

## RESEARCH ARTICLE

## A practical problem with Egger regression in Mendelian randomization

Zhaotong Lin<sup>1</sup>, Isaac Pan<sup>2</sup>, Wei Pan<sup>1\*</sup>

**1** Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America, **2** Pomona College, Claremont, California, United States of America

\* [panxx014@umn.edu](mailto:panxx014@umn.edu)

## Abstract

Mendelian randomization (MR) is an instrumental variable (IV) method using genetic variants such as single nucleotide polymorphisms (SNPs) as IVs to disentangle the causal relationship between an exposure and an outcome. Since any causal conclusion critically depends on the three valid IV assumptions, which will likely be violated in practice, MR methods robust to the IV assumptions are greatly needed. As such a method, Egger regression stands out as one of the most widely used due to its easy use and perceived robustness. Although Egger regression is claimed to be robust to directional pleiotropy under the instrument strength independent of direct effect (InSIDE) assumption, it is known to be dependent on the orientations/coding schemes of SNPs (i.e. which allele of an SNP is selected as the reference group). The current practice, as recommended as the default setting in some popular MR software packages, is to orientate the SNPs to be all positively associated with the exposure, which however, to our knowledge, has not been fully studied to assess its robustness and potential impact. We use both numerical examples (with both real data and simulated data) and analytical results to demonstrate the practical problem of Egger regression with respect to its heavy dependence on the SNP orientations. Under the assumption that InSIDE holds for some specific (and *unknown*) coding scheme of the SNPs, we analytically show that other coding schemes would in general lead to the violation of InSIDE. Other related MR and IV regression methods may suffer from the same problem. Cautions should be taken when applying Egger regression (and related MR and IV regression methods) in practice.

## OPEN ACCESS

**Citation:** Lin Z, Pan I, Pan W (2022) A practical problem with Egger regression in Mendelian randomization. PLoS Genet 18(5): e1010166. <https://doi.org/10.1371/journal.pgen.1010166>

**Editor:** Zoltán Kutalik, University Hospital of the Canton Vaud (CHUV), SWITZERLAND

**Received:** December 14, 2021

**Accepted:** March 25, 2022

**Published:** May 4, 2022

**Copyright:** © 2022 Lin et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The GWAS summary datasets are publicly available at <https://doi.org/10.6084/m9.figshare.19669701.v1> The R code for simulations and analyses is available at [https://github.com/ZhaotongL/Egger\\_coding](https://github.com/ZhaotongL/Egger_coding).

**Funding:** ZL was supported by NIH grant R01AG065636. WP was supported by NIH grants R01AG065636, R01AG069895, RF1 AG067924, U01AG073079, R01HL116720, R01GM113250 and R01GM126002. This research was also supported by the Minnesota Supercomputing Institute at the University of Minnesota. The funders had no role in study design, data collection

## Author summary

Egger regression (MR-Egger) has been increasingly applied in Mendelian randomization (MR) analyses for its easy use and perceived robustness, though MR-Egger requires the InSIDE assumption, which in turn depends on the orientation of SNPs. The implications of this dependence to its practical use may not be well understood yet. In particular, it is unrealistic to assume that the InSIDE assumption holds for many or all arbitrarily chosen coding schemes of the SNPs; instead, it is more reasonable to assume that InSIDE holds for only one specific, but usually unknown, coding scheme of SNPs, under which,

and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

however, we show that use of other coding schemes of SNPs in general leads to the violation of InSIDE, and thus to biased causal estimates. The technical reason is due to the seemingly non-restrictive assumption of the random direct effects with a non-zero mean (i.e. directional pleiotropy) in MR-Egger (and related methods), which depends on the orientation of SNPs. This problem persists for many other related MR and instrumental variable regression methods, regardless whether GWAS summary data or individual-level data are used. Given the popularity of MR-Egger in practice, one should be aware of this issue and hence be cautious when applying MR-Egger.

## Introduction

With the increasing availability of large-scale GWAS summary data nowadays, Mendelian randomization (MR) has become a useful tool in epidemiologic studies for identifying determinants or causes of a complex trait or disease [1–3]. In particular, the validity of MR findings relies on three important instrumental variable (IV) assumptions, in which a valid IV used in MR must be

- (i). associated with the exposure  $X$  (relevance assumption);
- (ii). not associated with any hidden confounder  $U$  (independence assumption);
- (iii). not associated with the outcome  $Y$  conditional on the exposure and hidden confounder (exclusion restriction).

While assumption (i) is more likely to hold by selecting IVs strongly associated with the exposure, violations of assumptions (ii) and/or (iii) are more common in practice due to the wide-spread horizontal pleiotropy. In particular, violation of assumption (ii) introduces the so-called correlated pleiotropy (i.e. the pleiotropic effects of SNPs on  $Y$  are correlated with their effects on  $X$ ); uncorrelated pleiotropy results if assumption (iii) is violated and the direct effects on  $Y$  are uncorrelated with those on  $X$ . Egger regression is an MR method that could give a consistent estimate when the exclusion restriction assumption is violated for *all* IVs, but requiring a milder so-called InSIDE assumption, that is, Instrument Strength Independent of Direct Effect [4]. In general, the InSIDE assumption does not hold if assumption (ii) is violated. But other reasons such as a bidirectional relationship between  $X$  and  $Y$  could also cause the violation of InSIDE assumption [5]. Due to both its simplicity and weaker assumptions (in allowing all IVs to have direct effects on the outcome with directional pleiotropy), Egger regression has become one of the most popular MR methods: as one evidence, the number of the citations of one key reference [4] has been increasing every year, totaling over 2000, since its publication in 2015.

Despite its claimed robustness to (uncorrelated) pleiotropy, some authors have noted that Egger regression is dependent on the orientation (or coding) of each SNP/IV [6]. If a (usually biallelic) SNP has two alleles, say an A allele and a G allele, its association value with a trait using A allele as the reference allele would be the opposite of that using G as the reference. Usually we do not expect the analysis conclusion to vary with the coding of a SNP (or any other variables). This property of Egger regression is both surprising and undesirable; we will show in this paper that it is indeed problematic. As to be shown in the real data analysis, when we applied Egger regression to 48 risk factor-disease pairs, the results can be largely different with various orientations or coding schemes of the SNPs being used. One may wonder whether this phenomenon is just due to finite sample sizes. We point out that this problem

exists even for large samples; the GWAS sample sizes in our example of the Height-CAD pair are 253288 and 547261 respectively, and the number of IVs used is 986. We will confirm this analytically and via simulations.

Some authors recognized this problem, mainly from its influence on estimating and interpreting the intercept term (i.e. average pleiotropic effect) in the Egger regression model, and thus proposed the *default* orientation of the SNPs so that they are all positively associated with the exposure as recommended and implemented in some popular MR software packages [7, 8], while (implicitly) imposing the InSIDE assumption being satisfied with this *specific* default orientation [6]. We'd argue however that perhaps this problem is not yet as fully and widely appreciated as it should be, including its implications in practice, as the InSIDE assumption being used needs to be clarified. We believe that it is more realistic to assume that the InSIDE assumption holds only for some unknown *oracle* coding; under this assumption, how would various orientations of SNPs impact the causal estimate? If there are biases, are they going to disappear as the sample size increases? In this paper, we will investigate this problem using simulation studies, followed by analytical explanations, and hopefully raise attention to this problem. We show that the problem carries over to similar IV regression methods with individual-level data [9].

## Methods

### Data and model

Let  $(\hat{\beta}_{x_j}, \hat{\sigma}_{x_j})$  and  $(\hat{\beta}_{y_j}, \hat{\sigma}_{y_j})$  denote the estimate and its standard error for SNP-exposure and SNP-outcome associations respectively from two independent GWAS summary datasets. By default we assume that the second IV assumption (independence assumption) holds unless specified otherwise. We consider the following true causal model (Fig 1):

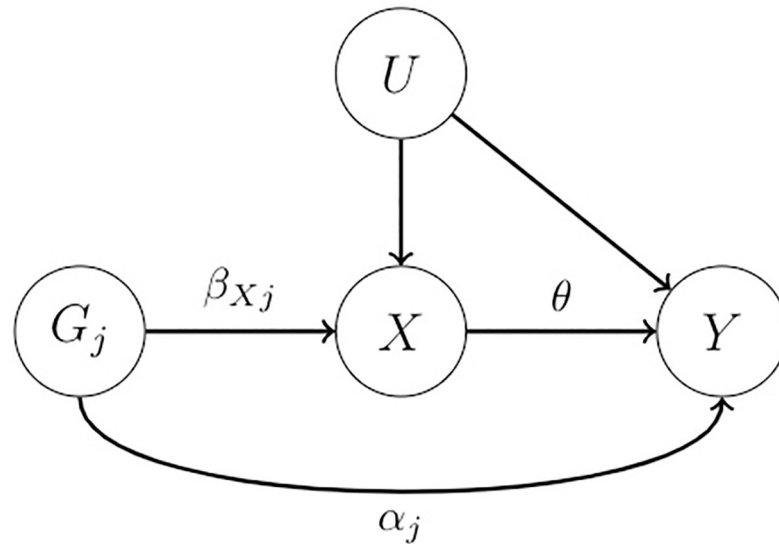
$$\begin{aligned} X &= \sum_{j=1}^m \beta_{x_j} G_j + U + \epsilon_X, \\ Y &= \sum_{j=1}^m \alpha_j G_j + \theta X + U + \epsilon_Y, \end{aligned} \tag{1}$$

where  $\epsilon_X$  and  $\epsilon_Y$  are independent random errors, and  $(\epsilon_X, \epsilon_Y) \perp\!\!\!\perp (G_1, \dots, G_m, U)$ ,  $U$  is an unmeasured (aggregated) confounder, independent with  $\epsilon_X$  and  $\epsilon_Y$ ;  $\theta$  is the causal effect of interest, and  $\alpha_j$  is the direct or pleiotropic effect of  $G_j$  on the outcome  $Y$  not mediated through the exposure  $X$ . Throughout the paper, we assume that the  $m$  SNPs are mutually independent. The association between SNP  $G_j$  and the outcome  $Y$ ,  $\beta_{y_j}$ , can be decomposed as:

$$\beta_{y_j} = \alpha_j + \theta \beta_{x_j}. \tag{2}$$

SNP  $G_j$  is an invalid IV with a pleiotropic/direct effect if  $\alpha_j \neq 0$ .

The GWAS summary statistics  $\hat{\beta}_{x_j}$  and  $\hat{\beta}_{y_j}$ ,  $j = 1, \dots, m$ , are usually computed from simple linear regressions:  $\hat{\beta}_{x_j} = \widehat{\text{cov}}(X, G_j) / \widehat{\text{var}}(G_j)$  and  $\hat{\beta}_{y_j} = \widehat{\text{cov}}(Y, G_j) / \widehat{\text{var}}(G_j)$ , where  $\widehat{\text{cov}}()$  and  $\widehat{\text{var}}()$  are the sample covariance and sample variance respectively. When  $G_1, \dots, G_m$  are mutually independent, we obtain that  $\hat{\beta}_{x_j}$  and  $\hat{\beta}_{y_j}$  follow the asymptotic normal distributions with



**Fig 1. The true causal diagram.**

<https://doi.org/10.1371/journal.pgen.1010166.g001>

means  $\beta_{Xj}$  and  $\beta_{Yj}$  respectively. Given large sample sizes of GWAS as usual, we have

$$\widehat{\beta}_{Xj} = \beta_{Xj} + \epsilon_{Xj}, \quad \epsilon_{Xj} \sim \mathcal{N}(0, \sigma_{Xj}^2), \quad j = 1, \dots, m, \tag{3}$$

$$\widehat{\beta}_{Yj} = \alpha_j + \theta\beta_{Xj} + \epsilon_{Yj}, \quad \epsilon_{Yj} \sim \mathcal{N}(0, \sigma_{Yj}^2), \quad j = 1, \dots, m. \tag{4}$$

Throughout the paper, as usual for Egger regression, we impose the no-measurement error (NOME) assumption on the SNP-exposure estimates, i.e.,  $\sigma_{Xj} = 0$  and  $\widehat{\beta}_{Xj} = \beta_{Xj}$ . We also assume that  $\sigma_{Yj}$  is known or well estimated as  $\widehat{\sigma}_{Yj}$ .

### Inverse-variance weighted (IVW) method

The inverse-variance weighted (IVW) method can be viewed as a meta-analysis of the ratio estimates of the causal parameter  $\theta$ ,  $\widehat{\beta}_{Yj}/\widehat{\beta}_{Xj}$ , across all SNPs [10]. Under the NOME assumption on SNP-exposure associations, the overall estimator of the causal effect can be obtained by averaging across all  $m$  SNPs using inverse-variance weighting with weights  $\widehat{\beta}_{Xj}^2/\widehat{\sigma}_{Yj}^2$ :

$$\widehat{\theta}_{IVW} = \frac{\sum_{j=1}^m \widehat{\beta}_{Yj} \widehat{\beta}_{Xj} \widehat{\sigma}_{Yj}^{-2}}{\sum_j \widehat{\beta}_{Xj}^2 \widehat{\sigma}_{Yj}^{-2}}. \tag{5}$$

The same estimator can be also obtained from a weighted linear regression of  $\widehat{\beta}_{Yj}$  on  $\widehat{\beta}_{Xj}$  with weights  $(\widehat{\sigma}_{Yj}^{-2})$  and with the intercept constrained to be zero. When there is no heterogeneity in the Wald ratio estimates of  $\theta$  based on the individual IVs, the variance of  $\widehat{\theta}_{IVW}$  is  $1/\sum_j \widehat{\beta}_{Xj}^2 \widehat{\sigma}_{Yj}^{-2}$ , which corresponds to a fixed-effect (FE) meta-analysis, denoted IVW(FE). Otherwise, a multiplicative random-effect model, or IVW(RE), should be preferred and the variance of  $\widehat{\theta}_{IVW}$  is  $\sigma_I/\sum_j \widehat{\beta}_{Xj}^2 \widehat{\sigma}_{Yj}^{-2}$ , where  $\sigma_I$  is an overdispersion parameter to be estimated from the residuals in the weighted linear regression described above [6, 11]. In the presence of (balanced) pleiotropy, the

over-dispersion parameter ( $\sigma_I$ ) allows the variance of  $\hat{\theta}_{IVW}$  to increase so that IVW(RE) could control the type-I error, but the point estimates of IVW(FE) and IVW(RE) are still the same [11].

While requiring the InSIDE assumption, IVW(RE), as another popular MR method closely related to Egger regression, is however invariant to and thus has no problem with various orientations of SNPs. This can be seen from Eq (5) that both the numerator and the denominator are invariant to different orientations of SNPs. This advantage of IVW(RE) and other approaches in modeling the mean of  $\alpha_j$ 's as 0 has been noted by others [12].

### Egger regression (MR-Egger)

Egger regression (MR-Egger) is a simple modification of IVW(RE) by adding an intercept term to capture the non-zero (weighted) average pleiotropic effect [4]:

$$\hat{\beta}_{Y_j} = r + \theta \hat{\beta}_{X_j} + \epsilon_{Ej}; \quad \epsilon_{Ej} \sim \mathcal{N}(0, \sigma_E^2 \hat{\sigma}_{Y_j}^2), \tag{6}$$

where  $\epsilon_{Ej}$  is a random error and  $\sigma_E^2 \geq 1$  is an unknown overdispersion parameter (as  $\sigma_I^2$  used in IVW(RE)). The model can be derived from Eqs (3) and (4): under the NOME assumption, we have

$$\hat{\beta}_{Y_j} = \alpha_j + \theta \hat{\beta}_{X_j} + \epsilon_{Yj}.$$

Treating  $\alpha_j$  as random, we have

$$E(\hat{\beta}_{Y_j} | \hat{\beta}_{X_j}) = E(\alpha_j | \hat{\beta}_{X_j}) + \theta \hat{\beta}_{X_j} = E(\alpha_j) + \theta \hat{\beta}_{X_j} = r + \theta \hat{\beta}_{X_j},$$

where the second equality follows from the InSIDE assumption, and the third from the definition of  $E(\alpha_j) = r$ . Each  $\hat{\beta}_{X_j}$  is treated as fixed (as a covariate) in fitting the Egger regression Eq (6). It is also clear that other parameters  $\theta$  and  $r$  are treated as fixed.

When the intercept term  $r$  is equal to zero, that is, the average pleiotropic effect is zero (known as balanced pleiotropy), the MR-Egger and IVW estimators coincide and both are consistent estimators of the causal effect under the InSIDE assumption, which will be discussed next. When the average pleiotropic effect is not zero (known as directional pleiotropy), it is known that the IVW estimator is not consistent anymore while the MR-Egger estimator is consistent under the InSIDE assumption [4, 6].

### The InSIDE assumption and orientation of SNPs

Throughout this paper, the InSIDE assumption is defined in model Eq (4) for a fixed number of IVs; specifically, the InSIDE assumption holds if the *sample* weighted covariance between the pleiotropic effects and SNP-exposure associations equals to zero for the set of SNPs used in the MR analysis [6, 13]:

$$\text{cov}_w(\boldsymbol{\alpha}, \boldsymbol{\beta}_X) = \frac{\sum_{j=1}^m \hat{\sigma}_{Y_j}^{-2} (\alpha_j - \bar{\boldsymbol{\alpha}}) (\beta_{X_j} - \bar{\boldsymbol{\beta}}_X)}{\sum_{j=1}^m \hat{\sigma}_{Y_j}^{-2}} = 0, \tag{7}$$

where  $\bar{\boldsymbol{\alpha}}$  and  $\bar{\boldsymbol{\beta}}_X$  are the weighted sample averages of  $\alpha_j$  and  $\beta_{X_j}$  respectively, with weights equal to  $1/\hat{\sigma}_{Y_j}^2$ . The MR-Egger estimator is consistent under this InSIDE assumption [6]:

$$\hat{\theta}_{Egger} = \frac{\text{cov}_w(\hat{\boldsymbol{\beta}}_Y, \hat{\boldsymbol{\beta}}_X)}{\text{var}_w(\hat{\boldsymbol{\beta}}_X)} \xrightarrow{n \rightarrow \infty} \theta + \frac{\text{cov}_w(\boldsymbol{\alpha}, \boldsymbol{\beta}_X)}{\text{var}_w(\boldsymbol{\beta}_X)} = \theta, \tag{8}$$

where  $n$  is the sample size of the GWAS data.

In Eq (2), if we use the other allele of SNP  $G_j$  as the reference, i.e., if we flip the coding of SNP  $G_j$ , we will have its associated  $\beta_{x_j}^* = -\beta_{x_j}$  and  $\beta_{y_j}^* = -\beta_{y_j}$ , and the pleiotropic effect will also change accordingly:  $\alpha_j^* = -\alpha_j$ . Consequently, the average pleiotropy (i.e.,  $\bar{\alpha}$ ), and thus possibly whether there is balanced or directional pleiotropy, will change. That is, the definition of directional pleiotropy depends on the SNP coding. More importantly, it will impact whether the InSIDE assumption holds or not in Eq (4). It is clear that the InSIDE assumption is defined with respect to a *specific* coding scheme of the set of SNPs. With the intercept constrained to be 0 in IVW, flipping the coding of any SNPs will not change the resulting IVW estimate, but it will in general impact the resulting MR-Egger estimates of the causal parameter (i.e. the slope parameter in Eq (6)) as well as the intercept. One exception is when we flip the coding of *all* SNPs, the causal estimate from MR-Egger will remain the same. This issue has been recognized by some authors; accordingly they proposed re-orientating all the SNPs to be positively associated with the exposure [6, 14]. Although this current practice of MR-Egger allows the users to obtain the same causal estimate under the same coding scheme for the same GWAS data, notably it requires the assumption that InSIDE holds for *this* specific default orientation of the SNPs; if this assumption does not hold, MR-Egger may not perform well as to be shown in simulations. In the real analysis, we will show that the results of MR-Egger largely depend on the specific coding being used. We can also see from Eq (8) that the variance of the causal estimate in MR-Egger is inversely proportional to the weighted variance of  $\hat{\beta}_x$ , which also depends on the orientation of SNPs.

### The NOME assumption and orientations of SNPs

The standard implementations of IVW and Egger regression as discussed in Inverse-variance weighted (IVW) method and Egger regression (MR-Egger) both assume no-measurement-errors (NOME) for the SNP-exposure associations. However, in practice with finite samples, this could never hold, which can lead to biased causal estimates, even when InSIDE holds. The impact of violation of NOME assumption on IVW and MR-Egger has been studied extensively before [13, 15, 16]. In particular, for an unweighted Egger regression (when  $\hat{\sigma}_{y_j}^2$ s are the same), as shown in Eq.(3) in [13],  $E(\hat{\theta}_{Egger}) \approx \theta \text{var}(\beta_x) / \text{var}(\hat{\beta}_x)$ , where  $\text{var}()$  is the sample variance calculated on the set of  $m$  IVs used in the analysis. [13] proposed to use  $I^2 = (Q - (m - 1))/Q$  to estimate the ratio  $\text{var}(\beta_x) / \text{var}(\hat{\beta}_x)$ , where  $Q = \sum_{j=1}^m (\hat{\beta}_{x_j} - \bar{\hat{\beta}}_x)^2 / \hat{\sigma}_{x_j}^2$  and  $\bar{\hat{\beta}}_x$  is the mean of  $\hat{\beta}_x$  weighted by  $1/\hat{\sigma}_{x_j}^2$ . It is clear that the degree of NOME violation in MR-Egger also depends on the orientations of SNPs. As  $\hat{\beta}_{x_j}$ 's are more widely dispersed,  $I^2$  is closer to one, and the impact of NOME violation is smaller. However, the default coding makes all  $\hat{\beta}_{x_j}$  to be positive and thus  $I^2$  would be the smallest among all SNP coding schemes. As to be shown later in simulations, this will lead to larger biases in the causal estimates using the default coding even when the InSIDE assumption is satisfied.

### Radial Egger regression

A closely related method, Radial-Egger [17], introduces a new form of MR-Egger:

$$\hat{\theta}_j \sqrt{w_j} = r' + \theta \sqrt{w_j} + \epsilon_{Rj}; \quad \epsilon_{Rj} \sim \mathcal{N}(0, \sigma_R^2),$$

where  $\hat{\theta}_j = \hat{\beta}_{y_j} / \hat{\beta}_{x_j}$  and we used the first-order weights  $w_j = \hat{\beta}_{x_j}^2 / \hat{\sigma}_{y_j}^2$ . To derive the above model, as for MR-Egger, we start from the true model Eqs (3) and (4) with the NOME

assumption, obtaining

$$\hat{\theta}_j \sqrt{w_j} = \alpha_j \sqrt{w_j} / \hat{\beta}_{X_j} + \theta \sqrt{w_j} + \epsilon_{Y_j} \sqrt{w_j} / \hat{\beta}_{X_j},$$

$$E(\hat{\theta}_j \sqrt{w_j} | \sqrt{w_j}) = E(\alpha_j \sqrt{w_j} / \hat{\beta}_{X_j} | \sqrt{w_j}) + \theta \sqrt{w_j},$$

which reduces to the Radial-Egger regression model if

$$E(\alpha_j \sqrt{w_j} / \hat{\beta}_{X_j} | \sqrt{w_j}) = E(\alpha_j \text{sign}(\hat{\beta}_{X_j}) \hat{\sigma}_{Y_j}^{-1} | \sqrt{w_j}) = E(\alpha_j \text{sign}(\hat{\beta}_{X_j}) \hat{\sigma}_{Y_j}^{-1}) = r',$$

where the second equality requires a new form of the InSIDE assumption that the (weighted) pleiotropic effects (with respect to the default exposure-increasing coding as adopted here) are independent of the Radial weights [17] (and as usual assuming that  $\hat{\sigma}_{Y_j}^{-1}$  is fixed). This InSIDE assumption is similar to the one used (for the default coding) in MR-Egger. As to be shown in the simulation, Radial-Egger indeed performed similarly to MR-Egger with the default coding.

## GWAS summary data

We examined the SNP coding issue of Egger regression on 48 risk factor-disease pairs, including 12 cardiometabolic risk factors and 4 diseases. The 12 risk factors were triglycerides (TG), low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) [18], Height [19], body-mass index (BMI) [20], body fat percentage (BF) [21], birth weight (BW) [22], diastolic blood pressure (DBP), systolic blood pressure (SBP) [23], fasting glucose (FG) [24], Smoke and Alcohol [25]. The 4 diseases were coronary artery disease (CAD) [26], stroke [27], type 2 diabetes (T2D) [28] and asthma [29].

The sample sizes for the 16 GWAS datasets ranged from 46 186 to 1 232 091, with a median of 220 933. The numbers of IVs used in the 48 MR analyses ranged from 9 to 1345, with a median of 126.

## Simulation set-ups

We simulated data according to the true causal model Eq (1) with  $G_j \sim \text{Binomial}(2, 0.3)$ , and  $U, \epsilon_X, \epsilon_Y \sim \mathcal{N}(0, 1)$  independently. We simulated  $\beta_{X_j}$  from (a) a uniform distribution on  $(-0.2, -0.1) \cup (0.1, 0.2)$ ; (b) a uniform distribution on  $(-0.1, -0.03) \cup (0.1, 0.2)$  and (c) a uniform distribution on  $(0.1, 0.3)$ . We will refer them to Simulation (a), (b) and (c) respectively later. In all three simulation set-ups, we considered 0%, 30%, 70% or 100% invalid IVs with balanced pleiotropy  $\alpha_j \sim \mathcal{N}(0, 0.1^2)$ , or with directional pleiotropy  $\alpha_j \sim \mathcal{N}(0.1, 0.1^2)$ . It is noted that for each simulation set-up, we generated IV strengths  $\beta_{X_j}$ 's and direct effects  $\alpha_j$ 's from *independent* distributions. Although for each specific simulated dataset, the sample covariance of  $\beta_{X_j}$  and  $\alpha_j$  ( $j = 1, \dots, m$ ) (Eq (7)) might not be exactly equal to zero, across all simulated datasets, the average sample covariance between  $\beta_{X_j}$  and  $\alpha_j$  will be (nearly) zero (see S1 Text). Following [13], we refer this as 'weak' InSIDE assumption.

The causal effect  $\theta$  was set to 0 or 0.2, and the number of IVs,  $m$ , to 30 or 100. For Simulation (b), we also considered 500 IVs. The summary data for genetic associations were calculated for the exposure and the outcome on non-overlapping samples of individuals, each consisting of  $n = 50\,000$  or  $100\,000$  individuals. The *oracle* coding referred to the specific coding that was used to generate the simulated data under the (weak) InSIDE assumption.

In Simulation (a), the mean of SNP-exposure associations was zero, a special case we will discuss later; Simulation (b) was more general with both positive and negative SNP-exposure associations with a non-zero mean; Simulation (c) was also a special case with all SNP-

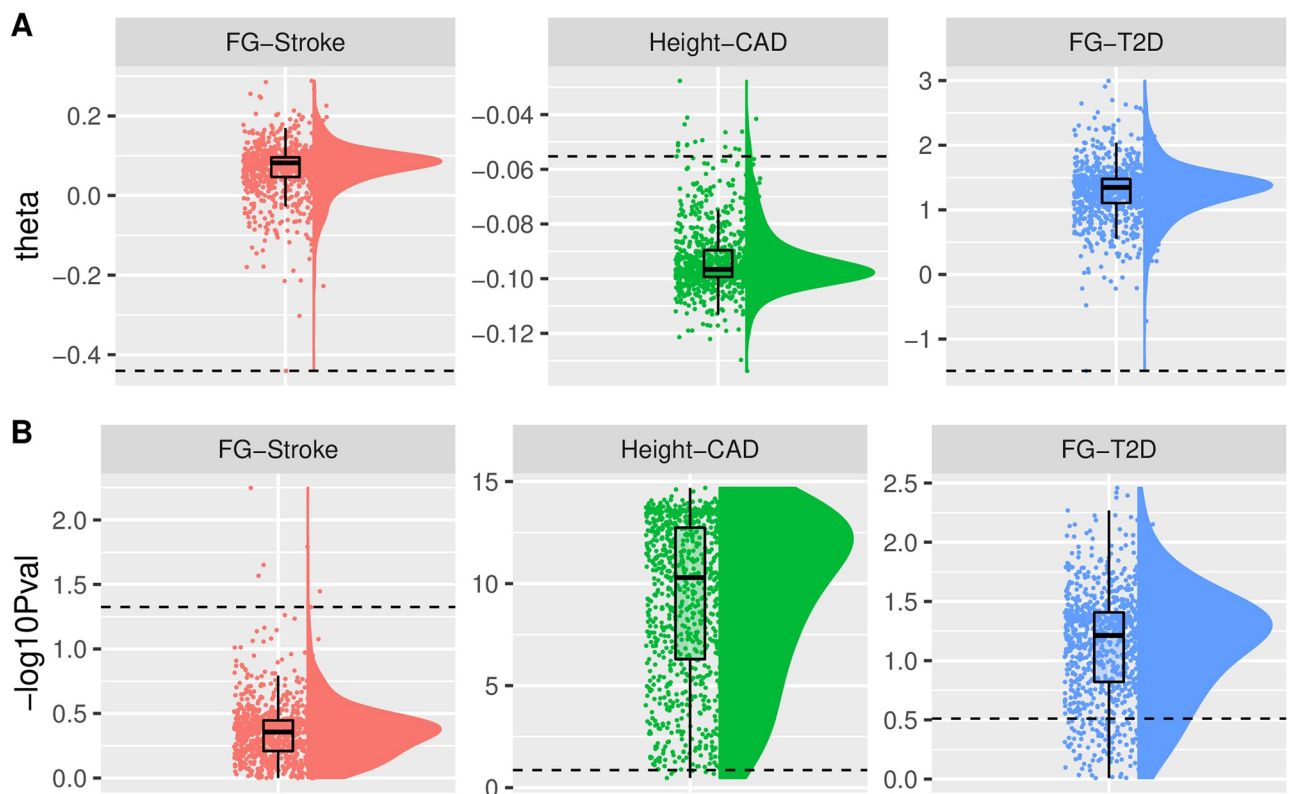
exposure associations being positive. Here the SNP-exposure associations as well as the pleiotropy were defined with respect to the *oracle* coding scheme.

We ran 1000 replications for each simulation set-up. For each simulated dataset, we applied MR-Egger with (i) the *default* coding (as adopted in the current practice): we orientated SNPs so that  $\hat{\beta}_{x_j}$  were all positive in Eq (6); (ii) the *oracle* coding: we used the coding generating the simulated data, under which the weak InSIDE assumption was satisfied; (iii) the *random* coding: we randomly flipped the coding of some SNPs. We also applied IVW(RE) and Radial-Egger for comparison.

## Results

### Real data example

As a motivating example, we applied MR-Egger to some large-scale GWAS summary data of 48 risk factor-disease pairs [30] using the *default* coding scheme (i.e., we orientated the SNPs so that they were all positively associated with the exposure as recommended and implemented in the popular TwoSampleMR software [7]) and a *random* coding scheme (i.e. we randomly selected the reference allele for each SNP). With the default coding, Egger regression identified 7 significant pairs, whereas with the whatever coding given in the original GWAS datasets from [30], Egger regression identified 17 significant pairs (after the Bonferroni correction). Fig 2 shows some representative results for three pairs: Fasting glucose (FG)-Stroke, body height-coronary artery disease (CAD) and FG-type 2 diabetes (T2D), in which we tried 999



**Fig 2. Three real data examples showing different results from different coding schemes in Egger regression.** Panel A: estimates of the causal effect; Panel B:  $-\log_{10}(p\text{-value})$ 's of the causal effect. The dashed line in each plot corresponds to the result from the *default* coding.

<https://doi.org/10.1371/journal.pgen.1010166.g002>

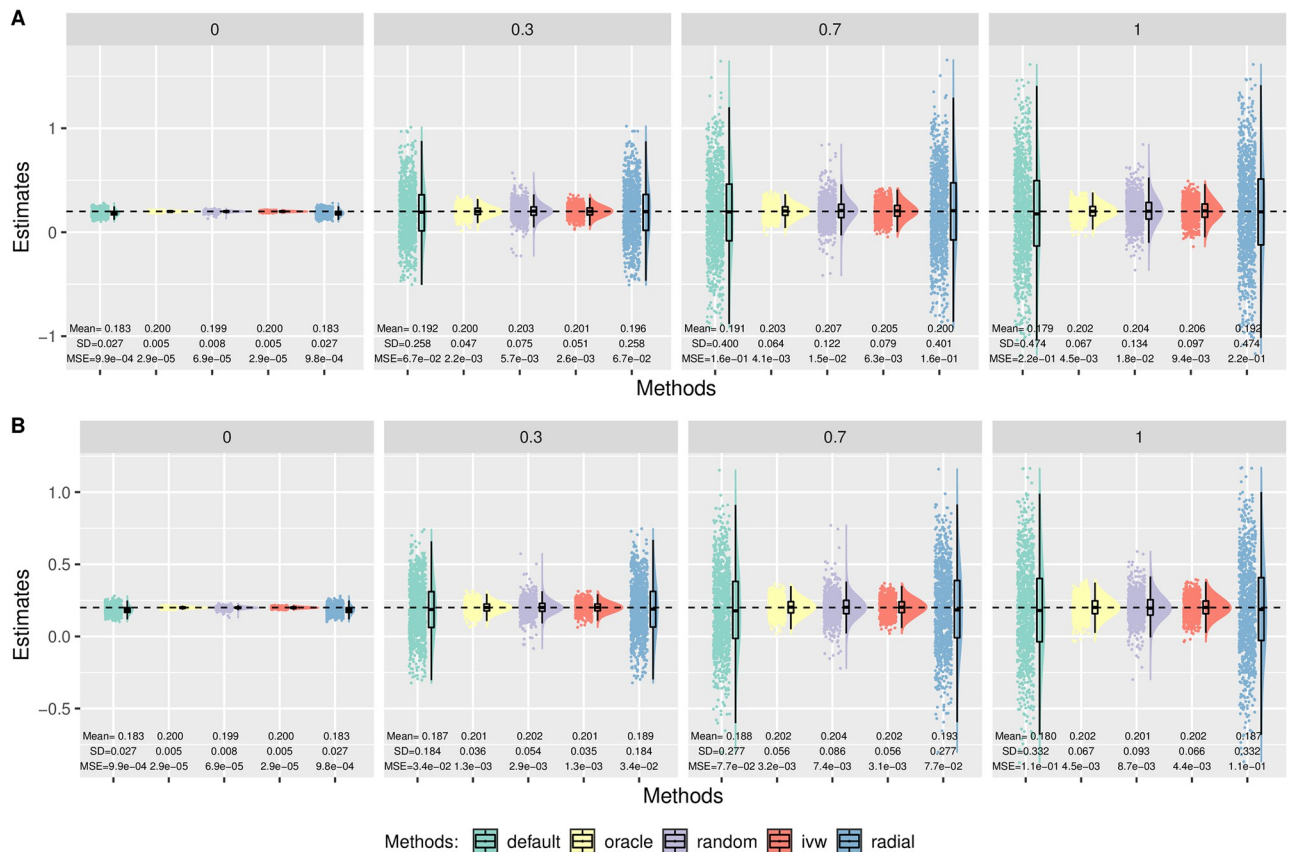


random (and unique) codings plus the default one (as the dashed lines in the plot). We can see that with different coding schemes in Egger regression, not only the point estimates of the causal parameter varied (even from negative to positive), but also the p-values (from insignificant to highly significant), giving possibly opposite conclusions. For example, for the FG-Stroke pair, using the default coding suggested a negative relationship ( $\hat{\theta} = -0.44$  with p-value = 0.047) while using the original coding in the GWAS dataset suggested a positive relationship ( $\hat{\theta} = 0.17$  with p-value = 0.010). These results clearly demonstrate the critical and possibly dramatic dependence of MR-Egger on the orientation of SNPs.

We also investigated whether a larger intercept estimate (in absolute value) was obtained when the default (i.e. exposure-increasing) coding was used in MR-Egger. We found that using the default coding often yielded more extreme intercept estimates than using other (random) coding schemes, but not necessarily more significant p-values (because the standard errors of the intercept estimates were usually larger under the default coding). The details are given in [S1 Text](#).

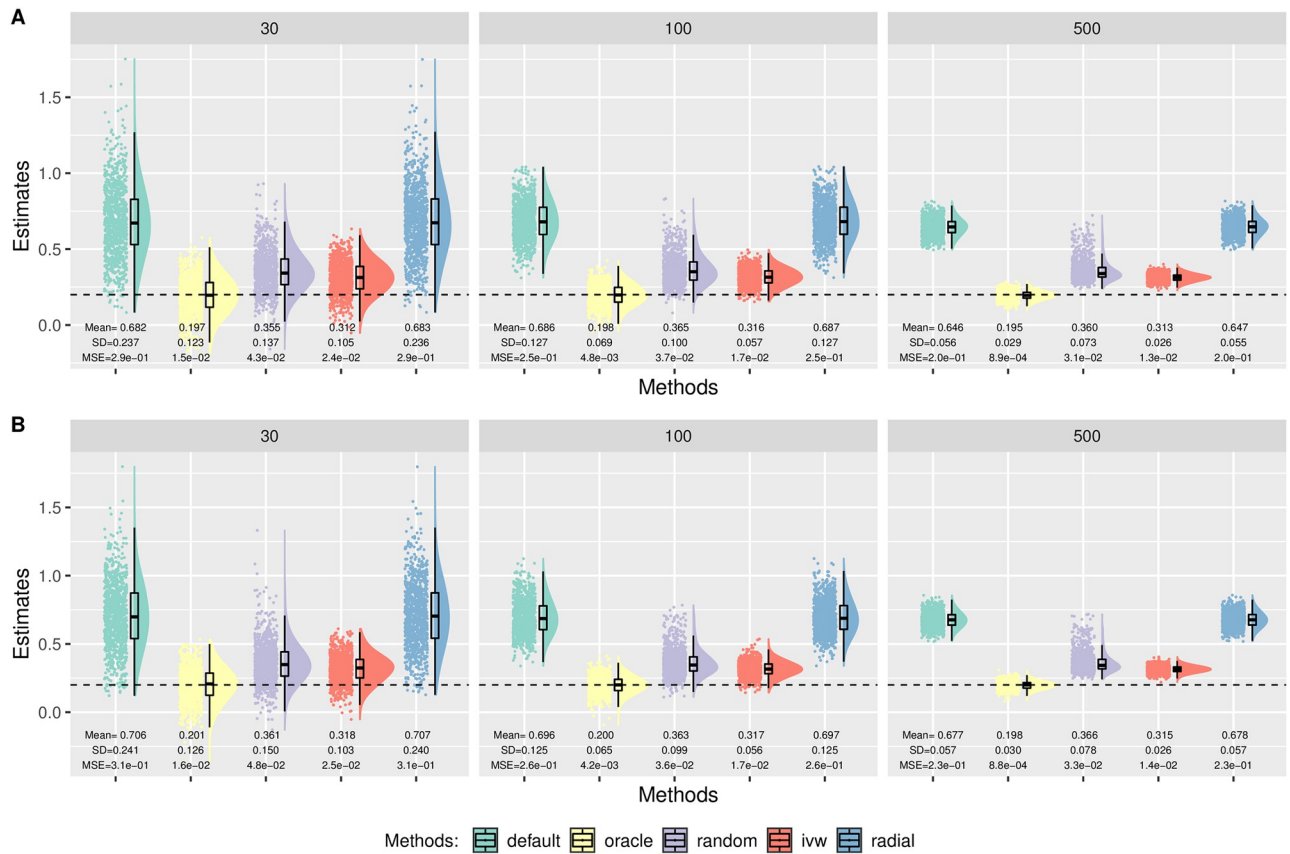
### Simulation results

**Simulation (a): SNP-exposure associations with mean 0.** In Simulation (a), we generated  $\beta_{Xj}$  from a uniform distribution on  $(-0.2, -0.1) \cup (0.1, 0.2)$ . Here we only show some representative results while others are given in [S1 Text](#). [Fig 3](#) shows the distributions of the causal



**Fig 3. Simulation (a) results with  $n = 100,000$ ,  $\theta = 0.2$ ,  $m = 100$ .** Empirical distributions of the estimates of the causal effect  $\theta$  by the methods. Each column corresponds to 0%, 30%, 70% or 100% invalid IVs. A: Directional pleiotropy. B: Balanced pleiotropy.

<https://doi.org/10.1371/journal.pgen.1010166.g003>

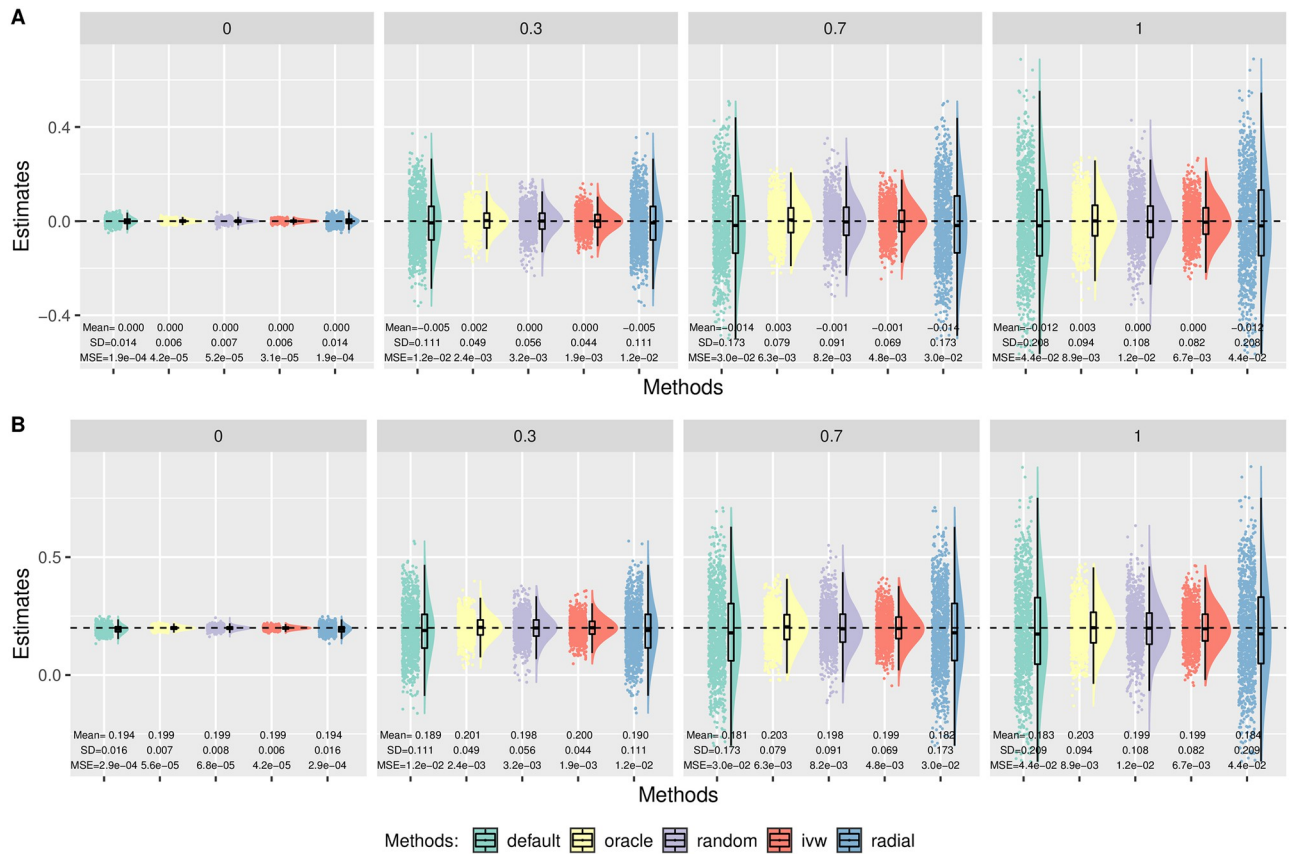


**Fig 4. Simulation (b) results with directional pleiotropy.** Empirical distributions of the estimates of the causal effect  $\theta$  by the methods with 30% invalid IVs and  $\theta = 0.2$ . Each column corresponds to  $m = 30, 100$  or  $500$  IVs. A:  $n = 50\ 000$ . B:  $n = 100\ 000$ .

<https://doi.org/10.1371/journal.pgen.1010166.g004>

estimates by each method with  $m = 100, n = 100\ 000, \theta = 0.2$  in the presence of directional (Panel A) and balanced pleiotropy (Panel B). First, MR-Egger with the oracle coding performed the best with unbiased estimates and smallest variances across all different scenarios. In the case of balanced pleiotropy, MR-Egger with the oracle coding and IVW coincided with each other as expected. In the meantime, MR-Egger with the default coding gave slightly biased estimates towards the null. This was perhaps due to the violation of NOME assumption with the average estimated  $I^2$  statistic about 0.921. On the other hand, the average estimated  $I^2$  statistic was 0.997 with the oracle coding. (More simulation results studying the NOME assumption are given in [S1 Text](#).) Despite the slight bias due to violation of NOME assumption, we note that the (weak) InSIDE assumption still held under the default coding in this simulation set-up, which will be shown in Analysis. As for Radial-Egger, we can see that it performed similarly to the default coding in MR-Egger. Also, perhaps surprisingly, IVW yielded the unbiased estimates even in the case of directional pleiotropy ([Fig 3A](#)), and we will show the reason later. In addition, even though the (weak) InSIDE assumption still held, using the default coding magnified the extent of the violation of NOME assumption, leading to larger finite-sample biases. Finally, MR-Egger with the default coding yielded the largest variance for the causal estimate, implying its low power.

**Simulation (b): SNP-exposure associations with a non-zero mean.** In Simulation (b), we generated  $\beta_{X_j}$  from a uniform distribution on  $(-0.1, -0.03) \cup (0.1, 0.2)$ . [Fig 4](#) shows the distributions of the causal estimates by each method with  $\theta = 0.2$  in the presence of directional



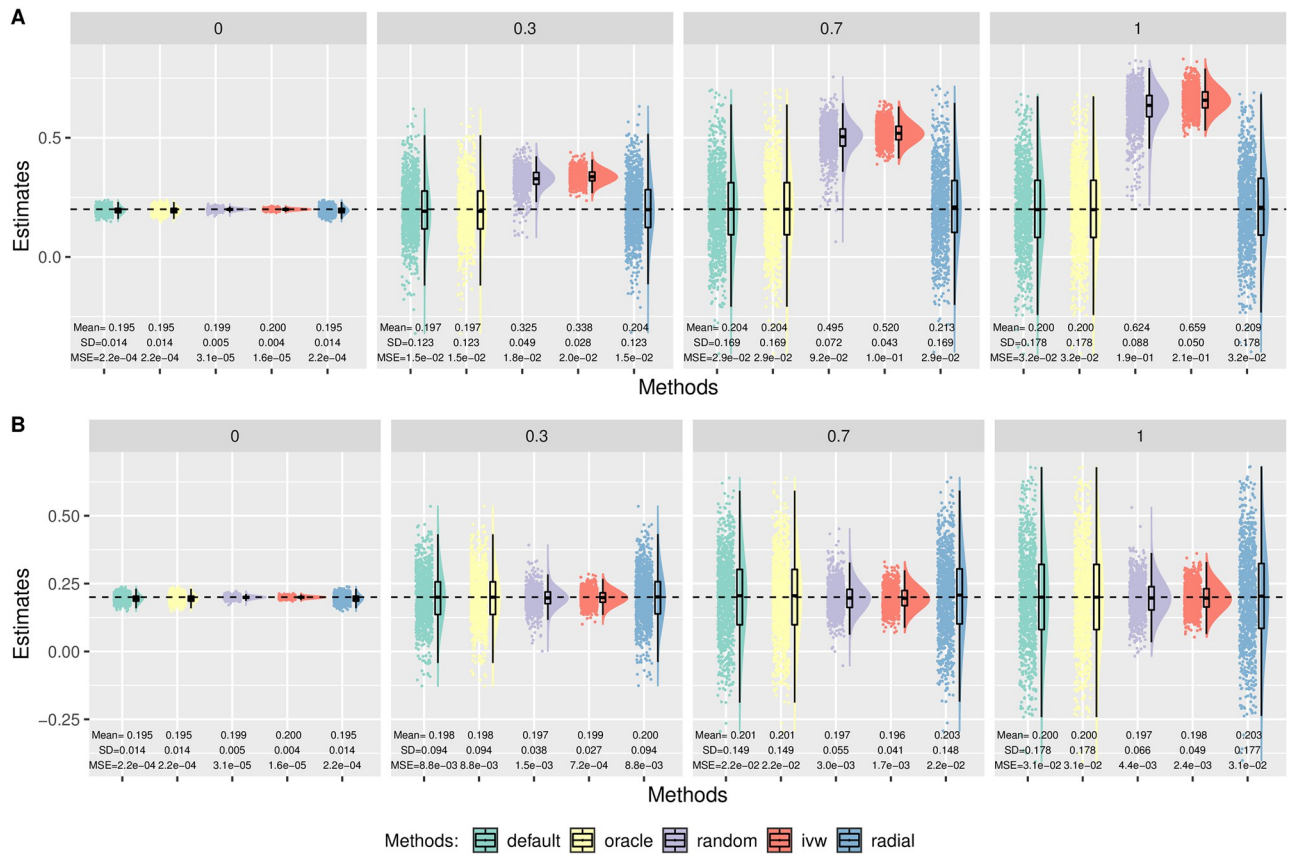
**Fig 5. Simulation (b) results with balanced pleiotropy,  $n = 100\,000$ ,  $m = 100$ .** Empirical distributions of the estimates of the causal effect  $\theta$  by the methods. Each column corresponds to 0%, 30%, 70% or 100% invalid IVs. A:  $\theta = 0$ . B:  $\theta = 0.2$ .

<https://doi.org/10.1371/journal.pgen.1010166.g005>

pleiotropy with 30% invalid IVs. The top row corresponds to  $n = 50\,000$  and the bottom corresponds to a larger sample size of  $n = 100\,000$ . As we can see, only MR-Egger with the oracle coding gave unbiased estimates. Radial-Egger and MR-Egger with the default coding performed similarly with the largest bias. Moreover, the bias did not disappear as the sample size and the number of IVs increased. This result may seem to contradict the common belief that MR-Egger is robust to directional pleiotropy (under InSIDE), but we will show later that the current practice of flipping SNPs actually led to the violation of the InSIDE assumption in this scenario, thus yielding the biased causal parameter estimates. Furthermore, this was *not* due to the violation of the NOME assumption either. When we used the true  $\beta_{X_j}$ , instead of the estimated  $\hat{\beta}_{X_j}$ , the bias still persisted for the default coding in MR-Egger; the details are given in [S1 Text](#).

[Fig 5](#) shows the results with  $m = 100$  and  $n = 100\,000$  in the case of balanced pleiotropy. In this case, MR-Egger with the oracle coding and IVW yielded unbiased estimates with the smallest variance. Again, though the (weak) InSIDE assumption held under the default coding here (with balanced pleiotropy) as to be shown in Analysis, it yielded slight under-estimation perhaps due to the violation of NOME.

**Simulation (c): All positive SNP-exposure associations.** In Simulation (c), we generated  $\beta_{X_j}$  to be all positive from a uniform distribution on (0.1, 0.3). As a result, the default coding coincided with the oracle coding in this case. As shown in [Fig 6](#), the default and oracle codings



**Fig 6. Simulation (c) results with  $n = 100\,000$ ,  $\theta = 0.2$ ,  $m = 100$ .** Empirical distributions of the estimates of the causal effect  $\theta$  by the methods. Each column corresponds to 0%, 30%, 70% or 100% invalid IVs. A: Directional pleiotropy. B: Balanced pleiotropy.

<https://doi.org/10.1371/journal.pgen.1010166.g006>

in MR-Egger had the same results with approximately unbiased estimates, as expected. Again, Radial-Egger gave the similar results to that of the default coding in MR-Egger. On the other hand, IVW yielded biased estimates in the presence of directional pleiotropy (panel A), but gave unbiased estimates with balanced pleiotropy (panel B).

### Analysis

In this session, we will dive into the relationship between the orientation of SNPs and the InSIDE assumption, and the impact on the IVW and MR-Egger estimates of the causal parameter. For simplicity of notation, we assume that the  $m$  SNPs have the same minor allele frequency so that  $\hat{\sigma}_{Y_j}^2$  are the same and thus no need to use the weighted covariance in the definition of InSIDE; a similar argument carries over for the general case. First, the InSIDE assumption in Eq (7) for the oracle coding becomes:

$$m \cdot \text{cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}_X) = \sum_{j=1}^m \alpha_j \beta_{X_j} - m \bar{\alpha} \bar{\beta}_X = 0 \tag{9}$$

If we flip the coding of some SNPs, say the first  $0 < k < m$  SNPs, and denote the new data as  $(\beta_{X_j}^*, \beta_{Y_j}^*)$ . That is, for  $j = 1, \dots, k$ ,  $\beta_{X_j}^* = -\beta_{X_j}$ ,  $\beta_{Y_j}^* = -\beta_{Y_j}$  and  $\alpha_j^* = -\alpha_j$ , while for others,

$\beta_{X_j}^* = \beta_{X_j}$ ,  $\beta_{Y_j}^* = \beta_{Y_j}$  and  $\alpha_j^* = \alpha_j$ . Then in general we have

$$m \cdot \text{cov}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}_X^*) = \sum_{j=1}^m \alpha_j^* \beta_{X_j}^* - m \bar{\boldsymbol{\alpha}}^* \bar{\boldsymbol{\beta}}_X^* = \sum_{j=1}^m \alpha_j \beta_{X_j} - m \bar{\boldsymbol{\alpha}}^* \bar{\boldsymbol{\beta}}_X^* \neq 0, \tag{10}$$

where the last inequality is due to  $\sum_j \alpha_j \beta_{X_j} = m \bar{\boldsymbol{\alpha}} \bar{\boldsymbol{\beta}}_X \neq m \bar{\boldsymbol{\alpha}}^* \bar{\boldsymbol{\beta}}_X^*$  unless under some special situations. This suggests that the default coding or an arbitrarily chosen coding in MR-Egger is likely to lead to the violation of the InSIDE assumption, and by (8) and (10), to inconsistent estimates.

In general, the (asymptotic) bias of the MR-Egger estimate will not diminish even for a large  $m$  after flipping the coding of  $k$  SNPs on the basis of the oracle coding. Under the theoretical model that both  $\alpha_j$ 's and  $\beta_{X_j}$ 's are iid from two continuous distributions each of a bounded domain and non-zero mean, and that  $k/m \rightarrow c \in (0, 1)$  as  $m \rightarrow \infty$ . It is easy to verify that as  $m \rightarrow \infty$ , we have  $\bar{\boldsymbol{\alpha}} \rightarrow \mu_\alpha$  say, and  $\bar{\boldsymbol{\alpha}}^* \rightarrow (1 - 2c)\mu_\alpha \neq \mu_\alpha$ , and similarly for  $\bar{\boldsymbol{\beta}}_X$  and  $\bar{\boldsymbol{\beta}}_X^*$ . Then with probability one we have  $\bar{\boldsymbol{\alpha}} \bar{\boldsymbol{\beta}}_X - \bar{\boldsymbol{\alpha}}^* \bar{\boldsymbol{\beta}}_X^* \not\rightarrow 0$  as  $m \rightarrow \infty$ .

In contrast, the IVW estimate in Eq (5) after flipping the coding becomes

$$\hat{\theta}_{IVW}^* = \frac{\sum_{j=1}^m \hat{\beta}_{Y_j} \hat{\beta}_{X_j}^*}{\sum_{j=1}^m (\hat{\beta}_{X_j}^*)^2} = \frac{\sum_{j=1}^m \hat{\beta}_{Y_j} \hat{\beta}_{X_j}}{\sum_{j=1}^m \hat{\beta}_{X_j}^2} = \hat{\theta}_{IVW} \xrightarrow{m \rightarrow \infty} \theta + \frac{\sum_{j=1}^m \alpha_j \beta_{X_j}}{\sum_{j=1}^m \beta_{X_j}^2}, \tag{11}$$

showing that the IVW estimate is invariant to re-orientation of SNPs, and it is consistent when  $\sum_j \alpha_j \beta_{X_j}$  is zero. Under the InSIDE assumption in Eq (9), this holds when  $\bar{\boldsymbol{\alpha}} = 0$ , i.e., balanced pleiotropy, and/or  $\bar{\boldsymbol{\beta}}_X = 0$ . The latter condition explains why IVW still yielded unbiased estimates even in the case of directional pleiotropy in Simulation (a), where we generated  $\beta_{X_j}$  from a uniform distribution on  $(-0.2, -0.1) \cup (0.1, 0.2)$ .

In summary, the key issue here is that even though the InSIDE assumption holds for some unknown oracle coding, in general it does not for the default coding or an arbitrarily chosen coding, leading to inconsistent estimates in MR-Egger. Notably the InSIDE assumption is difficult to check [31, 32].

### The problem remains with the use of individual-level data

Instead of applying MR to GWAS summary data, one can apply IV regression to model (1) with individual-level data. With high-dimensional IVs, i.e. a large  $m$ , the InSIDE assumption is required [9]:  $\text{cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}_X) = 0$  (or more generally,  $\rightarrow 0$  as  $m \rightarrow \infty$ ) for some oracle coding. Using the same argument as before, if the SNPs/IVs are recoded with the corresponding  $\boldsymbol{\alpha}^*$  and  $\boldsymbol{\beta}_X^*$ , under general conditions we have  $\text{cov}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}_X^*) \neq 0$  (or  $\not\rightarrow 0$ ), leading to the violation of the InSIDE assumption and thus an inconsistent estimate. Corresponding to MR-Egger, we can implement the 2-stage IV regression by imposing  $\alpha_j \sim \mathcal{N}(r, \sigma_\alpha^2)$  iid. As shown in S1 Text, it was confirmed in the simulations that when applied to individual-level data, such an IV regression method behaved similarly to MR-Egger, yielding biased estimates of  $\theta$  for non-oracle coding schemes. In contrast, the method imposing  $r = 0$  performed similarly to IVW(RE), invariant to re-orientations of the SNPs.

### Results for other related methods

We note that other methods that incorporate Egger regression, such as MV-MR-Egger [14], LDA MR-Egger [33], MV-IWAS-Egger [34], PMR-Egger [35], mixIE [36], strictly speaking, would also inherit the limitations of MR-Egger, but possibly to varying extents. We applied both PMR-Egger and mixIE to simulated data as detailed in S1 Text. Under some general

conditions PMR-Egger might not perform well. On the other hand, mixIE was much more robust to various orientations of SNPs except when the proportion of invalid IVs was extremely high (e.g. close to 100%), which might be rare in practice. This is because mixIE depends on the IVW estimate (based on detected valid IVs) to a larger degree than on the MR-Egger estimate (based on detected invalid IVs), and often it could correctly identify valid IVs. Furthermore, mixIE is also more robust to mild to moderate violations of the InSIDE assumption [36].

We also applied some robust MR methods that are invariant to allele coding, including MR-cML [37] and MR-RAPS [12]. In simulations these methods performed well when the proportion of invalid IVs were not high; otherwise, e.g. when all IVs were invalid, only MR-Egger (with the oracle coding) performed well. The results are shown in [S1 Text](#).

### Irrelevant IVs

Albeit not the main point here, we point out a related issue that MR-Egger, or more specifically the InSIDE assumption, is not robust to the presence of irrelevant IVs (i.e. with the IV relevance assumption violated). Suppose that we mistakenly use  $m_0 \geq 1$  irrelevant IVs with  $\beta_{X_j} = 0$  for  $j = m + 1, \dots, m + m_0$ , in addition to  $m$  IVs used before. Even if the oracle coding is known and the InSIDE holds for the first  $m$  IVs, we have

$$(m + m_0)\text{cov}(\boldsymbol{\alpha}, \boldsymbol{\beta}_X) = \sum_{j=1}^{m+m_0} \alpha_j \beta_{X_j} - (m + m_0)\bar{\boldsymbol{\alpha}}\bar{\boldsymbol{\beta}}_X = \sum_{j=1}^m \alpha_j \beta_{X_j} - (m + m_0)\bar{\boldsymbol{\alpha}}\bar{\boldsymbol{\beta}}_X \neq 0, \quad (12)$$

because, by (9),  $\sum_{j=1}^m \alpha_j \beta_{X_j} = (\sum_{j=1}^m \alpha_j)(\sum_{j=1}^m \beta_{X_j})/m \neq (\sum_{j=1}^{m+m_0} \alpha_j)(\sum_{j=1}^m \beta_{X_j})/(m + m_0)$  in general unless under special cases such as  $\sum_{j=1}^m \beta_{X_j} = 0$ , or  $\sum_{j=1}^m \alpha_j = \sum_{j=1}^{m+m_0} \alpha_j = 0$ . Note that this conclusion holds regardless whether the irrelevant SNPs have direct effects or not (unless under some special cases). Hence, in general the InSIDE assumption would be violated if all  $m + m_0$  IVs are used, leading to an inconsistent estimate; this was confirmed in the simulations detailed in [S1 Text](#). Section G. This non-robustness property of MR-Egger is in contrast to some other methods, such as MR-RAPS [12] and MR-cML [37], whose consistency will not be influenced by the presence of a few irrelevant IVs (without and with pleiotropy respectively); see the conditions for Theorem 3.3 in [12], which also holds for cML. This is relevant because some authors [12] have advocated using various larger sets of IVs, possibly including some weak or irrelevant IVs, to increase the estimation efficiency and assess the robustness of a causal conclusion in an MR analysis, including from MR-Egger [38].

### Testing the intercept in MR-Egger

We also performed simulations to study the performance of testing the intercept term with the null hypothesis  $H_0: r = 0$  versus  $H_1: r \neq 0$  in MR-Egger [Eq \(6\)](#) using different SNP coding schemes. This test can be useful in suggesting the presence of invalid IVs with pleiotropic effect. We summarize our main findings here with all details given in [S1 Text](#). Section H. First, when all IVs are valid, the intercept term in [Eq \(6\)](#) is expected to be zero, no matter what coding scheme is used. Thus using any coding scheme could maintain a correct type I error rate. Second, the estimated intercept in MR-Egger based on the default coding tended to be larger (in absolute values) than those using other coding schemes, but the power could still be relatively low because of the lower precision of the estimate. Third, a non-zero intercept could mean several different things, such as the presence of correlated pleiotropy, the presence of uncorrelated directional pleiotropy, or both. In general the intercept term should not be simply interpreted as the average pleiotropic effect in practice [6]. As shown in Section H.3 in [S1 Text](#), in the

presence of correlated pleiotropy, even in the scenario where the pleiotropic effects of all invalid IVs were positive under the default coding, the estimated intercept could still be negative.

## Discussion

Although the phenomenon that MR-Egger depends on the coding of SNPs has been noticed and the default coding as a remedy has been recommended and widely applied [6, 14], to our best knowledge, there has been no other assessment and analysis of its implications and impact. In this paper, we have examined the influence of SNPs' coding on MR-Egger. Our findings could be summarized as follows. First, the current practice of orientating SNPs to be all positively associated with the exposure (referred as the default coding in MR-Egger) will be problematic unless its corresponding and coding-specific InSIDE assumption holds. Assuming that there is a true oracle coding under which the InSIDE assumption holds, the InSIDE assumption under the default (or another) coding will still hold and the current practice will yield a consistent estimate in the case of balanced pleiotropy (with respect to the oracle coding) and/or the SNP-exposure associations under the oracle coding have mean zero. When the SNP-exposure associations under the oracle coding have the same sign (all positive or negative), the default coding coincides with the oracle coding, thus will also give a consistent estimate; this is what is imposed in the current practice of MR-Egger, different from that the InSIDE assumption holds for some unknown oracle coding, under which, more generally and more likely, the current practice of applying MR-Egger will yield biased estimates. We also point out that this is not a finite-sample problem. In addition to our analysis, we have shown in the motivating real data examples and simulation studies that even with large  $n$  (and  $m$ ), this issue persisted. Second, even in the (special and unlikely) case when using the default coding in MR-Egger could still give consistent estimates, its variance is usually large because of the small ranges of SNP-exposure association effects after reorientation [6]. This also contributes to the low power with the default coding as noticed previously [37]. A small range of SNP-exposure associations would also magnify the degree of NOME violation, leading to a larger bias. Third, compared with another popular method IVW(RE), as shown in our simulation studies and analysis, the current practice of MR-Egger would only have an advantage when the default coding is the oracle coding and there is directional pleiotropy with respect to the default (oracle) coding. Fourth, most importantly, in practice we don't know the oracle coding under which the InSIDE assumption holds (if so) and the InSIDE assumption is very difficult to test, hence cautions should be taken when applying MR-Egger.

One may wonder whether the oracle coding, under which the InSIDE assumption holds, can be identified in practice. We tried a model selection approach to select the "best" coding for the Egger regression model (Eq (6)), but it did not always work, sometimes not only failing to select the best model but also leading to inflated type I errors. Alternatively, we tried to choose a coding scheme giving the minimum evidence of the violation of the InSIDE assumption. This turned out to be quite challenging too, involving a circular reasoning—to test whether a coding scheme under which the InSIDE assumption is satisfied requires a reliable/valid causal estimate, which however relies on the InSIDE assumption. It is noted that, under any non-oracle coding, although MR-Egger may give a biased estimate, the corresponding regression model is indeed "correct" in the usual sense that the specified regression model may still fit well the given data. A possible approach is to use one of other robust MR methods that do not require the InSIDE assumption to obtain a reliable causal estimate, such as MR-cML [37] and MRMix [39]. Then we could use this estimate to assess the InSIDE assumption for a given coding of SNPs. However, those methods have their own assumptions. In particular, when all the SNPs are invalid with directional pleiotropy, those methods all break down, while

MR-Egger with the oracle coding still works, as shown in [S1 Text](#). Furthermore, such a practice appears unnecessary if we feel confident in having already obtained a reliable causal estimate via another method.

Treating some nuisance parameters as random is a common and often effective way to reduce the number of the parameters to be estimated. For example, by modeling the direct effects  $\alpha_j$ 's as Normal random effects in MR-Egger, we do not need to estimate them but their mean  $r$  (and variance). In longitudinal and clustered data analysis, subject-specific effects are modeled as random in generalized linear mixed-effects models (GLMMs). However, modeling nuisance parameters as random usually imposes another important but largely neglected assumption: the distribution of the random effects is independent of other covariates; in MR-Egger, the other covariates are SNP-exposure associations  $\hat{\beta}_{x_j}$ , and the assumption is equivalent to the InSIDE assumption. The violation of this assumption can happen, leading to biased estimates in GLMMs [40]. [40] also showed that, by treating the subject-specific effects as fixed, instead of random, then applying a conditional likelihood approach (that eliminates the subject-specific effects from the conditional likelihood by conditioning on their sufficient statistics) avoids the problem. In the current context, if we treat the direct effects as fixed, the model is over-specified and the parameters are not estimable while it is unclear how to apply a conditioning argument; however, under other assumptions, notably that some  $\alpha_j = 0$  in the framework of MR-cML, one can allow the violation of the InSIDE assumption (and more generally allow correlated pleiotropy) for a subset of the IVs [37]. Furthermore, with a mean zero assumption on the random effects in GLMMs and IVW(RE), there is no issue of the dependence of the result on the coding of the covariates/SNPs; in contrast, it becomes problematic by assuming a non-zero mean of the random effects in MR-Egger.

In summary, we have studied the impact of SNP coding/orientation on Egger regression (and similar IV regression methods requiring the InSIDE assumption [9]). We emphasize that, since the InSIDE assumption is defined with respect to a specific coding scheme of the SNPs, even if it holds for some unknown (oracle) coding scheme, generally it does not hold for the default (exposure-increasing) coding (and many other codings) unless under some special and unlikely scenarios (such as when the default coding coincides with the oracle coding). The violation of the InSIDE assumption leads to the inconsistent estimator of the causal effect in MR-Egger. Thus, it is important for practitioners to keep in mind that, when applying MR-Egger with the default exposure-increasing allele coding, the interpretation of the causal effect estimate depends crucially on the non-violation of the InSIDE assumption under the default coding. We suggest that this SNP coding-specific assumption should be stated clearly when interpreting the results. How to fix the problem does not appear obvious since the InSIDE assumption is difficult to test and selecting the oracle coding is challenging. We have also confirmed that SNP coding in MR-Egger impacts the precision of the causal estimate as well as the extent of the NOME violation. The default coding gives the smallest range of IV-exposure associations, which tends to increase the variance of the causal estimate and magnify the degree of the NOME violation. Moreover, MR-Egger is not robust to outliers due to its use of the squared error loss function; it will be more robust to use other robust loss functions [12]. Until a better solution appears, we should be cautious when applying MR-Egger (and other related MR and IV regression methods for either GWAS summary or individual-level data) in data analysis.

## Supporting information

**S1 Text. Supplementary file with additional real data analysis results and additional simulation results.**

(PDF)



## Author Contributions

**Conceptualization:** Zhaotong Lin, Isaac Pan, Wei Pan.

**Data curation:** Zhaotong Lin, Isaac Pan.

**Formal analysis:** Zhaotong Lin, Isaac Pan, Wei Pan.

**Funding acquisition:** Wei Pan.

**Investigation:** Zhaotong Lin, Isaac Pan, Wei Pan.

**Methodology:** Zhaotong Lin, Wei Pan.

**Project administration:** Wei Pan.

**Resources:** Wei Pan.

**Software:** Zhaotong Lin.

**Supervision:** Wei Pan.

**Validation:** Zhaotong Lin.

**Visualization:** Zhaotong Lin.

**Writing – original draft:** Zhaotong Lin.

**Writing – review & editing:** Isaac Pan, Wei Pan.

## References

1. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*. 2003; 32(1):1–22. <https://doi.org/10.1093/ije/dym289>
2. Burgess S, Thompson SG. *Mendelian randomization: methods for using genetic variants in causal estimation*. CRC Press; 2015.
3. Zhu X. Mendelian randomization and pleiotropy analysis. *Quantitative Biology*. 2021; 9(2):122–132. <https://doi.org/10.1007/s40484-020-0216-3> PMID: 34386270
4. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology*. 2015; 44(2):512–525. <https://doi.org/10.1093/ije/dyv080> PMID: 26050253
5. Xue H, Pan W. Robust Inference of Bi-Directional Causal Relationships in Presence of Correlated Pleiotropy with GWAS Summary Data. *bioRxiv*. 2022;. <https://doi.org/10.1101/2022.02.28.482305> PMID: 35262077
6. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *European journal of epidemiology*. 2017; 32(5):377–389. <https://doi.org/10.1007/s10654-017-0255-x> PMID: 28527048
7. Hemani G, Zheng J, Elsworth B, Wade K, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife*. 2018; 7:e34408. <https://doi.org/10.7554/eLife.34408> PMID: 29846171
8. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International journal of epidemiology*. 2017; 46(6):1734–1739. <https://doi.org/10.1093/ije/dyx034> PMID: 28398548
9. Kolesár M, Chetty R, Friedman J, Glaeser E, Imbens GW. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*. 2015; 33(4):474–484. <https://doi.org/10.1080/07350015.2014.978175>
10. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*. 2013; 37(7):658–665. <https://doi.org/10.1002/gepi.21758> PMID: 24114802

11. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in medicine*. 2017; 36(11):1783–1802. <https://doi.org/10.1002/sim.7221> PMID: 28114746
12. Zhao Q, Wang J, Hemani G, Bowden J, Small DS. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*. 2020; 48(3):1742–1769. <https://doi.org/10.1214/19-AOS1866>
13. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I<sup>2</sup> statistic. *International journal of epidemiology*. 2016; 45(6):1961–1974. <https://doi.org/10.1093/ije/dyw220> PMID: 27616674
14. Rees JM, Wood AM, Burgess S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics in medicine*. 2017; 36(29):4705–4718. <https://doi.org/10.1002/sim.7492> PMID: 28960498
15. Burgess S, Thompson SG, genetics collaboration CC. Avoiding bias from weak instruments in Mendelian randomization studies. *International journal of epidemiology*. 2011; 40(3):755–764. <https://doi.org/10.1093/ije/dyr036> PMID: 21414999
16. Bowden J, Del Greco M F, Minelli C, Zhao Q, Lawlor DA, Sheehan NA, et al. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International journal of epidemiology*. 2019; 48(3):728–742. <https://doi.org/10.1093/ije/dyy258> PMID: 30561657
17. Bowden J, Spiller W, Del Greco M F, Sheehan N, Thompson J, Minelli C, et al. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. *International journal of epidemiology*. 2018; 47(4):1264–1278. <https://doi.org/10.1093/ije/dyy101> PMID: 29961852
18. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nature genetics*. 2013; 45(11):1274. <https://doi.org/10.1038/ng.2797> PMID: 24097068
19. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*. 2014; 46(11):1173–1186. <https://doi.org/10.1038/ng.3097> PMID: 25282103
20. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015; 518(7538):197–206. <https://doi.org/10.1038/nature14177> PMID: 25673413
21. Lu Y, Day FR, Gustafsson S, Buchkovich ML, Na J, Bataille V, et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature communications*. 2016; 7(1):1–15. <https://doi.org/10.1038/ncomms10495> PMID: 26833246
22. Horikoshi M, Beaumont RN, Day FR, Warrington NM, Kooijman MN, Fernandez-Tajes J, et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature*. 2016; 538(7624):248–252. <https://doi.org/10.1038/nature19806> PMID: 27680694
23. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature genetics*. 2018; 50(10):1412–1425. <https://doi.org/10.1038/s41588-018-0205-x> PMID: 30224653
24. Dupuis J, Langenberg C, Prokopenko I, Saxena R, Soranzo N, Jackson AU, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics*. 2010; 42(2):105–116. <https://doi.org/10.1038/ng.520> PMID: 20081858
25. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics*. 2019; 51(2):237–244. <https://doi.org/10.1038/s41588-018-0307-5> PMID: 30643251
26. van der Harst P, Verweij N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research*. 2018; 122(3):433–443. <https://doi.org/10.1161/CIRCRESAHA.117.312086> PMID: 29212778
27. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nature genetics*. 2018; 50(4):524–537. <https://doi.org/10.1038/s41588-018-0058-3> PMID: 29531354
28. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segre AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*. 2012; 44(9):981. <https://doi.org/10.1038/ng.2383> PMID: 22885922
29. Demenais F, Margerite-Jeannin P, Barnes KC, Cookson WO, Altmüller J, Ang W, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature genetics*. 2018; 50(1):42–53. <https://doi.org/10.1038/s41588-017-0014-7> PMID: 29273806

30. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature genetics*. 2020; 52(7):740–747. <https://doi.org/10.1038/s41588-020-0631-4> PMID: 32451458
31. Bowden J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. *International journal of epidemiology*. 2017; 46(6):2097–2099. <https://doi.org/10.1093/ije/dyx192> PMID: 29025021
32. Bowden J, Burgess S, Smith GD. Difficulties in testing the instrument strength independent of direct effect assumption in Mendelian randomization. *JAMA cardiology*. 2017; 2(8):929–930. <https://doi.org/10.1001/jamacardio.2017.1572> PMID: 28564679
33. Barfield R, Feng H, Gusev A, Wu L, Zheng W, Pasaniuc B, et al. Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genetic epidemiology*. 2018; 42(5):418–433. <https://doi.org/10.1002/gepi.22131> PMID: 29808603
34. Knutson KA, Deng Y, Pan W. Implicating causal brain imaging endophenotypes in Alzheimer's disease using multivariable IWAS and GWAS summary data. *NeuroImage*. 2020; 223:117347. <https://doi.org/10.1016/j.neuroimage.2020.117347> PMID: 32898681
35. Yuan Z, Zhu H, Zeng P, Yang S, Sun S, Yang C, et al. Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. *Nature communications*. 2020; 11(1):1–14. <https://doi.org/10.1038/s41467-020-17668-6> PMID: 32737316
36. Lin Z, Deng Y, Pan W. Combining the strengths of inverse-variance weighting and Egger regression in Mendelian randomization using a mixture of regressions model. *PLoS genetics*. 2021; 17(11): e1009922. <https://doi.org/10.1371/journal.pgen.1009922> PMID: 34793444
37. 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*. 2021; 108(7):1251–1269. <https://doi.org/10.1016/j.ajhg.2021.05.014> PMID: 34214446
38. Wang J, Zhao Q, Bowden J, Hemani G, Smith GD, Small DS, et al. Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS genetics*. 2021; 17(6):e1009575. <https://doi.org/10.1371/journal.pgen.1009575> PMID: 34157017
39. Qi G, Chatterjee N. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature communications*. 2019; 10(1):1–10. <https://doi.org/10.1038/s41467-019-09432-2> PMID: 31028273
40. Pan W. A note on the use of marginal likelihood and conditional likelihood in analyzing clustered data. *The American Statistician*. 2002; 56(3):171–174. <https://doi.org/10.1198/00031300292>