data were from meta-analysis performed using data from the Psychiatric Genomics Consortium [51] on 130k EUR individuals (53k cases). BP GWAS data were from [52] that had a total sample size of 413k (42k cases) EUR individuals, where the outcome phenotype was defined as either lifetime Bipolar I or II disorder. More complete descriptions of all GWAS data used in MR are available in the **Supplement**.

2.6.2 MR instrumental variables

Many exposure GWAS could only identify a few (e.g., 2 for CUD) SNPs that reached genome-wide significance ($P<5x10^{-8}$) and were independent ($r^2<0.01$, 1Mb window) with other SNPs using a EUR LD reference panel from 1000G [50]. We therefore considered all independent SNPs with exposure GWAS $P<5x10^{-5}$ in univariable MR analysis. Use of this threshold (vs a stricter threshold of e.g. $P<5x10^{-8}$) may introduce some weak instrument bias, but MRBEE can correct for it. We then excluded 3 IVs whose minor allele frequency differed by more than 0.10 from all other exposures. All independent IVs from univariable MR were initially considered for multivariable MR. We then filtered this list of 3,346 IVs to only include those associated with the 7 exposures selected for multivariable MR at $P<5x10^{-8}$ in a 7-degree of freedom joint test for association with the 7 selected MR exposures. We did not remove any IVs with large S_{pleio} statistics because tests of global unbalanced horizontal pleiotropy using Q_{pleio} (Equation 20) did not suggest any imbalance (see **Results**) for BP or SCZ. This resulted in 504 IVs that were used in multivariable MR. IV estimates were standardized by standard error and not by the methods in [16] because the GWAS sample size for left handedness was considerably larger (n=1.7 million) than those for the remaining exposures. Since the standardization in [16] is based explicitly on sample size, we standardized only by standard error for the results presented in the **Results** section to ensure the scale of causal effect estimates was not heavily influenced by exposure GWAS sample size. Results from standardization using the methods in [16] are presented in the **Supplement** and the inferences are consistent with the results presented in the **Results** section.

2.7 Joint testing of causal effects

In Real Data Analysis II, we performed joint and contrast testing of causal effects for schizophrenia (SCZ) and bipolar disorder (BP) in the following way. Denote $\hat{\mathbf{\Theta}}_{(2)}$ as the $(7+1)\times 2$ matrix of causal effect estimates produced by MRBEE (+1 for intercept terms), where the first column of $\hat{\mathbf{\Theta}}_{(2)}$ corresponds

to the exposure effects on SCZ and the second on BP. Let $\hat{\mathbf{\Lambda}}_{(2)}$ denote the corresponding variance-covariance matrix of $\text{vec}(\hat{\mathbf{\Theta}}_{(2)})$. Since the same IVs were used for SCZ and BP, elements of $\mathbf{\Theta}_{(2)}$ and $\mathbf{\Lambda}_{(2)}$ were estimated using a single multivariate MRBEE model. We wished to test the general linear hypotheses of the following form:

$$H_0 : \mathbf{C}^\top \mathbf{\Theta}_{(2)} \mathbf{L} = \boldsymbol{\mu} \qquad \text{vs} \qquad H_1 : \mathbf{C}^\top \mathbf{\Theta}_{(2)} \mathbf{L} \neq \boldsymbol{\mu}, \tag{21}$$

where $\mathbf{C}$ and $\mathbf{L}$ were respectively $g \times h$ and $s \times l$ fixed design matrices with elements chosen for specific hypotheses that are indicated in the **Results** section. Let $\mathbf{H} = (\mathbf{L} \otimes \mathbf{C})^\top$ and $\hat{\boldsymbol{\vartheta}} = \text{vec}(\hat{\mathbf{\Theta}}_{(2)})$ and we tested different H_0 in Equation 21 using the statistic

$$(\mathbf{H}\hat{\boldsymbol{\vartheta}})^\top (\mathbf{H}\hat{\mathbf{\Lambda}}\mathbf{H}^\top)^{-1} (\mathbf{H}\hat{\boldsymbol{\vartheta}}) \xrightarrow{D} \chi^2(hl). \tag{22}$$

The convergence statement in Equation 22 is proven in the **Supplement**.

2.7.1 Genome-wide horizontal pleiotropy testing

We performed genome-wide horizontal pleiotropy testing with $S_{\texttt{pleio}}$ using all MR exposures with a significant (i.e., $P < 0.05$) causal effect on either SCZ or BP. Including non-significant exposures (i.e., left handedness and ADHD) in genome-wide pleiotropy testing would have effectively only increased the variance term used in $S_{\texttt{pleio}}$ (see Equation 17) and not otherwise affected the inferences we could make. Genome-wide testing with $S_{\texttt{pleio}}$ was performed separately for SCZ and BP. Lead SNPs in independent loci (r^2<0.01, 1Mb, P<5x10^{-8}) in the original SCZ and BP GWAS were used for comparisons with results from genome-wide horizontal pleiotropy testing to produce the results in Table 2.

3 Results

3.1 Bias in MR

In multivariable MR with IVW, the direct causal effects of multiple exposures on an outcome phenotype are estimated with a set of m independent IVs that have evidence of association with at least one of the p exposures and are assumed to not have any horizontally pleiotropic effects. Using the notation in Section 2.1 of **Methods**, the bias in multivariable IVW can be written as the product of two terms:

$$\text{bias}(\hat{\boldsymbol{\theta}}_{\texttt{IVW}}) = \underbrace{\left\{ \mathbf{\Sigma}_{\beta\beta}^{(m)} + \sum_{j=1}^{m} \mathbf{\Sigma}_{\mathbf{w}_{\beta_j}\mathbf{w}_{\beta_j}} \right\}^{-1}}_{\text{Weak instruments bias}} \underbrace{\left\{ \sum_{j=1}^{m} \boldsymbol{\sigma}_{\mathbf{w}_{\beta_j} w_{\alpha_j}} - \mathbf{\Sigma}_{\mathbf{w}_{\beta_j}\mathbf{w}_{\beta_j}} \boldsymbol{\theta} \right\}}_{\text{Sample overlap \& Measurement error}}, \tag{23}$$

where $\boldsymbol{\Sigma}_{\beta\beta}^{(m)}$ is the $p \times p$ matrix of genetic covariance between the exposures using the m IVs, $\boldsymbol{\Sigma}_{\mathbf{w}_{\beta_j}\mathbf{w}_{\beta_j}} := \texttt{Cov}(\mathbf{w}_{\beta_j})$, and $\boldsymbol{\sigma}_{\mathbf{w}_{\beta_j} w_{\alpha_j}} := \texttt{Cov}(\mathbf{w}_{\beta_j}, w_{\alpha_j})$. The first term in Equation 23 represents bias from weak instruments because $m^{-1}\boldsymbol{\Sigma}_{\beta\beta}^{(m)}$ represents the average instrument strength, where it can be observed that each diagonal element in $\boldsymbol{\Sigma}_{\beta\beta}^{(m)}$ represents the SNP heritability explained by the m IVs for each exposure. When there is sample overlap between an exposure and the outcome in GWAS, $\boldsymbol{\sigma}_{\mathbf{w}_{\beta_j} w_{\alpha_j}}$ is nonzero if any of the phenotypic correlations are nonzero [19]. The term $\boldsymbol{\Sigma}_{\mathbf{w}_{\beta_j}\mathbf{w}_{\beta_j}}\boldsymbol{\theta}$ represents the contribution of the measurement error variance (which is inversely proportion to the sample sizes of exposure GWAS [33]) for the exposures unless there are no causal effects of the exposures on the outcome. It is clear from Equation 23 that measurement error biases and bias from sample overlap are working in opposite directions since they have different signs.

These sources of bias are independent of UHP and CHP biases because the values in Equation 23 are completely determined by GWAS sample size, overlap proportions, true phenotypic correlations, and causal effects. These biases can go in any direction when multiple exposures are included in MR, may affect the power of horizontal pleiotropy testing as in MRPRESSO [23] and IMRP [21], Type I/II error rates of causal effect hypothesis testing (**Supplement**), and will increase as the number of IVs increases (Figure 2C). As the number of IVs increases, the IV set on average becomes weaker, thereby enhancing biases from measurement error and sample overlap. When there is only one exposure, the bias term is simplified to

$$\text{bias}(\hat{\theta}_{\text{IVW}}) = \left(\frac{m}{n_1 h_X^2 + m\sigma_X^2}\right)\left(\frac{n_0\sigma_{XY}}{n_2} - \sigma_X^2\theta\right), \tag{24}$$

where n_1 and n_2 are the respective exposure and outcome GWAS sample sizes, n_0 is the number of overlapping GWAS participants, h_X^2 is the SNP heritability in the exposure explained by the m IVs, σ_X^2 is the phenotypic variance of the exposure, and σ_{XY} is the phenotypic covariance between the exposure and outcome (see **Supplement**).

Figure 2A lists the currently available MR methods and their capability to deal with these and other sources of bias. Existing MR methods cannot simultaneously address all known sources of bias. Including multiple genetically correlated exposures in MR may reduce CHP bias from omitted exposures, but will not necessarily reduce biases from weak instruments, sample overlap, or measurement error (Figure 2B). In our simulations, we observed that sample overlap can make the false positive rate for testing no causal effect of the exposure on an outcome approach 100% even if GWAS sample sizes are reasonably large (∼30k) (Figure 2D), suggesting that the implications of these biases can be large.

Figure 3 demonstrates that MRBEE is the only MR method that can estimate the total causal effect of a single exposure without bias as UHP, sample

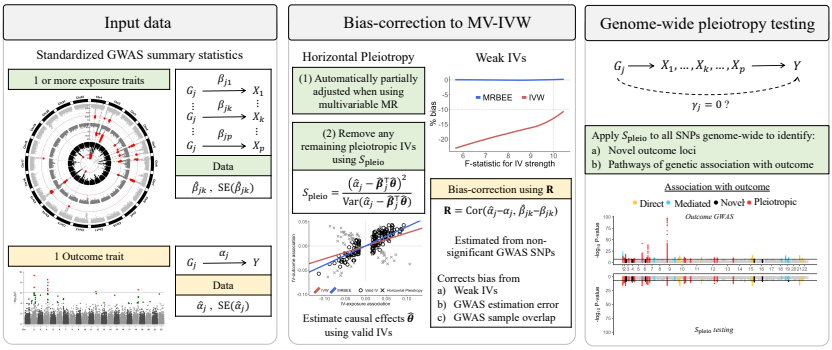


Fig. 1 Flowchart illustrating the MRBEE method. (left) MRBEE uses standarized, but potentially non-harmonized, GWAS summary statistics and can include multiple exposures. (middle) MRBEE corrects for bias from weak IVs, GWAS measurement/estimation error, and sample overlap using a bias-corrected estimator; MRBEE corrects for bias from horizontal pleiotropy by including multiple exposures in MR simultaneously and applying the S_{pleio} test adapted from the IMRP [21] algorithm to remove pleiotropic IVs (explained further in Sections 2.1, 2.3, and 3.1). (right) The S_{pleio} test can also be applied genome-wide, which may find novel outcome loci and explain how some loci are associated with the outcome. The Manhattan plot in the bottom portion of the right panel are results from Real Data Analysis 1 (see **Methods**, **Results**), showing how epidemiological inference from genome-wide pleiotropy testing can be made in practice for coronary artery disease (explained further in Sections 2.3 and 2.5.3).

overlap, GWAS sample sizes, and weak instrument bias varies compared to IVW [22], dIVW [27], weighted median [24], MR-Robust [63], IMRP [21], MR-CML [28], MRMix [16], MR-Corr [20], and MR-CUE [17] methods. MRBEE is therefore the only method of these that has controlled coverage frequencies (Type I error is also controlled, see Supplementary Section S3). MRBEE can also unbiasedly estimate true causal effects when only a small proportion of exposure variance is explained by the IVs (see Supplementary Section S1.1) since MRBEE corrects weak instrument bias. The traditional IVW method is biased in this case [11].

Figure 4A demonstrates that, comapared to the alternative methods included in Figure 3, MRBEE is the only MR method that can estimate direct causal effects without bias in the presence of CHP. Multivariable MR methods (i.e., weighted median, IVW, MR-Robust) are generally less biased than univariable MR methods (i.e., MR-CUE, MR-CML, MR-Corr, MR-Mix), but still they cannot consistently estimate direct causal effects because of uncontrolled biases from weak instruments, measurement error, and sample overlap. Since every other MR method except MRBEE is biased, their coverage proportions are generally low. For example, the coverage proportion for MR-CML is less than 1% in every case in simulations where horizontal pleiotropy, sample overlap, and the number of IVs varied (Figure 4B), which may lead to incorrect causal inference in practice. MRBEE was the only MR method to obtain optimal coverage frequencies in all simulation settings. The univariable MR-Corr often has coverage proportions around or equal to 1 but it also has large standard errors.

3.1.1 Genome-wide application of S_{pleio}

The horizontal pleiotropy test statistic S_{pleio} can also be applied genome-wide to search for genomic loci with evidence of horizontal pleiotropy. The genome-wide application of univariable versions of S_{pleio} has previously successfully identified novel genomic regions undetected by standard GWAS testing [18]. When multiple exposures are included in MR, S_{pleio} can also be used to infer pathways of genetic association with a phenotype. That is, S_{pleio} can identify genomic regions whose association with a phenotype are completely mediated by a set of exposures. Figure 5B shows the four types of inference that can be made with S_{pleio} and Figure 5C demonstrates how S_{pleio} was applied to real data in the study of coronary artery disease (CAD) in Europeans with a set of nine exposures (see below).

3.2 Real Data Analyses

We performed two independent analyses of real data to demonstrate how MRBEE works and compares to existing MR methods in practice. We first estimated direct causal effects of 9 cardiometabolic risk factors on CAD risk in East Asian (EAS) and European (EUR) populations (Real Data Analysis 1). We next estimated direct causal effects of 7 known risk factors for schizophrenia (SCZ) or bipolar disorder (BP) in a European population (Real Data Analysis 2). We compared causal estimates from MRBEE to IVW* (defined below), MR-Robust (equivalent to MR-Lasso), and multivariable weighted median methods since these are the only multivariable MR methods that exist. IVW* refers to the multivariable IVW causal estimates after removing specific IVs with significant S_{pleio} statistics at a FDR-adjusted (for the number of IVs) Type I error rate of 5% using Benjamini-Hochberg [34] in the MRBEE procedure.

3.2.1 Real Data Analysis 1 (Coronary Artery Disease)

Univariable MR results suggested nonzero causal effects of all exposures on CAD in EAS or EUR populations. However, there was widespread evidence of unbalanced horizontal pleiotropy as indicated by large differences in causal estimates between UHP-adjusted and UHP-naïve MR methods. For example, the causal effect of DBP on CAD in EAS was estimated to be 2.03 (OR; P=$2.8\text{x}10^{-11}$) using IMRP but only 1.43 (P=0.140) using MR-Egger, where estimates from these two methods differ only by how UHP is addressed. Full univariable MR results are presented in the **Supplement**.

Multivariable MR estimates (see Figure 6A) using MRBEE were generally consistent between EAS and EUR populations. All 9 exposures had evidence of nonzero causal effect on CAD in EAS or EUR in population-specific testing. IVW*, MR-Robust, and weighted median generally had similar causal estimates. LDL had the strongest causal effect estimate in both EAS (MRBEE OR=2.09; P=$1.1\text{x}10^{-20}$) and EUR (OR=1.76, P=$4.1\text{x}10^{-20}$), which was undetected for EUR in [48] because of bias. Unlike in [48], we included HDL

in multivariable MR and found negative nonzero effects on CAD in EAS (OR=0.85, P=0.011) and EUR (OR=0.77, P=3.7x10^{-5}). In EAS, all other multivariable MR methods may underestimated the direct causal effect of LDL on CAD. For example, IVW* and MR-Robust produced odds ratio estimates of 1.39 (P=2.2x10^{-25}) and 1.46 (P=1.2x10^{-22}). The direct causal effect of SBP on CAD in EAS was similarly underestimated by IVW* compared to MRBEE (MRBEE OR=2.29, P=1.6x10^{-11} vs IVW* OR=1.56, P=5.0x10^{-23}).

In EAS, the total causal effect of BMI on CAD (MRBEE OR=1.44, =2.0x10^{-25}) was completely mediated by SBP (P=0.220 for total mediation; see **Supplement**). Evidence of nonzero direct causal effect of BMI on CAD in EUR may be explained by the GWAS data used for SBP. In EUR, the SBP GWAS included BMI as a covariate and so SBP could not statistically act as a mediator for BMI in multivariable MR with CAD. The BMI result displayed in Figure 6A therefore reflects the effect of BMI on CAD through all other exposures except SBP. This phenomenon – that including one exposure as a covariate in the GWAS for another can preclude consistent direct causal effect estimation in multivariable MR – is confirmed in simulation (see **Supplement**) and reported in [58].

Differences between MRBEE and IVW* causal estimates were strongly correlated with the bias in Equation 23 due to weak instruments, sample overlap, and measurement error (EAS Pearson r=0.92, P=4.6x10^{-4}; EUR r=0.65, P=0.058; Figure 7A). Since there was no evidence of unbalanced horizontal pleiotropy (EAS P=0.528 for a test of balance using Q_{pleio}, see **Methods**; EUR P=0.231), this suggested that differences between IVW* and MRBEE causal estimates were due to uncontrolled bias in IVW*. Since causal estimates made by IVW* were generally very similar to those made by MR-Robust and weighted median methods, a similar interpretation can be made for them, too.

3.2.2 Real Data Analysis 2 (Schizophrenia and Bipolar Disorder)

Univariable MR results suggested nonzero total causal effects of cannabis use disorder (CUD), Attention-Deficit/Hyperactivity Disorder (ADHD), left handedness, neuroticism, sleep duration, intelligence, and education on BP or SCZ. Consistent with [35] but not [36, 37], we found a strong negative causal effect of left handedness on BP risk (MRBEE OR=0.70, P=8.9x10^{-34}), which is of opposite sign for SCZ (OR=1.36, P=6.2x10^{-24}). Whether this is the result of inherent bias in univariable MR cannot be confirmed so full univariable results are reserved for the **Supplement**.

Multivariable MR with MRBEE and exposures CUD, ADHD, left handedness, neuroticism, sleep duration, intelligence, and education identified nonzero causal effects for all exposures on BP and/or SCZ except ADHD and left handedness in outcome-specific testing. IVW*, MR-Robust, and weighted median MR methods generally produced very similar causal estimates. Compared to MRBEE, IVW* underestimated the direct causal effect of CUD on SCZ, where IVW* and MRBEE respectively produced odds ratio estimates of 2.02

(P=2.2x10^{-16}) and 4.00 (P=1.8x10^{-7}), the latter of which is most consistent with the literature (i.e., OR for CUD association is 3.90, 95% CI: 2.84-5.34 in [40]). IVW* additionally underestimated causal effects for education, intelligence, left handedness, and neuroticism compared to MRBEE for SCZ and BP (see Figure 6B). CUD had the largest causal effect on BP and SCZ, explaining as much as 16% and 39% of the genetic variance in BP and SCZ, respectively (see **Methods**).

Differences between IVW* and MRBEE causal estimates were almost perfectly correlated with the bias in Equation 23 due to weak instruments, sample overlap, and measurement error (Figure 7B; BP Pearson r=0.96, P=5.1x10^{-4}; SCZ r=0.97, P=3.2x10^{-4}). Since there was no evidence of unbalanced horizontal pleiotropy (global P=0.625 for joint test of balance for BP or SCZ using Q_{pleio}), this suggested that MRBEE appropriately adjusted for these sources of bias while the IVW*, MR-Robust, and weighted median methods did not.

Follow-up testing using Equation 21 in Methods indicated that all exposures except left handedness and ADHD had nonzero direct causal effects on BP or SCZ (see **Methods**). These 7 exposures on average have stronger causal effects on SCZ than BP (P=2.5x10^{-5}), though the effect of sleep duration (P=0.698) and ADHD (P=0.415) on BP and SCZ may be similar (P-values correspond to a test of difference). There was no evidence that modifiable risk factors including CUD, sleep duration, and education had in total greater direct causal effects than the largely non-modifiable risk factors (all other exposures) on both BP (P=0.265 for difference) and SCZ (P=0.267).

3.2.3 Genome-wide Horizontal Pleiotropy Testing

We finally performed genome-wide horizontal pleiotropy testing for CAD, SCZ and BP using S_{pleio}, for which a summary of the results are displayed in Table 2.

Of the 65 and 39 loci (1Mb, r^2<0.01, P<5x10^{-8}) identified by standard GWAS testing for CAD in EAS and EUR, respectively, horizontal pleiotropy testing correspondingly identified 27 ($\lambda_{GC}^{\text{pleio}}$=1.08 vs $\lambda_{GC}^{\text{GWAS}}$=1.16) and 41 ($\lambda_{GC}^{\text{pleio}}$=1.01 vs $\lambda_{GC}^{\text{GWAS}}$=1.00 in GWAS) loci. In EUR, nine loci that were detected in horizontal pleiotropy testing were not detected in the original CAD GWAS. Seven of these were replicated (P<0.05 for lead SNP) in an independent CAD GWAS in EUR (UKBB; [56]), all of which could only be detected in a larger CAD GWAS [38]. In EUR and EAS, we respectively identified only 10 and 18 loci that are directly associated (i.e., not indirectly through the MR exposures) with CAD and 19 (EUR) and 5 (EAS) with evidence of simultaneous association (i.e., horizontal pleiotropy) with the MR exposures and CAD conditional on the exposures.

We also identified one locus for SCZ using S_{pleio} genome-wide that was not discovered in the original SCZ GWAS (i.e., GWAS P=9.3x10^{-3}, S_{pleio} P=3.5x10^{-8}). This locus maps to the *ATXN2L* gene which is known to be associated with intelligence [41], brain morphology [42], and autoimmune disease [43]. Of the 152 and 44 loci identified for SCZ and BP in the original GWAS,

151 (98%) and 42 (95%) were not associated at the level of genome-wide significance respectively with SCZ and BP through pathways that did not include CUD, intelligence, education, sleep duration, or neuroticism (see **Methods**). This suggests that a large portion of the heritability of these two traits may be due to genetic associations with their causal risk factors and not directly with the psychiatric disorders themselves.

4 Discussion

Our study provides strong evidence that the existing univariable and multivariable MR approaches are vulnerable to biases from weak instruments, measurement error, UHP, CHP, sample overlap, and omitted exposures. One suggested solution to this problem that is currently being heavily practiced in the literature is to use multiple MR methods and appraise the evidence in aggregate more highly than evidence from any one method alone [39]. Our applications of MRBEE to real data demonstrated that multiple MR methods can be biased in similar ways, rendering any aggregated inference from multiple biased methods no less subject to mistake than inference from any one method alone. In contrast, the multivariable MRBEE we developed here is generally robust to the above biases and may be a useful tool in practice.

We demonstrated the practical utility of MRBEE in two independent applications to the study of CAD in (i) EAS and EUR and (ii) schizophrenia (SCZ) and bipolar disorder (BP). MRBEE was able to explain and correct inconsistent causal estimates made by existing MR methods. Causal risk factors were generally consistent for CAD between EAS and EUR and between SCZ and BP in EUR, where any differences between MRBEE estimates and those made by alternative methods were the results of uncontrolled bias in other methods. For example, the causal estimate of LDL on CAD in EAS was expected to have 55.3% downward bias from Equation 23 and indeed the horizontal pleiotropy-robust IVW causal estimate was 55.7% smaller than the MRBEE estimate (Figure 6A). In Real Data Analysis 1 with CAD, we observed that the total causal effect of BMI on CAD was completely mediated by blood pressure and partially by uric acid in EAS, though the GWAS data in EUR precluded testing of this kind. In Real Data Analysis 2 with schizophrenia (SCZ) and bipolar (BP), we observed that cannabis use disorder (CUD) has very large direct causal effects on SCZ and BP risk (consistent with the literature [40], but that existing MR methods may underestimate the sizes of these effects.

We finally introduced a multivariable horizontal pleiotropy test (using statistic S_{pleio}) that, when applied genome-wide, identified the epidemiological pathways through which many genomic loci were associated with CAD in EAS and EUR, SCZ, and BP. The majority of known genetic associations with all disease endpoints were non-direct (Table 2), suggesting that a large portion of the heritability of these complex traits may be conferred indirectly through their causal risk factors. This test also identified 9 novel loci (undetected in standard GWAS) for CAD in EUR – seven of which were replicated

in UKBB – and one for SCZ in EUR, for which no adequate independent replication data exists. This method of pleiotropy testing using S_{pleio} is therefore a valuable tool both for gaining better insight into how genetic risk of disease is conferred and in detecting new risk loci.

MRBEE has the following limitations. As with all MR methods, the reliability of causal estimates produced by MRBEE depend on the quality of GWAS data used in MR. For example, biases in GWAS from assortative mating or dynastic effects may propagate through to MR and bias causal estimation [60, 61]. Second, the extent to which differences between univariable and multivariable causal estimates differ because of different total vs direct effect sizes, CHP, or the different sets of IVs used can be difficult to determine in practice. Third, high multicollinearity in our real data analyses prevented us from including some exposures we otherwise would have if MRBEE was not vulnerable to this problem. Future work that can expand the application of MRBEE to the high-dimensional setting may help address this challenge. Fourth, MRBEE may not be able to identify CHP that neither acts through the included MR exposures nor behaves as outliers in regression. We note that no existing method can fully address CHP in this scenario, but indeed it may be the subject of future work.

In conclusion, single-exposure MR is inherently limited in its ability to reduce bias or provide precise epidemiological inference, but univariable MR methods and their application have so far dominated the literature compared to multivariable analyses. We developed multivariable MRBEE to reduce known biases in MR and estimate direct causal effects of multiple exposures in robust way. MRBEE can be a useful tool in studying causality between risk factors and disease outcomes as more large GWAS summary statistics are made publicly available.

4.1 Software

The software used to perform all simulations and analyze the real data used above is available at github.com/noahlorinczcomi/MRBEE and http://hal.case.edu/~xxz10/zhu-web/. The software contains all functions needed to use MRBEE and perform all their associated tests in practice.

Supplementary information. Please refer to the **Supplement** for additional derivations, simulation results, and details of real data analyses.

Acknowledgments. We acknowledge the help of Zhengxi Chen in providing the authors with background literature on the association of cannabis use with schizophrenia and bipolar disorder risk.

Declarations

4.2 Funding

This work was supported by grant HG011052 (to XZ) from the National Human Genome Research Institute (NHGRI). NLC was partially supported by grant T32 HL007567 from the National Heart, Lung, and Blood Institute (NHLBI).

4.3 Conflict of interest

No authors have any relevant conflicts of interest to disclose.

4.4 Ethics approval

The study was approved by the institutional review board (IRB number: STUDY20180592) at Case Western Reserve University.

4.5 Consent to participate

Not applicable.

4.6 Consent for publication

Not applicable.

4.7 Availability of data and materials

All GWAS data used for the analyses were retrieved from publicly available repositories whose online locations are presented in Supplementary Tables 1S and 6S. Genomic loci detected in either the original genome-wide association studies or in genome-wide horizontal pleiotropy testing in Real Data Analyses 1 & 2 are available at github.com/noahlorinczcomi/MRBEE.

4.8 Code availability

R [57] code used in simulations and real data analyses are available at github.com/noahlorinczcomi/MRBEE.

4.9 Software availability

The MRBEE software, written in both the R language [57] and Python, is available at github.com/noahlorinczcomi/MRBEE.

4.10 Author's contributions

NLC, YY, and XZ concevied of and developed MRBEE, NLC and YY performed simulations, NLC performed all real data analyses with the support of GL for Real Data Analysis 2, NLC drafted the manuscript and XZ and YY revised it. XZ provided guidance and support in all aspects of this work.

References

[1] Burgess, S., Timpson, N., Ebrahim, S. & Davey Smith, G. Mendelian randomization: where are we now and where are we going?. *International Journal Of Epidemiology.* **44** pp. 379-388 (2015)

[2] Sanderson, E., Glymour, M., Holmes, M., Kang, H., Morrison, J., Munafò, M., Palmer, T., Schooling, C., Wallace, C., Zhao, Q. & Others Mendelian randomization. *Nature Reviews Methods Primers.* **2**, 1-21 (2022)

[3] Zhu, X. Mendelian randomization and pleiotropy analysis. *Quantitative Biology.* pp. 1-11 (2020)

[4] Gill, D., Georgakis, M., Walker, V., Schmidt, A., Gkatzionis, A., Freitag, D., Finan, C., Hingorani, A., Howson, J., Burgess, S. & Others Mendelian randomization for studying the effects of perturbing drug targets. *Wellcome Open Research.* **6** (2021)

[5] Graaf, A., Claringbould, A., Rimbert, A., Westra, H., Li, Y., Wijmenga, C. & Sanna, S. Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids. *Nature Communications.* **11**, 1-12 (2020)

[6] VanderWeele, T., Tchetgen, E., Cornelis, M. & Kraft, P. Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.).* **25**, 427 (2014)

[7] Kang, H., Zhang, A., Cai, T. & Small, D. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal Of The American Statistical Association.* **111**, 132-144 (2016)

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

[9] Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N. & Thompson, J. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *International Journal Of Epidemiology.* **45**, 1961-1974 (2016)

[10] Burgess, S., Davies, N. & Thompson, S. Bias due to participant overlap in two-sample Mendelian randomization. *Genetic Epidemiology.* **40**, 597-608 (2016)

[11] Burgess, S., Thompson, S. & Collaboration, C. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal Of Epidemiology.* **40**, 755-764 (2011)

[12] Sanderson, E., Davey Smith, G., Windmeijer, F. & Bowden, J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal Of Epidemiology.* **48**, 713-727 (2019)

[13] Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M. & Others A cross-population atlas of genetic associations for 220 human phenotypes. *Nature Genetics.* **53**, 1415-1424 (2021)

[14] Bulik-Sullivan, B., Finucane, H., Anttila, V., Gusev, A., Day, F., Loh, P., Duncan, L., Perry, J., Patterson, N., Robinson, E. & Others An atlas of genetic correlations across human diseases and traits. *Nature Genetics.* **47**, 1236-1241 (2015)

[15] Grant, A. & Burgess, S. Pleiotropy robust methods for multivariable Mendelian randomization. *Statistics In Medicine.* **40**, 5813-5830 (2021)

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

[17] Cheng, Q., Zhang, X., Chen, L. & Liu, J. Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. *Nature Communications.* **13**, 6490 (2022)

[18] Zhu, X., Zhu, L., Wang, H., Cooper, R. & Chakravarti, A. Genome-wide pleiotropy analysis identifies novel blood pressure variants and improves its polygenic risk scores. *Genetic Epidemiology.* **46**, 105-121 (2022)

[19] LeBlanc, M., Zuber, V., Thompson, W., Andreassen, O., Frigessi, A. & Andreassen, B. A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC Genomics.* **19**, 1-15 (2018)

[20] Cheng, Q., Qiu, T., Chai, X., Sun, B., Xia, Y., Shi, X. & Liu, J. MR-Corr2: a two-sample Mendelian randomization method that accounts for correlated horizontal pleiotropy using correlated instrumental variants. *Bioinformatics.* **38**, 303-310 (2022)

[21] Zhu, X., Li, X., Xu, R. & Wang, T. An iterative approach to detect pleiotropy and perform Mendelian Randomization analysis using GWAS summary statistics. *Bioinformatics.* **37**, 1390-1400 (2021)

[22] Burgess, S. & Bowden, J. Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods. *ArXiv Preprint ArXiv:1512.04486*. (2015)

[23] Verbanck, M., Chen, C., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*. **50**, 693-698 (2018)

[24] Bowden, J., Davey Smith, G., Haycock, P. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*. **40**, 304-314 (2016)

[25] Burgess, S., Foley, C., Allara, E., Staley, J. & Howson, J. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*. **11**, 1-11 (2020)

[26] Sanderson, E., Spiller, W. & Bowden, J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Statistics In Medicine*. **40**, 5434-5452 (2021)

[27] Ye, T., Shao, J. & Kang, H. Debiased inverse-variance weighted estimator in two-sample summary-data Mendelian randomization. *The Annals Of Statistics*. **49**, 2079-2100 (2021)

[28] Xue, H., Shen, X. & Pan, W. Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal Of Human Genetics*. **108**, 1251-1269 (2021)

[29] Burgess, S. & Thompson, S. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal Of Epidemiology*. **181**, 251-260 (2015)

[30] Rees, J., Wood, A. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics In Medicine*. **36**, 4705-4718 (2017)

[31] Grant, A. & Burgess, S. Pleiotropy robust methods for multivariable Mendelian randomization. *Statistics In Medicine*. **40**, 5813-5830 (2021)

[32] Lin, Z., Xue, H. & Pan, W. Robust multivariable Mendelian randomization based on constrained maximum likelihood. *The American Journal Of Human Genetics*. (2023)

[33] Yang, Y., Lorincz-Comi, N. & Zhu, X. Unbiased estimation and asymptotically valid inference in multivariable Mendelian randomization with many weak instrumental variables. *ArXiv Preprint ArXiv:2301.05130*. (2023)

[34] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal Of The Royal Statistical Society: Series B (Methodological)*. **57**, 289-300 (1995)

[35] Scully, P., Owens, J., Kinsella, A. & Waddington, J. Epidemiology and pathobiology of bipolar disorder, and their exploration within a complete catchment area population. *Acta Neuropsychiatrica*. **12**, 73-76 (2000)

[36] Bellani, M., Marzi, C., Savazzi, S., Perlini, C., Cerruti, S., Ferro, A., Marinelli, V., Sponda, S., Rambaldelli, G., Tansella, M. & Others Laterality effects in schizophrenia and bipolar disorder. *Experimental Brain Research*. **201**, 339-344 (2010)

[37] Savitz, J., Van Der Merwe, L., Solms, M. & Ramesar, R. Lateralization of hand skill in bipolar affective disorder. *Genes, Brain And Behavior*. **6**, 698-705 (2007)

[38] Aragam, K., Jiang, T., Goel, A., Kanoni, S., Wolford, B., Weeks, E., Wang, M., Hindy, G., Zhou, W., Grace, C. & Others Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *MedRxiv*. (2021)

[39] Burgess, S., Smith, G., Davies, N., Dudbridge, F., Gill, D., Glymour, M., Hartwig, F., Holmes, M., Minelli, C., Relton, C. & Others Guidelines for performing Mendelian randomization investigations. *Wellcome Open Research*. **4** (2019)

[40] Marconi, A., Di Forti, M., Lewis, C., Murray, R. & Vassos, E. Meta-analysis of the association between the level of cannabis use and risk of psychosis. *Schizophrenia Bulletin*. **42**, 1262-1269 (2016)

[41] Sniekers, S., Stringer, S., Watanabe, K., Jansen, P., Coleman, J., Krapohl, E., Taskesen, E., Hammerschlag, A., Okbay, A., Zabaneh, D. & Others Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nature Genetics*. **49**, 1107-1112 (2017)

[42] Meer, D., Frei, O., Kaufmann, T., Shadrin, A., Devor, A., Smeland, O., Thompson, W., Fan, C., Holland, D., Westlye, L. & Others Understanding the genetic determinants of the brain with MOSTest. *Nature Communications*. **11**, 1-9 (2020)

[43] Li, Y., Li, J., Zhao, S., Bradfield, J., Mentch, F., Maggadottir, S., Hou, C., Abrams, D., Chang, D., Gao, F. & Others Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine.* **21**, 1018-1027 (2015)

[44] Burgess, S. & Thompson, S. Interpreting findings from Mendelian randomization using the MR-Egger method. *European Journal Of Epidemiology.* **32**, 377-389 (2017)

[45] Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T. & Others Overview of the BioBank Japan Project: study design and profile. *Journal Of Epidemiology.* **27**, S2-S8 (2017)

[46] Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., Sakaue, S., Matoba, N., Low, S., Okada, Y. & Others Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nature Genetics.* **52**, 669-679 (2020)

[47] Nikpey, M., Goal, A., Won, H., Hall C. Willenberg, L.M., Kanoni, S., Saleheen, D. & Others A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics.* **47**, 1121-1130 (2015)

[48] Wang, K., Shi, X., Zhu, Z., Hao, X., Chen, L., Cheng, S., Foo, R. & Wang, C. Mendelian randomization analysis of 37 clinical factors and coronary artery disease in East Asian and European populations. *Genome Medicine.* **14**, 1-15 (2022)

[49] Chang, C., Chow, C., Tellier, L., Vattikuti, S., Purcell, S. & Lee, J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* **4**, s13742-015 (2015)

[50] Fairley, S., Lowy-Gallego, E., Perry, E. & Flicek, P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research.* **48**, D941-D947 (2020)

[51] Trubetskoy, V., Pardiñas, A., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T., Bryois, J., Chen, C., Dennison, C., Hall, L. & Others Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature.* **604**, 502-508 (2022)

[52] Mullins, N., Forstner, A., O'Connell, K., Coombes, B., Coleman, J., Qiao, Z., Als, T., Bigdeli, T., Børte, S., Bryois, J. & Others Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics.* **53**, 817-829 (2021)

[53] Demontis, D., Walters, R., Martin, J., Mattheisen, M., Als, T., Agerbo, E., Baldursson, G., Belliveau, R., Bybjerg-Grauholm, J., Bækvad-Hansen, M. & Others Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics*. **51**, 63-75 (2019)

[54] Cuellar-Partida, G., Tung, J., Eriksson, N., Albrecht, E., Aliev, F., Andreassen, O., Barroso, I., Beckmann, J., Boks, M., Boomsma, D. & Others Genome-wide association study identifies 48 common genetic variants associated with handedness. *Nature Human Behaviour*. **5**, 59-70 (2021)

[55] Zhu, X., Feng, T., Tayo, B., Liang, J., Young, J., Franceschini, N., Smith, J., Yanek, L., Sun, Y., Edwards, T. & Others Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal Of Human Genetics*. **96**, 21-36 (2015)

[56] Neale Lab. UKBB GWAS round 2, http://www.nealelab.is/uk-biobank (2015)

[57] R Core Team R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing,2019), https://www.R-project.org

[58] Gilbody, J., Borges, M., Davey Smith, G. & Sanderson, E. Multivariable MR can mitigate bias in two-sample MR using covariable-adjusted summary associations. *MedRxiv*. pp. 2022-07 (2022)

[59] Mounier, N. & Kutalik, Z. Bias correction for inverse variance weighting Mendelian randomization. *BioRxiv*. pp. 2021-03 (2021)

[60] Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F., Harrison, S., Vie, G., Cho, Y., Howe, L., Hughes, A., Boomsma, D. & Others Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications*. **11**, 3519 (2020)

[61] Hartwig, F., Davies, N. & Davey Smith, G. Bias in Mendelian randomization due to assortative mating. *Genetic Epidemiology*. **42**, 608-620 (2018)

[62] Burgess, S., Butterworth, A. & Thompson, S. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*. **37**, 658-665 (2013)

[63] Burgess, S., Foley, C., Allara, E., Staley, J. & Howson, J. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*. **11**, 376 (2020)

[64] Rees, J., Wood, A., Dudbridge, F. & Burgess, S. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PloS One*. **14**, e0222362 (2019)

[65] Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T. & Others Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*. **54**, 437-449 (2022)

[66] Kanoni, S., Graham, S., Wang, Y., Surakka, I., Ramdas, S., Zhu, X., Clarke, S., Bhatti, K., Vedantam, S., Winkler, T. & Others Implicating genes, pleiotropy, and sexual dimorphism at blood lipid loci through multi-ancestry meta-analysis. *Genome Biology*. **23**, 1-42 (2022)

[67] Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A., Jiang, Y., Raghavan, S. & Others A saturated map of common genetic variants associated with human height. *Nature*. **610**, 704-712 (2022)

[68] Sadreev, I., Elsworth, B., Mitchell, R., Paternoster, L., Sanderson, E., Davies, N., Millard, L., Smith, G., Haycock, P., Bowden, J. & Others Navigating sample overlap, winner's curse and weak instrument bias in Mendelian randomization studies using the UK Biobank. *MedRxiv*. (2021)

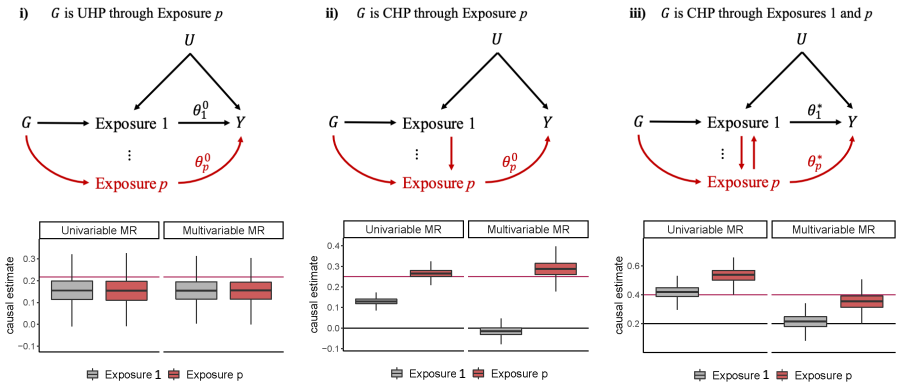
[69] Stock, J. & Trebbi, F. Retrospectives: Who invented instrumental variable regression?. *Journal Of Economic Perspectives*. **17**, 177-194 (2003)

[70] Yi, G. Statistical analysis with measurement error or misclassification: strategy, method and application. (Springer,2017)

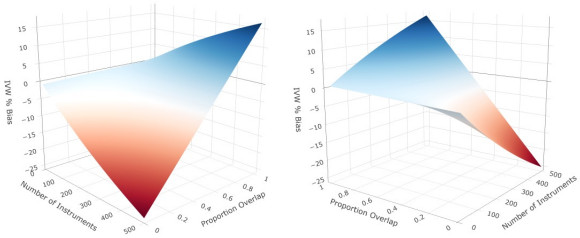
(A) Capabilities of existing MR methods in different conditions

	Weak IVs	Sample overlap	Measurement error	UHP	CHP	Additional exposures
Single exposure methods						
d-IVW	Y	N	Y	N	N	N
MRMix	N	N	Y	Y	Y	N
IMRP/MRPRESSO	N	N	N	Y	O_a	N
CAUSE/MR-CUE	Y	Y	Y	Y	O_b	N
MR-Corr	N	N	Y	Y	O_b	N
MR-CML	N	N	Y	Y	N	N
MR-Lap	Y	Y	Y	N	N	N
Multiple exposure methods						
IVW/MR-Egger	N	N	N	N	N	Y
MR-Robust/MR-Lasso	N	N	N	Y	O_c	Y
Weighted median	N	N	N	N	O_c	Y
MRBEE	Y	Y	Y	Y	O_c	Y

(B) Bias in univariable MR compared to multivariable MR



(C) Theoretical bias of univariable IVW



(D) Univariable IVW Type I Error

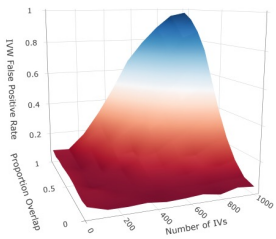


Fig. 2 **(A)**: Bias addressed by currently available MR methods. 'N': cannot address. 'Y': can fully address. To be 'Y', 'O_a' requires that all CHP IVs resemble outliers, 'O_b' requires proportion of CHP IVs to be <0.10, and 'O_c' requires that all genetically correlated exposures with CHP effects are included in MR or that all CHP resembles outliers. **(B)**: Situations in which univariable MR with IVW cannot reliably estimate direct causal effects. Multivariable IVW can more reliably estimate direct causal effects, but still suffers from large bias. Horizontal gray and red lines respectively indicate true direct effects of exposure 1 and p. Boxplots are causal estimates from simulation with true relationships represented by the directed acyclic graphs. **(C)**: Relative bias (bias divided by true causal effect times 100) using Equation 23 is displayed for different numbers of IVs and overlap proportions when the exposure and outcome GWAS sample sizes are 30k.

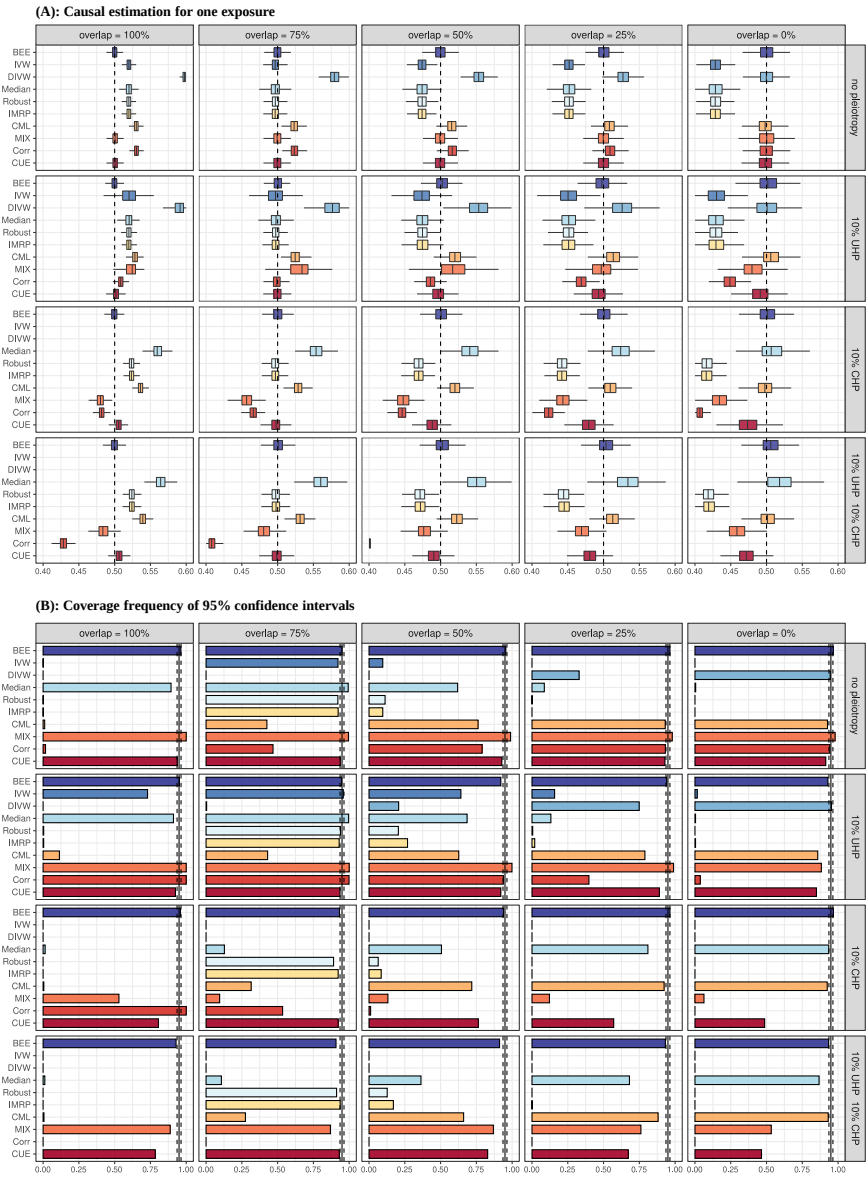


Fig. 3 **(A)** Biases of causal effect estimation in simulation with one exposure with true total causal effect θ =0.5 (indicated by the vertical dashed line) and there are 1,000 IVs. Sample sizes of exposure and outcome GWAS were both 20k, phenotypic correlation was 0.5 between the random errors in the exposure and outcome, and exposure/outcome SNP heritability was 0.3/0.15. "10% UHP" means that 10% of IVs were randomly selected to have UHP effects with variance 4/m, "10% CHP" means that 10% of IVs were randomly selected to have CHP effects with variance 4/m, and "10% UHP 10% CHP" means that 10% of IVs were randomly selected to have UHP effects with variance 4/m and an alternative 10% of IVs were randomly selected to have CHP effects with variance 4/m. To make UHP and CHP resemble outliers, we generated UHP and CHP only in the lower and upper 5% tail regions of the corresponding normal distribution. **(B)** 95% confidence interval coverage frequencies for simulation with settings described in in caption for panel **A**.

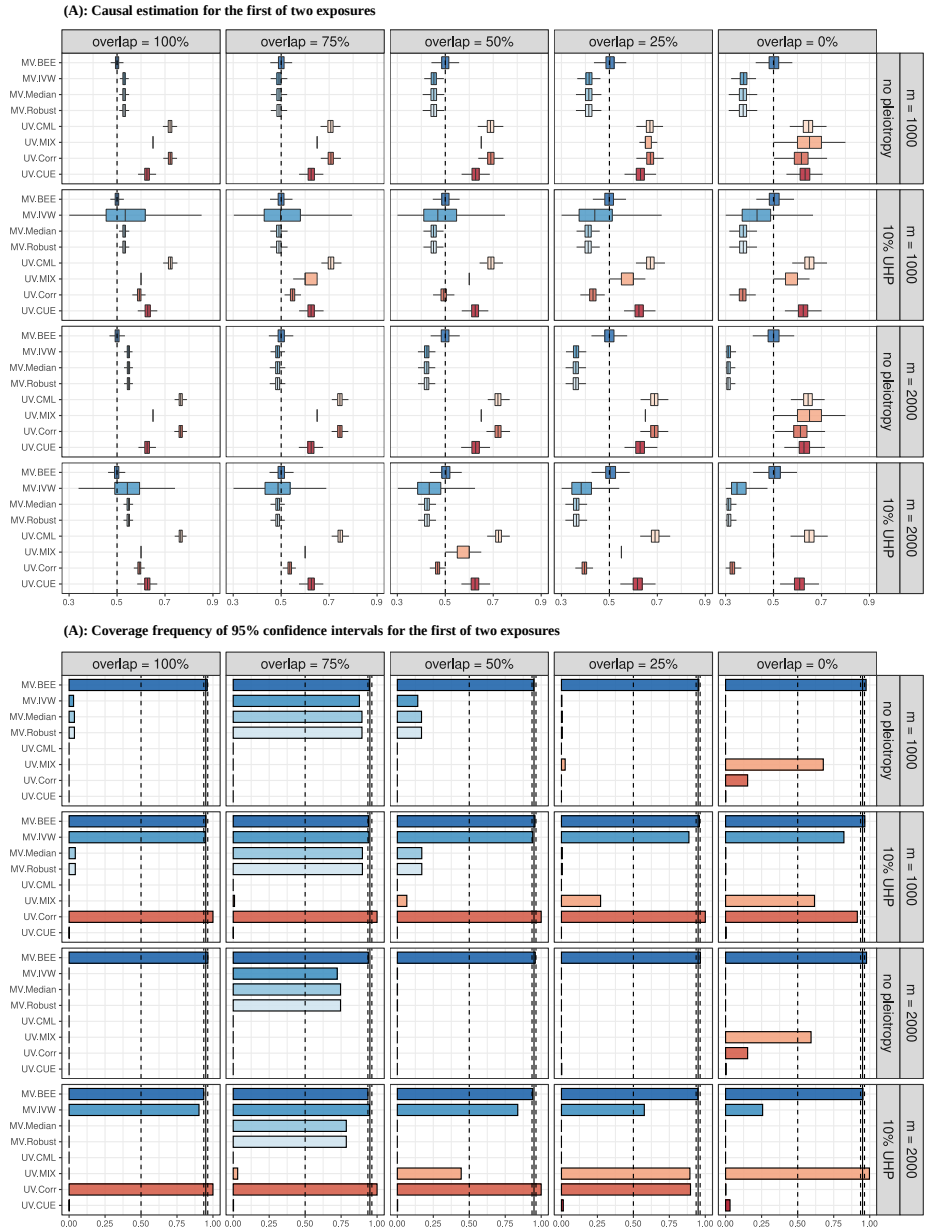


Fig. 4 **(A)**: Bias in causal estimation of the direct effect for the first of two true and genetically correlated exposures and one outcome. The true causal effect is indicated by the vertical dashed line (0.5). MR methods that could only include exposure 1 in MR have 'UV.' prefixes; multivariable methods that included both exposures have 'MV.' prefixes. We simulated two exposures with nonzero causal effects on a single outcome by randomly drawing GWAS summary statistics for GWAS samples of size 20k, phenotypic and genetic correlations of 0.5 between all phenotypes, and exposure/outcome heritability of 0.30/0.15. '10% UHP' means that 10% of IVs had UHP effects with variance 16/m. CHP is automatically introduced in univariable methods (UV.CML, UV.MIX, UV.CORR, UV.CUE) via the omitted exposure (see Supplementary Section S1.1.1 for full settings). **(B)**: Proportions of simulations in which the estimated 95% confidence interval of the causal estimate contained the true direct effect of exposure 1 (0.5).

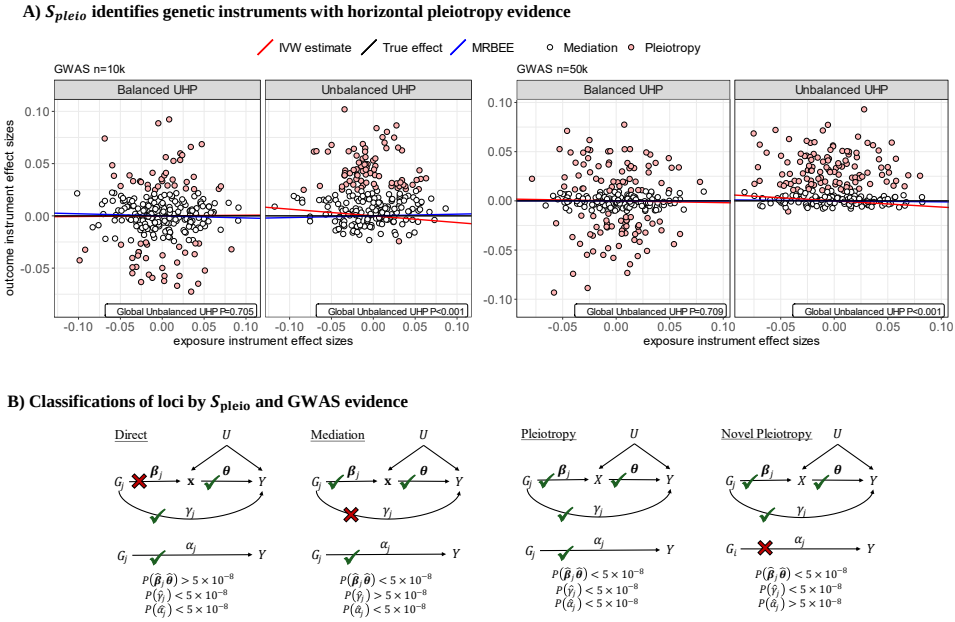


Fig. 5 **(A)**: Demonstration of how horizontal pleiotropy IVs are identified in MRBEE with the IMRP algorithm. A global test for balance is performed using statistic Q_{pleio} in Equation 20. An IV-specific test for horizontal pleiotropy is performed using statistic S_{pleio} in Equation 17. IVs represented by red points had a significant ($P<0.05/m$) S_{pleio} statistic and are removed from causal estimation; white points have small S_{pleio} and are used to estimate the causal effect. **(B)**: Classifications of genomic loci by evidence from original GWAS and genome-wide pleiotropy testing using S_{pleio}. Classifications are based on P-values [denoted as $P(\cdot)$] for testing null hypotheses of equality with 0 for a given parameter in practice. We display the standard threshold of $P<5\mathrm{x}10^{-8}$ for inference, but researchers can choose their own.

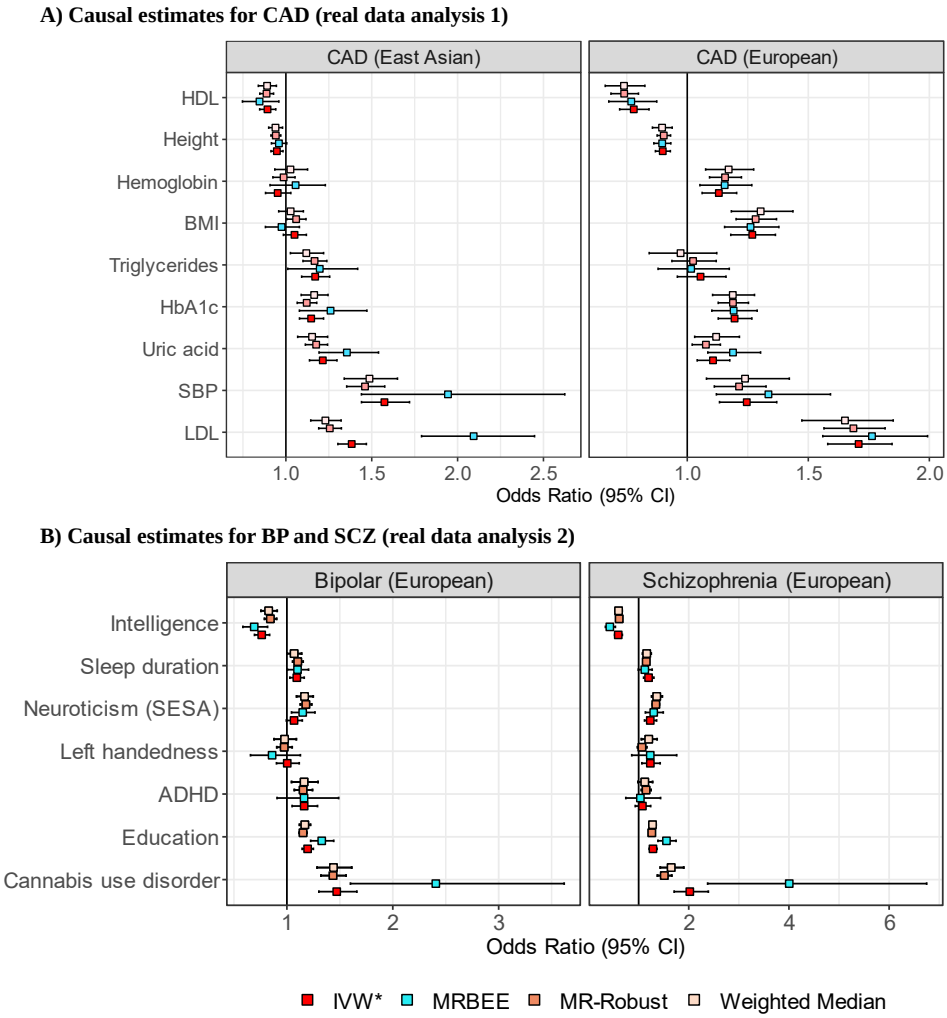


Fig. 6 Direct causal estimates from multivariable MR are obtained from IVs whose selection is described in **Methods**. 'IVW*' represents the multivariable IVW estimator after removing specific IVs with evidence of horizontal pleiotropy using S_{pleio} in MRBEE, adjusted for multiple testing using FDR. 'IVW*' is therefore more robust against UHP than IVW. We found no evidence of unbalanced horizontal pleiotropy in any analyses (see **Results**).

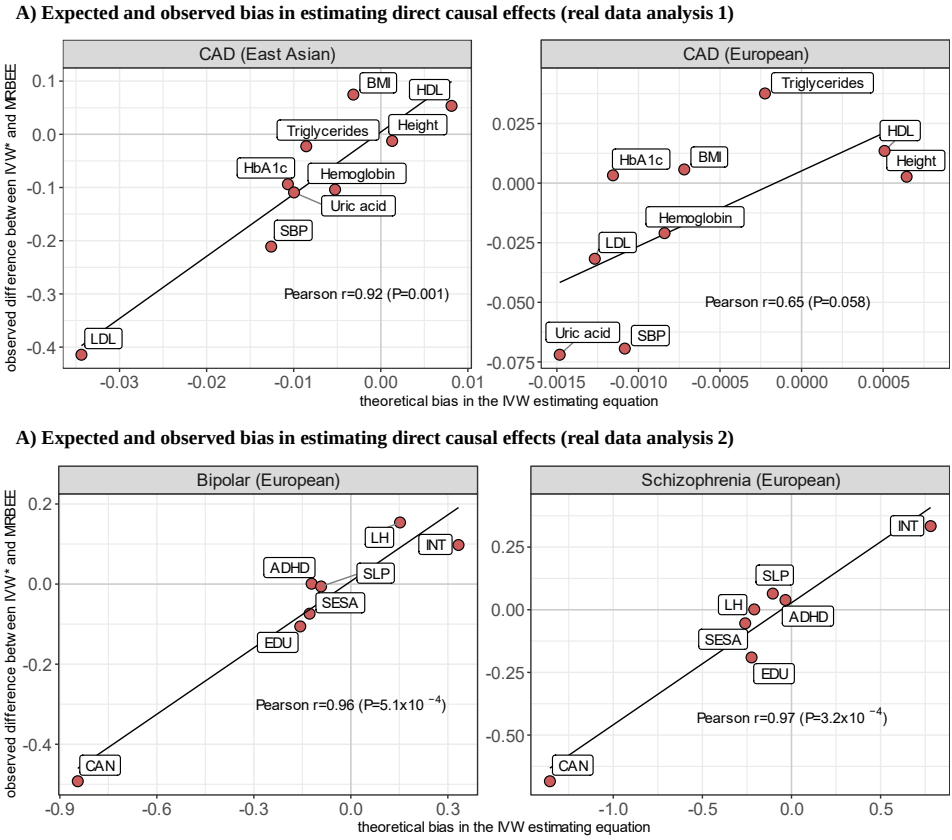


Fig. 7 The x-axes represent theoretical bias in the direct causal effect estimates of IVW* (horizontal pleiotropy-robust IVW; see Figure 6 legend, **Results**), which was calculated using the expectation of Equation 23 with the plugged-in MRBEE direct causal estimates. Y-axes are the observed difference between the IVW* and MRBEE direct causal estimates. Pearson's r values represent the linear correlation between values on the x- and y-axes. Corresponding P-values are for testing the null hypothesis that r=0.

Table 2 Counts of specific locus types from genome-wide horizontal pleiotropy testing

	Mediation	Direct	Pleiotropy	Novel[1]	Novel Replicated Genes
CAD (EUR)	9 (19%)	10 (21%)	19 (41%)	9 (19%)	*FN1,FGD5,PRDM8,FGF5,FURIN,CFDP1,AXL*
CAD (EAS)	37 (62%)	18 (30%)	5 (8%)	0 (0%)	NA
SCZ (EUR)	151 (98%)	1 (1%)	0 (0%)	1 (1%)	*ATXN2L*[2]
BP (EUR)	42 (95%)	2 (5%)	0 (0%)	0 (0%)	NA

.

Values represent counts of loci identified in genome-wide testing for the trait. Percentages within each row sum to 100 and represent the percentage of GWAS loci that can be classified into each of the four categories. Loci are determined with the following parameters: $r^2<0.01$, 1Mb, $P<5x10^{-8}$. Categories are determined using the algorithm in Figure 5B

[a]detected by horizontal pleiotropy testing but not GWAS signal within +-1Mb of the lead pleiotropy SNP was discovered

[b]Insufficient independent GWAS data exist to replicate the association of this gene with SCZ