PLOS GENETICS

RESEARCH ARTICLE

# Combining the strengths of inverse-variance weighting and Egger regression in Mendelian randomization using a mixture of regressions model

**Zhaotong Lin, Yangqing Deng, Wei Pan***

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, United States of America

* panxx014@umn.edu





**Citation:** Lin Z, Deng Y, Pan W (2021) Combining the strengths of inverse-variance weighting and Egger regression in Mendelian randomization using a mixture of regressions model. PLoS Genet 17(11): e1009922. https://doi.org/10.1371/journal.pgen.1009922

**Editor:** Stephen Burgess, University of Cambridge, UNITED KINGDOM

**Received:** May 23, 2021

**Accepted:** November 2, 2021

**Published:** November 18, 2021



**Data Availability Statement:** The GWAS summary datasets used in the primary real data example are publicly available at https://static-content.springer.com/esm/art%3A10.1038%2Fs41588-020-0631-4/MediaObjects/41588_2020_631_MOESM3_ESM.xlsx. The GWAS summary datasets used in the secondary real data example are available in TwoSampleMR at https://mrcieu.github.io/TwoSampleMR/, and some example R code used to extract the GWAS summary data is given in Supplementary S4.1 at https://ars.els-cdn.com/content/image/1-s2.0-S0002929721002196-

## Abstract

With the increasing availability of large-scale GWAS summary data on various traits, Mendelian randomization (MR) has become commonly used to infer causality between a pair of traits, an exposure and an outcome. It depends on using genetic variants, typically SNPs, as instrumental variables (IVs). The inverse-variance weighted (IVW) method (with a fixed-effect meta-analysis model) is most powerful when all IVs are valid; however, when horizontal pleiotropy is present, it may lead to biased inference. On the other hand, Egger regression is one of the most widely used methods robust to (uncorrelated) pleiotropy, but it suffers from loss of power. We propose a two-component mixture of regressions to combine and thus take advantage of both IVW and Egger regression; it is often both more efficient (i.e. higher powered) and more robust to pleiotropy (i.e. controlling type I error) than either IVW or Egger regression alone by accounting for both valid and invalid IVs respectively. We propose a model averaging approach and a novel data perturbation scheme to account for uncertainties in model/IV selection, leading to more robust statistical inference for finite samples. Through extensive simulations and applications to the GWAS summary data of 48 risk factor-disease pairs and 63 genetically uncorrelated trait pairs, we showcase that our proposed methods could often control type I error better while achieving much higher power than IVW and Egger regression (and sometimes than several other new/popular MR methods). We expect that our proposed methods will be a useful addition to the toolbox of Mendelian randomization for causal inference.

## Author summary

For causal inference, inverse-variance weighting (IVW) and Egger regression are two of the most widely applied Mendelian randomization methods nowadays. IVW is the most powerful under the perhaps too restrictive assumption that all IVs are valid, while Egger regression is often unnecessarily too flexible in assuming all IVs to be invalid with uncorrelated pleiotropic effects. In spite of their usefulness, we point out their limitations: an

mmc1.pdf. The proposed methods are implemented in R package mixIE, which is publicly available on GitHub at https://github.com/ZhaotongL/mixIE. All other MR methods used for comparison are in publicly available R packages TwoSampleMR, MendelianRandomization (https://cran.r-project.org/web/packages/MendelianRandomization/index.html, MRMix (https://github.com/gqi/MRMix) and MRcML (https://github.com/xue-hr/MRcML).

**Funding:** ZL was supported by NIH grant R01 AG065636; YD was supported by NSF grant DMS 1711226; WP was supported by NIH grants R01 AG069895, RF1 AG067924, R01 AG065636, R01 HL116720, R01 GM113250 and R01 GM126002, and by NSF grant DMS 1711226. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

IVW estimate of a causal effect would be biased if some/all IVs have directional pleiotropic effects, and an Egger regression estimate has too large a variance, leading to its loss of power. Accordingly we propose a mixture model to combine them to take advantage of their strengths while overcoming their major limitations. Furthermore, we propose a model-averaging approach and a novel data perturbation scheme to account for uncertainties in model/IV selection, leading to more robust statistical inference. Through simulations and applications to some publicly available large-scale GWAS summary data, we demonstrate the superiority of our methods over IVW and Egger regression (and over some other state-of-the-art MR methods in some scenarios).

## Introduction

Mendelian randomization (MR) has become a widely used technique to infer causal relationship between an exposure (e.g. a risk factor) and an outcome (e.g. a disease) using GWAS summary data, in which usually independent genetic variants (SNPs) are used as instrument variables (IVs) [1–3]. To guarantee correct inference, as shown in Fig 1, a valid IV used in MR must be:

1. associated with the exposure ($X$), i.e. $\gamma_i \neq 0$;
2. not associated with the outcome ($Y$) conditional on the exposure ($X$) and hidden confounder ($U$), i.e. $\alpha_i = 0$;
3. not associated with any hidden confounder ($U$), i.e. $\phi_i = 0$.

When all IVs used in Mendelian randomization are valid (and uncorrelated), the inverse-variance weighted (IVW) method is consistent and most powerful: it combines the IV-specific ratio estimates most efficiently by inverse-variance weighting [4]. However, in the presence of horizontal pleiotropy, where some or all IVs have direct effects on the outcome, the second IV

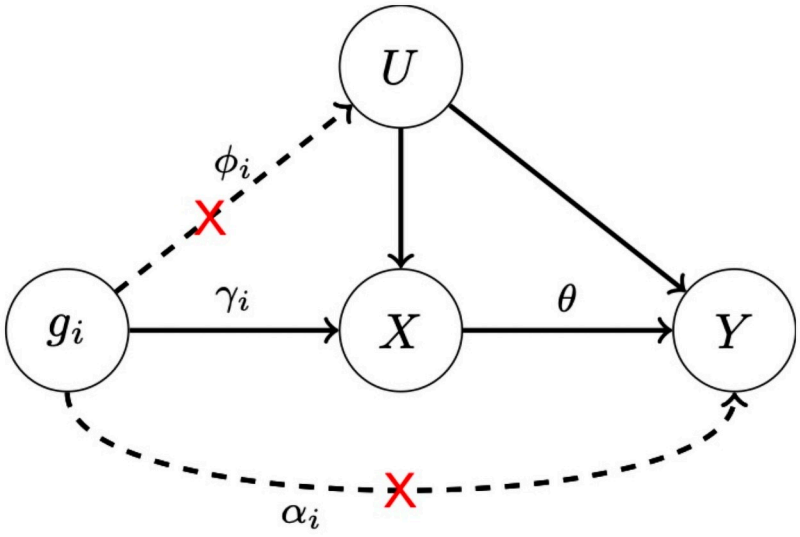


**Fig 1. A causal diagram illustrating the three assumptions on a valid IV $g_i$.** Dashed lines (which are marked with a red 'cross') correspond to violations of assumptions 2 and 3.

https://doi.org/10.1371/journal.pgen.1009922.g001

assumption is violated and IVW may not be consistent unless the mean of the direct effects is zero, a scenario of so-called "balanced pleiotropy". More generally, Egger regression is applied under a weaker assumption that the direct or pleiotropic effects of the genetic variants on the outcome are independent of the genetic associations with the exposure (so-called InSIDE assumption) [5]. It has been noted recently that Egger regression often suffers from a severe loss of power, because it assumes that all IVs are invalid, which may be too extreme. Many other methods have been proposed to deal with the violation of Assumption 2 or both Assumptions 2 and 3, but most of them require either the plurality or the majority assumption [6–9]. We note that Egger regression is the only method allowing all IVs to be invalid (with possibly directional pleiotropy), and is easy to apply, both of which perhaps explain its popularity.

Both IVW and Egger regression, as two of the most popular MR methods, impose too extreme assumptions: while IVW (fixed effects) assumes that all IVs are valid, both IVW (random effects) and Egger regression assume that all IVs are invalid; the truth is perhaps often between the two. We acknowledge and take account of the possibility of having from zero to all invalid IVs. Accordingly, we propose a two-component mixture of regressions model, denoted **mixIE**, which can be viewed as the mixture of IVW (fixed effects) and Egger regression, and thus may be more efficient and more robust to the violation of IV Assumption 2; as Egger regression, the proposed new method requires the IV Assumption 3 (so that InSIDE) to hold, though we will show that it is more robust than Egger regression when the assumption is violated. The model is fitted using a classification expectation-maximization (CEM) algorithm [10, 11], selecting valid and invalid IVs to be used by IVW (fixed effects) and Egger regression respectively. To account for uncertainties in model/IV selection, we propose a model-averaging approach based on a few top selected models, in addition to the default IVW (fixed effects) and Egger regression models. Furthermore, we propose a novel data perturbation scheme to deal with more challenging situations where the true model is only weakly identifiable, e.g. with many IVs having weak pleiotropic effects; it controls the type I error better in these challenging situations.

There are two other related methods based on mixture models as well, namely MR-ContMix [12] and MRMix [13]. MRMix assumes a four-component normal mixture model for the underlying bivariate effect size distribution of all the SNPs for a pair of traits, and MR-ContMix assumes a two-component normal mixture model for the ratio estimates (corresponding to valid and invalid IVs), while our proposed method is based on a mixture of *regressions* model for GWAS summary statistics directly. All three mixture models include a component for valid IVs and could potentially identify valid and invalid IVs. Another novel and powerful method based on constrained maximum likelihood (cML) also selects and (implicitly) removes invalid IVs by estimating the pleiotropic effect (if it exists) for each SNP and then only uses valid IVs for statistical inference [14]. However, [14] shows in their simulations that although MR-ContMix performed well in most scenarios, sometimes it performed poorly probably due to the challenge of its pre-selection of a fixed tuning parameter. At the same time, MRMix often performed well with mostly well controlled type 1 errors and high power, but it might have either largely inflated or too conservative type 1 errors while giving biased estimates. Here we will further compare mixIE with cML, MR-ContMix and MRMix in our numerical studies. At last, MRMix, MR-ContMix and cML all require the valid IV plurality condition, while mixIE, as Egger regression, could potentially handle situations when all IVs are invalid but requiring the InSIDE assumption to hold; we also show that mixIE is more robust than Egger regression when the InSIDE assumption is violated.

The rest of the paper is organized as follows. We first introduce mixIE and its variants based on model averaging and data perturbation. We then show the advantages of our proposed methods over IVW and Egger regression through extensive simulations. Finally we apply and compare various methods on 48 risk factor-disease pairs and 63 trait pairs using multiple publicly available GWAS summary datasets.

## Methods

### Overview

We consider the setup of two-sample Mendelian randomization. Suppose that we have $m$ independent SNPs, $G_1, \ldots, G_m$, as IVs, an exposure $X$, an outcome $Y$, and some hidden confounder $U$; $\theta$ is the causal effect of the exposure on the outcome and is the parameter of interest. The coefficient vector $\boldsymbol{\gamma}_{xg}$ represents the associations between SNPs and the exposure, and $\boldsymbol{\alpha}_{yg}$ the direct effects of SNPs on the outcome that are not mediated through the exposure. The true model is

$$\begin{aligned} \mathbf{X} &= \mathbf{G}\boldsymbol{\gamma}_{xg} + \mathbf{U} + \mathbf{e}_x, \\ \mathbf{Y} &= \theta\mathbf{X} + \mathbf{G}\boldsymbol{\alpha}_{yg} + \mathbf{U} + \mathbf{e}_y. \end{aligned} \tag{1}$$

where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{G}$ are the vectors for the observed exposure, outcome and genotypic scores respectively, and $\mathbf{e}_x$ and $\mathbf{e}_y$ are the vectors for independent errors.

Given two independent GWAS summary datasets of traits $X$ and $Y$ with sample sizes $n_x$ and $n_y$ respectively, we extract the data for $m$ independent SNPs that are significantly associated with $X$; that is, their marginal association parameter and variance estimates, $(\hat{\beta}_{Xi}, \hat{\sigma}^2_{Xi})$ and $(\hat{\beta}_{Yi}, \hat{\sigma}^2_{Yi})$, $i = 1, \ldots, m$. The IVW model is

$$\hat{\beta}_{Yi} = \theta\hat{\beta}_{Xi} + \epsilon_{Ii}; \quad \epsilon_{Ii} \sim \mathcal{N}(0, \sigma_I^2\hat{\sigma}^2_{Yi}).$$

Fitting the above model using weighted least squares lead to the IVW estimate $\hat{\theta}_{IVW}$, which is consistent for $\theta$ if all IVs are valid or have balanced pleiotropy. If all IVs are valid, $\sigma_I^2 = 1$, which is specified under the fixed-effect (FE) IVW; on the other hand, if some IVs are invalid but have balanced pleiotropy, $\sigma_I^2 > 1$ is estimated under the random-effect (RE) IVW. Note that IVW(FE) and IVW(RE) assume that all IVs are valid and invalid respectively. Throughout this paper, if IVW is used alone (as a single MR method), it always refers to its RE version; however, in our proposed mixIE, IVW(FE) is always used.

Egger regression is a simple modification to the above linear model without constraining the intercept to be zero:

$$\hat{\beta}_{Yi} = r + \theta\hat{\beta}_{Xi} + \epsilon_{Ei}; \quad \epsilon_{Ei} \sim \mathcal{N}(0, \sigma_E^2\hat{\sigma}^2_{Yi}). \tag{2}$$

As pleiotropic effects of genetic variants will lead to overdispersion, a random-effects analysis should always be preferred by estimating $\sigma_E^2 \geq 1$ in Egger regression [15]. Under the InSIDE assumption, the Egger estimate $\hat{\theta}_E$ of $\theta$ is consistent (as both the sample size and the number of genetic variants tend to infinity).

To take the advantage of both IVW and Egger regression, we propose a two-component mixture regression model for IVW and Egger regression respectively, called **mixIE** for short. The two components correspond to modeling valid IVs and invalid IVs respectively. If we know which IVs are valid and which are invalid, we would use the

corresponding IVW(FE) and Egger regression models respectively. In practice, since we do not know, we have a mixture of two regressions and the log-likelihood function (up to a constant) as

$$\begin{aligned} l(\theta, r, c, \pi; \{\hat{\beta}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{Yi}\}) &= \sum_{i=1}^{K} -\frac{1}{2}\left[\frac{(\hat{\beta}_{Yi} - \theta\hat{\beta}_{Xi} - r)^2}{c\hat{\sigma}^2_{Yi}} + \log(c\hat{\sigma}^2_{Yi})\right] \\ &\quad + \sum_{i=K+1}^{m} -\frac{1}{2}\left[\frac{(\hat{\beta}_{Yi} - \theta\hat{\beta}_{Xi})^2}{\hat{\sigma}^2_{Yi}} + \log(\hat{\sigma}^2_{Yi})\right], \end{aligned} \tag{3}$$

where, among the unknown parameters, $r$ is the average pleiotropic effect of invalid IVs, $c$ is the multiplicative over-dispersion parameter, and $\pi = K/m$ is the proportion of invalid IVs; we use $\{\hat{\beta}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{Yi}\}$ to represent the observed data. Here for simplicity of notation, we denote the first $K$ SNPs, $G_1, \ldots, G_K$, as invalid IVs, and the remaining $G_{K+1}, \ldots, G_m$ as valid IVs. In practice, we do not know which IVs are invalid (and all the parameter values), thus we apply the classification expectation-maximization (CEM), a variant of EM algorithm, to classify IVs (and estimate the parameters) [10, 11]. Since the result of EM or CEM depends on the choice of starting values and there may be multiple models fitting the data well, we propose a model averaging approach, called **mixIE-MA**. We also propose a data perturbation strategy to further take account of the uncertainty in model/IV selection, called **mixIE-MA-DP** later.

## Model fitting

The formulation of our approach lends itself to the classification EM algorithm [10]. Let $z_i$ denote the unobserved indicator of whether IV/SNP $i$ being invalid or not. For convenience, let $\boldsymbol{\alpha} = (\theta, r, c, \pi)$ denote the set of all unknown parameters, and $\mathbf{D}_i = (\hat{\beta}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{Yi})$ the observed data for SNP $i$. The $(t+1)$th iteration of CEM algorithm is defined as follows:

- E-step: calculate

$$\begin{aligned} \tau^{(t+1)}_{i,0} &:= P(z_i = 0 | \mathbf{D}_i; \boldsymbol{\alpha}^{(t)}) \\ &= \frac{f(\hat{\beta}_{Yi} - \theta^{(t)}\hat{\beta}_{Xi}, \hat{\sigma}^2_{Yi}) \cdot (1 - \pi^{(t)})}{f(\hat{\beta}_{Yi} - \theta^{(t)}\hat{\beta}_{Xi}, \hat{\sigma}^2_{Yi}) \cdot (1 - \pi^{(t)}) + f(\hat{\beta}_{Yi} - \theta^{(t)}\hat{\beta}_{Xi} - r^{(t)}, c^{(t)}\hat{\sigma}^2_{Yi}) \cdot \pi^{(t)}}, \\ \tau^{(t+1)}_{i,1} &:= P(z_i = 1 | \mathbf{D}_i; \boldsymbol{\alpha}^{(t)}) = 1 - \tau^{(t+1)}_{i,0}, \end{aligned}$$

  where $f(a, \sigma^2)$ is the density function value at $a$ for $\mathcal{N}(0, \sigma^2)$.

- C-step: classify SNP $i$ as invalid IV if $\tau^{(t+1)}_{i,1} \geqslant 0.5$, otherwise as valid IV, and let $\hat{K}$ denote the number of the classified invalid IVs. Again, for simplicity of notation, after possible rearrangement of the orders of the SNPs, we denote the first $\hat{K}$ SNPs as invalid IVs and the rest $m - \hat{K}$ as valid IVs.

- M-step: update the parameter estimates

$$\begin{aligned}
\pi^{(t+1)} &= \frac{\hat{K}}{m}, \\
r^{(t+1)} &= \frac{\sum_{i=1}^{\hat{K}} \frac{\hat{\beta}_{Yi} - \theta^{(t)}\hat{\beta}_{Xi}}{\hat{\sigma}^2_{Yi}}}{\sum_{i=1}^{\hat{K}} \frac{1}{\hat{\sigma}^2_{Yi}}}, \\
c^{(t+1)} &= \frac{\sum_{i=1}^{\hat{K}} \frac{(\hat{\beta}_{Yi} - \theta^{(t)}\hat{\beta}_{Xi} - r^{(t+1)})^2}{\hat{\sigma}^2_{Yi}}}{\hat{K}}, \\
\theta^{(t+1)} &= \frac{\sum_{i=1}^{\hat{K}} \frac{(\hat{\beta}_{Yi} - r^{(t+1)})\hat{\beta}_{Xi}}{c^{(t+1)}\hat{\sigma}^2_{Yi}} + \sum_{i=\hat{K}+1}^{m} \frac{\hat{\beta}_{Yi}\hat{\beta}_{Xi}}{\hat{\sigma}^2_{Yi}}}{\sum_{i=1}^{\hat{K}} \frac{\hat{\beta}^2_{Xi}}{c^{(t+1)}\hat{\sigma}^2_{Yi}} + \sum_{i=\hat{K}+1}^{m} \frac{\hat{\beta}^2_{Xi}}{\hat{\sigma}^2_{Yi}}}.
\end{aligned}$$

We obtain the final estimates $(\hat{\theta}, \hat{r}, \hat{c}, \hat{\pi}) = (\theta^{(t+1)}, r^{(t+1)}, c^{(t+1)}, \pi^{(t+1)})$ at the convergence. By default, as in our simulations and real data analyses, we set $r^{(0)} = 0$, $c^{(0)} = 1$, $\pi^{(0)} = 0.2$ and generate $\theta^{(0)}$ randomly from $-\max(|\hat{\beta}_{Yi}/\hat{\beta}_{Xi}|)$ to $\max(|\hat{\beta}_{Yi}/\hat{\beta}_{Xi}|)$ as well as including 0 and point estimates from IVW and Egger regression. The choice of $\pi^{(0)} = 0.2$ is because that there is usually a small proportion of invalid IVs in many real data examples as shown later and in [14]. When $r^{(0)} = 0$ and $c^{(0)} = 1$, the two components are the same, thus we run two iterations of the standard EM algorithm before starting the CEM.

To obtain the standard error of the estimated parameters, we take the approach proposed in [16], which requires the computation of the gradients and the information matrix based on the complete data log-likelihood. The details are given in Section A in S1 Text.

## Model averaging

It is well known that the EM algorithm is sensitive to the choice of starting values while there may be multiple good models, therefore we use model averaging to account for this uncertainty. Specifically, we use different starting values to reach a few top candidate models, and by default, we always include the fixed-effect IVW model (i.e. $\hat{\pi} = 0, \hat{c} = 1, \hat{r} = 0$) and the Egger regression model (i.e. $\hat{\pi} = 1$) in the list of the top candidate models. A model is judged by its Bayesian information criterion (BIC) [17]:

$$BIC = -2 \cdot l(\hat{\theta}, \hat{r}, \hat{c}, \hat{\pi}; \{\hat{\beta}_{Xi}, \hat{\beta}_{Yi}, \hat{\sigma}^2_{Yi}\}) + \log(n_y) \cdot (2 + \mathbb{1}\{\hat{r} \neq 0\} + \mathbb{1}\{\hat{c} > 1\}),$$

where the indicator function $\mathbb{1}\{A\} = 1$ or 0, depending on whether $A$ is true or not. Based on the BIC values of the various fitted models, we select up to 5 top models. Following [18], we define the weight for model $k = 1, \ldots, 5$ with BIC value $BIC_k$ as

$$w_k = \frac{\exp(-BIC_k/2)}{\sum_{j=1}^{5} \exp(-BIC_j/2)}.$$

Now we combine the estimates $\hat{\theta}_k$ from candidate models $k = 1, \ldots, 5$ to have the final model-

averaging estimate and its standard error as

$$\hat{\theta}_{MA} = \sum_k w_k \hat{\theta}_k,$$
$$SE(\hat{\theta}_{MA}) = \sum_k w_k \sqrt{SE(\hat{\theta}_k)^2 + (\hat{\theta}_k - \hat{\theta}_{MA})^2}.$$

## Data perturbation

To further take account of the uncertainty in model selection/averaging, we propose a data perturbation strategy [19]. Instead of applying the usual asymptotics that ignores the uncertainty in model selection/averaging, we perturb the data to mimic generating multiple samples from the data distribution. With each perturbed/generated sample, we repeat an estimation procedure (as applied to the original data) so that the uncertainty in model selection/averaging or other aspects can be taken account when, at the end, the empirical distribution of such estimates from multiple perturbed samples is used for inference. This is similar to and can even trace back to the little bootstrap method [20] in the context of model selection. We found that the (more general) data perturbation scheme proposed in [14] did not perform well for mixIE in some extreme situations (presumably because of the nature of Egger regression with the use of invalid IVs), so we propose a modified one for mixIE specifically.

Since mixIE classifies IVs as valid or invalid, estimating the causal effect $\hat{\theta}$ could be approximated via a fixed effect meta analysis of IVW estimate $\hat{\theta}_I$ and Egger estimate $\hat{\theta}_E$ based on the sets of valid IVs and of invalid IVs respectively:

$$\hat{\theta} = \left(\frac{\hat{\theta}_I}{SE(\hat{\theta}_I)^2} + \frac{\hat{\theta}_E}{SE(\hat{\theta}_E)^2}\right) \Big/ \left(\frac{1}{SE(\hat{\theta}_I)^2} + \frac{1}{SE(\hat{\theta}_E)^2}\right). \quad (4)$$

Similarly for mixIE-MA, we could classify IVs based on their model averaged posterior probabilities. Accordingly, we propose the following data perturbation scheme for mixIE-MA:

Step 1: Perturb the observed data by $\hat{\beta}_{Yi}^{(b)} = \hat{\beta}_{Yi} + \epsilon_i^*$, where $\epsilon_i^* \sim N(0, \hat{\sigma}_{Yi}^2)$, and $\hat{\sigma}_{Yi}^{(b)} = \sqrt{\text{var}(\hat{\beta}_{Yi}^{(b)})} = \sqrt{2} \cdot \hat{\sigma}_{Yi}$.

Step 2: Apply mixIE-MA algorithm on the $b$-th perturbed dataset $(\hat{\beta}_{Xi}, \hat{\beta}_{Yi}^{(b)}, \hat{\sigma}_{Yi}^{(b)})$ and obtain the corresponding estimated sets of valid IVs and invalid IVs. The estimated causal effect $\hat{\theta}^{(b)}$ could be approximated by Eq (4) with $(\hat{\theta}_I^{(b)}, SE(\hat{\theta}_I^{(b)}))$ and $(\hat{\theta}_E^{(b)}, SE(\hat{\theta}_E^{(b)}))$ estimated from IVW and Egger regression respectively.

Step 3: Further perturb the data for the set of invalid IVs by $\hat{\beta}_{Yi}^{(b)*} = \hat{\beta}_{Yi} + \hat{\sigma}^{(b)}\epsilon_i^*$, where $\hat{\sigma}^{(b)}$ is the estimated inflation factor in Egger regression model (2) for the set of invalid IVs identified above.

Step 4: Apply Egger regression to the new perturbed data $\hat{\beta}_{Yi}^{(b)*}$ and obtain $\hat{\theta}_E^{(b)*}$.

Step 5: Apply fixed-effect meta analysis to combine $\hat{\theta}_I^{(b)}$ and $\hat{\theta}_E^{(b)*}$, obtaining $\hat{\theta}^{(b)*}$ for the $b$-th perturbed dataset as follows:

$$\hat{\theta}^{(b)*} = \left(\frac{\hat{\theta}_I^{(b)}}{SE(\hat{\theta}_I^{(b)})^2} + \frac{\hat{\theta}_E^{(b)*}}{SE(\hat{\theta}_E^{(b)})^2}\right) \Big/ \left(\frac{1}{SE(\hat{\theta}_I^{(b)})^2} + \frac{1}{SE(\hat{\theta}_E^{(b)})^2}\right).$$

We repeat the above steps $B$ times and obtain the mean and standard deviation of $\{\hat{\theta}^{(b)*}\}_{b=1}^{B}$ as the final point estimate and standard error for mixIE-MA-DP. We could also use the proportion of the times when SNP $i$ being identified as invalid out of the $B$ data perturbations as the posterior probability estimate $\hat{\tau}_{i,1}$ for mixIE-MA-DP. By default we used $B = 200$ in our simulations and real data analysis.

It is noted that the DP on GWAS *summary* data is equivalent to the bootstrap on the corresponding *individual-level* data. Suppose the GWAS individual level data are $\{(Y_j, Z_{ij}): j = 1, 2, \ldots, n\}$ for outcome $Y$ and SNP/IV $i$ (both centered at mean 0). For model $Y_j = Z_{ij}\beta_{Yi} + e_{Yij}$ with iid $e_{Yij} \sim N(0, \sigma_{ei}^2)$, we obtain the ordinary least square (OLS), or equivalently maximum likelihood, estimate $\hat{\beta}_{Yi} \sim N(\beta_{Yi}, \sigma_{Yi}^2)$. If we apply the parametric bootstrap by generating $Y_j^{(b)} = Z_{ij}\hat{\beta}_{Yi} + e_{Yij}^{(b)}$ with iid $e_{Yij}^{(b)} \sim N(0, \sigma_{ei}^2)$, it is easy to verify that, conditional on $\hat{\beta}_{Yi}$, the corresponding OLS estimate $\hat{\beta}_{Yi}^{(b)} \sim N(\hat{\beta}_{Yi}, \sigma_{Yi}^2)$, the same distribution (with $\sigma_{Yi}^2$ replaced by its estimate $\hat{\sigma}_{Yi}^2$) used in Step 1 in our DP procedure. In fact, based on the results in [21] (Section 7.2.2), the conclusion holds (asymptotically) for other three types of the bootstrap: residual bootstrap (by resampling residuals), nonparametric bootstrap (by resampling pairs $(Y_j, Z_{ij})$) and external/wild/multiplier bootstrap.

## Goodness-of-fit testing

In mixIE, invalid IVs are modeled according to Egger regression, which requires the InSIDE assumption. Thus, in principle mixIE and the proposed data perturbation scheme would also require the InSIDE assumption to hold. On the other hand, it is known that the InSIDE assumption is difficult to test [22–24]. A general way for model checking is to compare our proposed model with some other methods that do not require the InSIDE assumption, such as cML, MRMix and MR-ContMix, though it only works if other assumptions for the latter methods hold (e.g. the plurality of valid IV assumption). Specifically, we evaluate the consistency between our estimates with that of another method; any inconsistency might be due to the violation of the InSIDE or other assumptions of the two methods being compared. Here we call it goodness-of-fit (GOF) testing, though it is perhaps more in line with and can be used for triangulation [25].

To compare two different methods on a given dataset, one has to account for the correlation between the two estimates from the two methods, which is not trivial. We propose using a general data perturbation scheme similar to that of [14] (but under no measurement error (NOME) assumption as adopted by IVW and Egger regression, thus by mixIE) for such a purpose. Specifically, we propose the following procedure: Staring with $b = 1$,

Step 1: Perturb the data to generate $\hat{\beta}_{Yi}^{(b)} = \hat{\beta}_{Yi} + \epsilon_{Yi}^{*}$ with $\epsilon_{Yi}^{*} \sim N(0, \hat{\sigma}_{Yi}^2)$ independently;

Step 2: Apply mixIE-MA and another method, such as cML-MA, to the perturbed dataset $(\hat{\beta}_{Xi}, \hat{\beta}_{Yi}^{(b)}, \hat{\sigma}_{Yi}^2)$ and obtain the corresponding $\theta_m^{(b)}$ and $\theta_c^{(b)}$;

Step 3: Calculate their difference, $\delta^{(b)} = \theta_m^{(b)} - \theta_c^{(b)}$, and let $b \leftarrow b + 1$.

We repeat the above steps for $B = 200$ times and obtain the empirical distribution of $\{\delta^{(b)*}\}_{b=1}^{B}$. If the 95% percentile interval of $\{\delta^{(b)*}\}$ does not cover 0, then we conclude that the results of mixIE-MA and the other method (not requiring InSIDE) are inconsistent with each other, suggesting possible violation of the InSIDE assumption for mixIE, or of some other assumptions required by the two methods. In such a case, cautions should be taken in interpreting the causal results.

## Other MR methods

We compared mixIE-MA and mixIE-MA-DP with other popular two-sample MR methods, including random-effect IVW model, Egger regression, MR-Mix [13], MR-ContMix [12], weighted-median [6] and a new method called constrained maximum likelihood (cML) [14]. We applied cML with its model averaging version based on BIC, called cML-MA for short, and its data perturbation version, called cML-MA-DP.

## GWAS data

**Primary real data example.** We applied our proposed methods and other MR methods to 48 pairs of risk factor-disease GWAS summary data following [26], including 12 risk factors and 4 diseases. For each risk factor-disease pair, we used the set of LD-independent SNPs as IVs as described in [26] (in their S4 Table), and applied all methods to the GWAS summary statistics of these SNPs.

**Secondary real data example.** Following [14], we also applied our proposed methods to 63 trait pairs whose genetic correlations are not significant, suggesting that they are unlikely to be causally related. For each pair, we used the code provided in its Supplementary to extract the GWAS summary statistics for analysis.

## Simulation set-ups

**Main simulations.** Similar to the set-ups in [27], we simulated data according to Eq (5),

$$\begin{aligned} \mathbf{U} &= \mathbf{G}\boldsymbol{\phi}_{ug} + \mathbf{e}_u, \\ \mathbf{X} &= \mathbf{G}\boldsymbol{\gamma}_{xg} + \mathbf{U} + \mathbf{e}_x, \\ \mathbf{Y} &= \theta\mathbf{X} + \mathbf{G}\boldsymbol{\alpha}_{yg} + \mathbf{U} + \mathbf{e}_y, \end{aligned} \tag{5}$$

in which

1. the genotype scores of $m$ SNPs/IVs ($\mathbf{G}$) were generated independently from Binomial(2,$f_i$), where for each SNP $i$ its MAF $f_i$ was generated independently from a uniform distribution $\mathcal{U}(0.1, 0.3)$;
2. the IV strengths $\boldsymbol{\gamma}_{xg}$ were generated from a left-truncated normal distribution: for $m = 10$, an IV strength was generated from $\mathcal{N}(0, 0.15^2)$ left-truncated at 0.15 (i.e. any value generated would be larger than 0.15); for $m = 30$, it was generated from $\mathcal{N}(0, 0.1^2)$ left-truncated at 0.1; for $m = 100$, it was from $\mathcal{N}(0, 0.05^2)$ left-truncated at 0.05;
3. for $K = m \cdot$ (p_invalid) invalid IVs, we consider three scenarios:

(a). Balanced pleiotropy and InSIDE satisfied, where pleiotropy effects $\boldsymbol{\alpha}_{yg}$ were generated independently from $\mathcal{N}(0, \sqrt{0.15}^2)$ and $\phi_{ug} = 0$;

(b). Directional pleiotropy and InSIDE satisfied, where $\boldsymbol{\alpha}_{yg}$ were generated from $\mathcal{N}(0.1, \sqrt{0.075}^2)$ and $\phi_{ug} = 0$;

(c). Directional pleiotropy and InSIDE violated, where $\boldsymbol{\alpha}_{yg}$ were generated from $\mathcal{N}(0.1, \sqrt{0.075}^2)$ and $\phi_{ug}$ were generated from $\mathcal{U}(0, b)$;

4. $\mathbf{e}_u$, $\mathbf{e}_x$, $\mathbf{e}_y$ were generated from $\mathcal{N}(0, 1)$ independently.

The summary data for genetic associations were calculated for the exposure and the outcome on non-overlapping sets of individuals, each consisting of $n$ individuals. For scenarios

(a) and (b), we varied $\theta$ from {0,0.2}, p_invalid from {0, 0.3, 0.5, 0.7, 1} and sample size $n_x = n_y = n = 10\,000$ or 50 000. To further compare the power, we also tried a smaller effect size $\theta$ from {±0.01, ±0.05, ±0.1, ±0.15}, p_invalid from {0.3,0.5,0.7} for $m = 30$ and $n = 50\,000$. For scenario (c), we considered different correlated pleiotropy effects by varying $b$ from {0.1,0.4,0.7}.

As one reviewer suggested, we also considered different sample sizes for the exposure and the outcome. Details are given in Section B.1.4 in S1 Text.

**Secondary simulations with weak invalid IVs.** Following [14], we simulated data with many invalid IVs with weak effects, called "weak invalid IVs" throughout. We generated $m = 50$ IVs with 60% invalid IVs and with sample size $n = 20\,000$. The IV strengths $\gamma_i$'s were generated independently from $\mathcal{N}(0, 0.01)$. Pleiotropic effects $\alpha_i$'s for the first 30 IVs were generated independently from $\mathcal{N}(0, h_y/m)$. Then we set $\hat{\sigma}_{Xi} = \hat{\sigma}_{Yi} = 1/\sqrt{n}$ and generated GWAS summary data $\hat{\beta}_{Xi} \sim \mathcal{N}(\gamma_i, \hat{\sigma}^2_{Xi})$ and $\hat{\beta}_{Yi} \sim \mathcal{N}(\theta \cdot \gamma_i + \alpha_i, \hat{\sigma}^2_{Yi})$, where $\theta$ is the causal effect of interest. We varied $h_y$ from {0.1,0.2,0.4} and $\theta$ from {-0.2,-0.1,-0.05,0,0.05,0.1,0.2}.

For each setup, we did 1000 simulations. We compare our proposed methods mixIE-MA and its data perturbation version mixIE-MA-DP with cML-MA, cML-MA-DP, Egger, IVW, MR-Mix, MR-ContMix and weighted-median. For mixIE-MA, mixIE-MA-DP and Egger regression, we used the original coding of SNPs throughout the simulations.

# Results

## Simulations

**Main simulations.** We compared our proposed methods with 6 most popular and new MR methods under three scenarios: InSIDE satisfied, directional pleiotropy; InSIDE satisfied, balanced pleiotropy and InSIDE violated, directional pleiotropy. Here we only show some representative results for $n = 50000$ while all others are given in Section B.1 in S1 Text.

**InSIDE satisfied.** Fig 2 shows the empirical type 1 error and power of different methods for directional pleiotropy under the InSIDE assumption. First, Egger regression was the only method that could control type 1 error across all scenarios, followed by mixIE-MA-DP and cML-MA-DP that could control type 1 error well except in the more extreme scenarios with all IVs being invalid. But Egger regression also had the lowest power across all scenarios. As expected, IVW had inflated type 1 error in the presence of directional pleiotropy (but not in that of balanced pleiotropy). However, as shown in Section B.1.2 in S1 Text, IVW had very low power in the scenarios of balanced pleiotropy even though it could control the type 1 error, while mixIE-MA-DP was able to control the type 1 error well and had much higher power except when all IVs were invalid. Our proposed method mixIE-MA had inflated type 1 error as the proportion of invalid IVs increased, but this could be improved by mixIE-MA-DP with data perturbation; the latter point was also reflected by cML-MA and cML-MA-DP. MR-Mix was able to control type 1 error in many scenarios except when $m = 10$ and/or all IVs were invalid, and it also performed too conservatively in the cases with all valid IVs. MR-ContMix performed well with 30% invalid IVs but began to have inflated type 1 error when more than 50% IVs were invalid; and weighted-median also had much inflated type 1 error when there were more than half of invalid IVs probably due to the violation of its majority of valid IV assumption. Second, data perturbation was able to further take account of the uncertainty in model selection. We see that mixIE-MA could have inflated type 1 error as the proportion of invalid IVs increased, probably because of incorrectly classifying IVs. In contrast, mixIE-MA-DP was able to control type 1 error in all scenarios but that of p_invalid = 1, and even when all IVs were invalid, it still had much lower type 1 error than the original method. On the other hand, data perturbation might lose some power as compared with the original

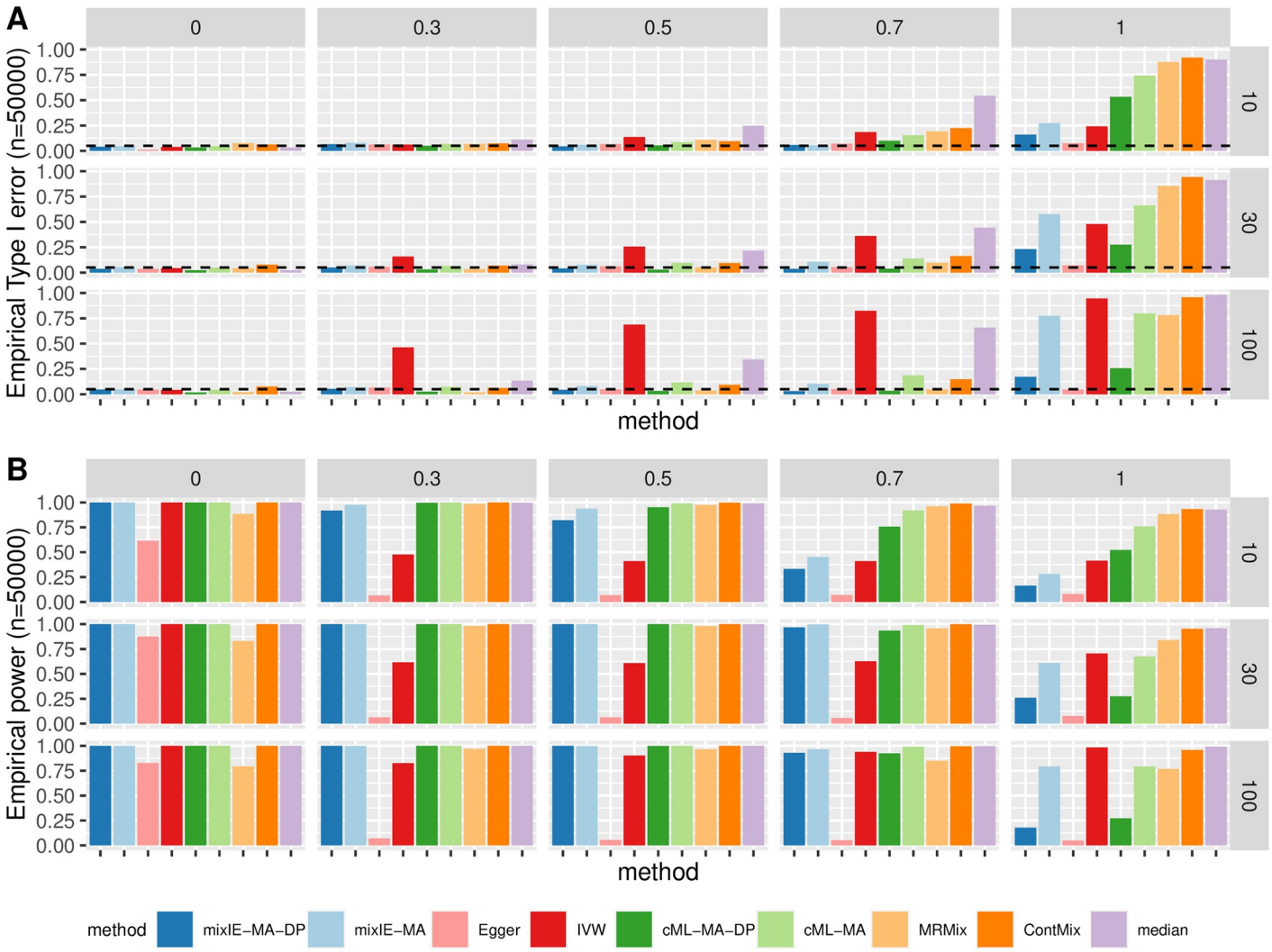


**Fig 2. Simulation results with directional pleiotropy and InSIDE satisfied.** A: Empirical type-I error; B: Power with sample size $n$ = 50 000. Each row corresponds to $m$ = 10, 30, 100 SNPs and each column corresponds to 0, 30%, 50%, 70%, 100% invalid IVs.

https://doi.org/10.1371/journal.pgen.1009922.g002

method, but the loss was acceptable especially when we had a large sample size as shown here. We can see that even when 70% IVs were invalid, there was not much power loss.

Fig 3 shows the distributions of the causal parameter estimates by each method when 70% IVs were invalid with directional pleiotropy under InSIDE. In general, cML-MA and cML-MA-DP had the smallest MSE as shown here and in S1 Text. The performance of our proposed methods was less stable with a small number of IVs (i.e. $m$ = 10). However, as $m$ increased, mixIE-MA-DP had comparable MSEs with that of cML-MA-DP. In addition, as shown in S1 Text, mixIE-MA-DP could have slightly higher power than cML-MA-DP when the sample sizes or effect sizes were small. MR-ContMix also yielded (almost) unbiased estimates with small variances as mixIE and cML methods, while MRMix was slightly biased towards 0 as shown in Fig 3B. And again, Egger regression and IVW gave a very wide range of estimates, and the IVW estimates were biased in the presence of directional pleiotropy.

Fig 4 shows the estimated proportions of invalid IVs by mixIE-MA in different scenarios. We can see that it was able to estimate the proportion of invalid IVs reasonably well, but

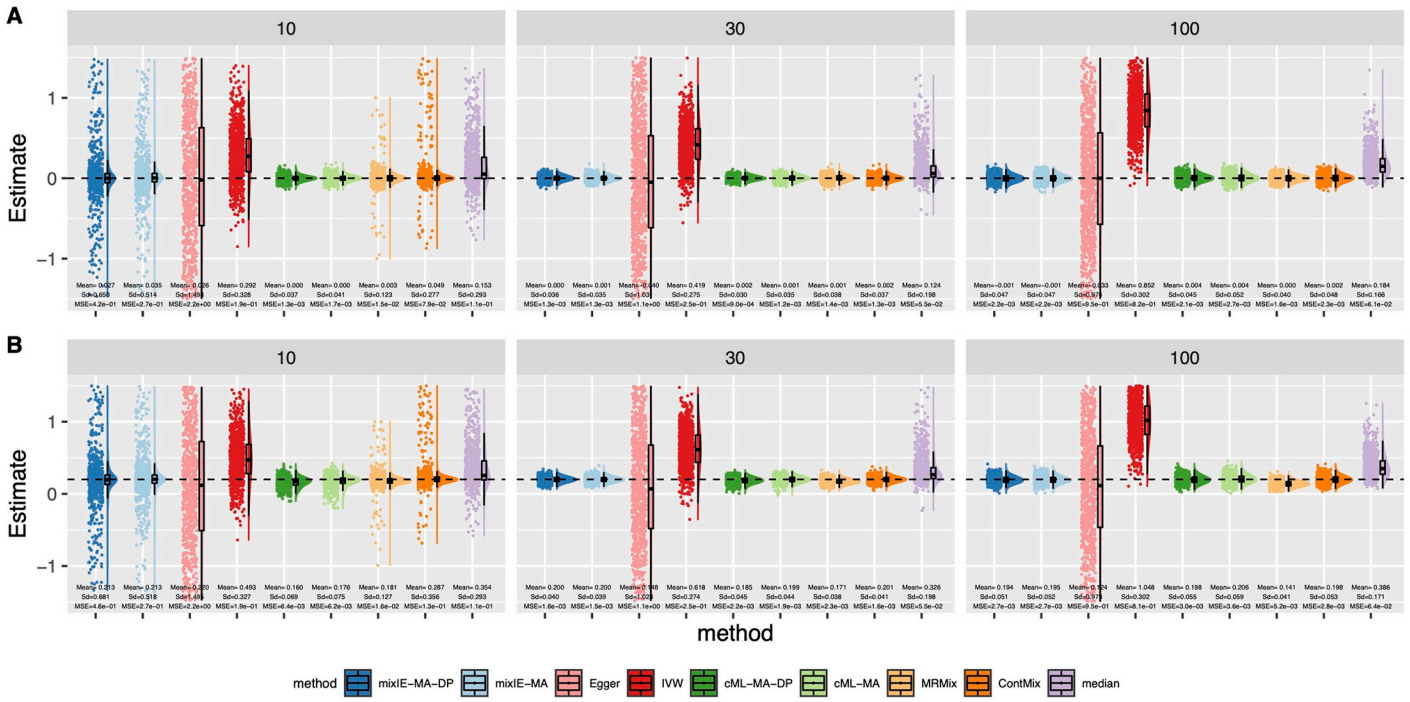


**Fig 3. Simulation results with directional pleiotropy and InSIDE satisfied.** Empirical distributions of the estimates of the causal effect $\theta$ by the methods with $n = 50000$ and 70% invalid IVs. A: $\theta = 0$. B: $\theta = 0.2$.

https://doi.org/10.1371/journal.pgen.1009922.g003

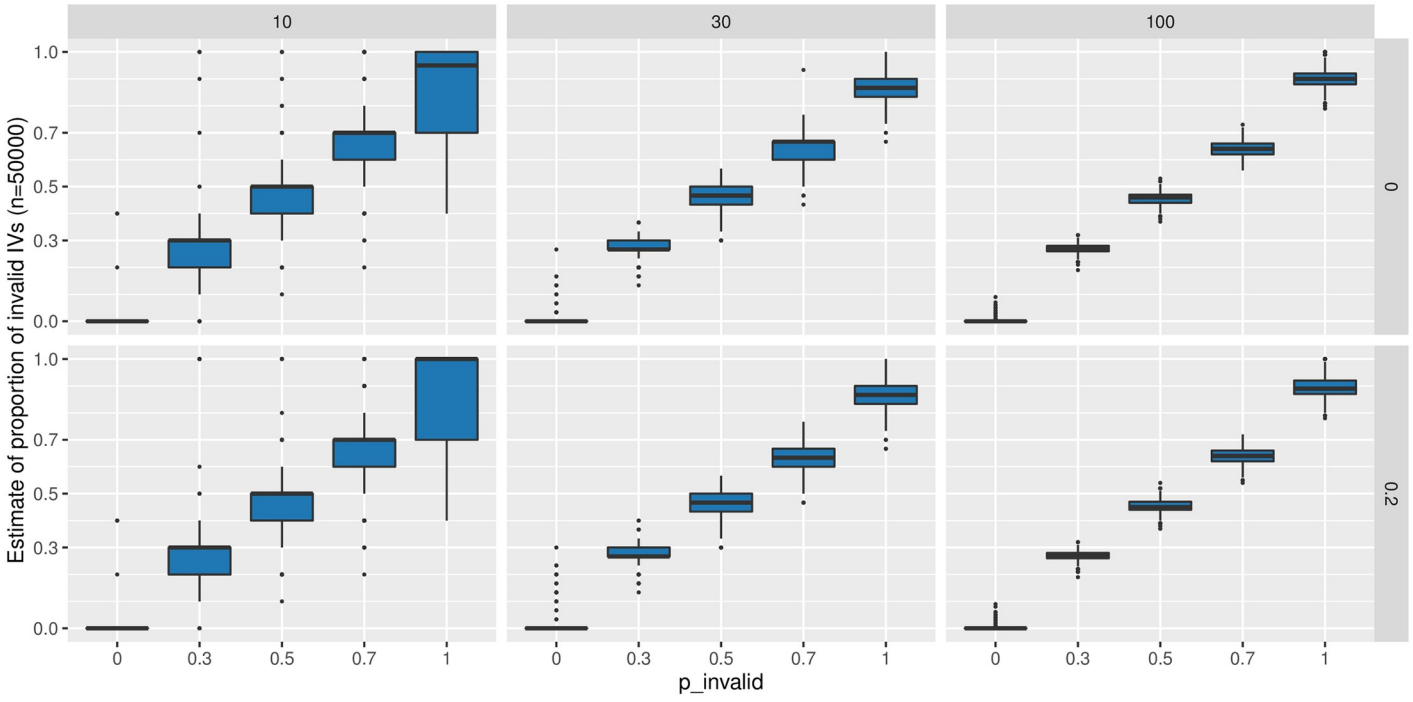


**Fig 4. Estimates of the proportion of invalid IVs by mixIE-MA with $n = 50\,000$ under directional pleiotropy and InSIDE satisfied.** The upper row corresponds to $\theta = 0$ and the lower one to $\theta = 0.2$; each column corresponds to $m = 10, 30, 100$ respectively.

https://doi.org/10.1371/journal.pgen.1009922.g004

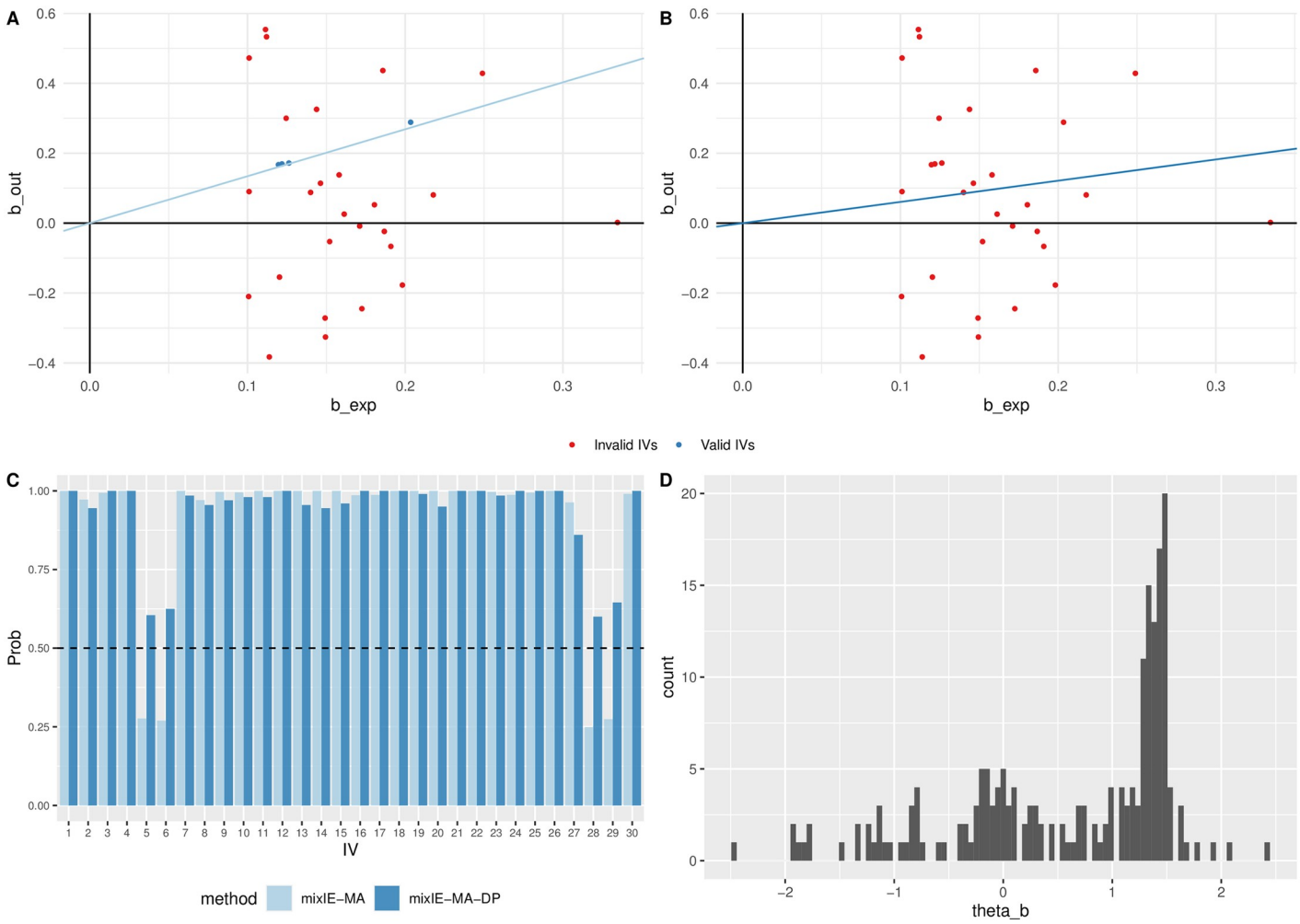


**Fig 5. A simulated data example with $n = 50\,000$, $m = 30$, $\theta = 0$, p_invalid = 1 under directional pleiotropy and InSIDE satisfied.** A: Causal estimate and identified invalid IVs by mixIE-MA. B: Causal estimate and identified invalid IVs by mixIE-MA-DP. C: Posterior probability of each IV being invalid. D: Histogram of $\hat{\theta}^{(b)*}$ from 200 perturbations.

https://doi.org/10.1371/journal.pgen.1009922.g005

under-estimated it when the proportion of invalid IVs was high. We argue that this was probably due to the weak identifiability of the mixture model: in such a scenario, the few points close to any line going through the origin point could be reasonably regarded as valid IVs. As shown in Fig 5A, although mixIE-MA identified most of the IVs to be invalid, the 4 (blue) points were identified to be valid IVs, driving a causal estimate of 1.34 with p-value $< 0.05$ and resulting in a type 1 error. While in Fig 5B, mixIE-MA-DP classified all IVs as invalid and gave a point estimate of 0.61 with p-value 0.54. In Fig 5C, using data perturbation, mixIE-MA-DP was able to identify more than 100 times out of 200 perturbations that those 4 IVs were invalid. As shown in the histogram of the causal estimates $\hat{\theta}^{(b)*}$ from 200 data perturbations in Fig 5D, there is a peak around the original estimate 1.34, a small peak around the true value 0, and some negative estimates. Again, from this example we can see that, first, data perturbation was able to account for some model selection uncertainties; second, when mixIE-MA estimated a high proportion of invalid IVs, the result might be unreliable.

**InSIDE violated.** Fig 6 shows the empirical type 1 error and power of different methods in the presence of directional pleiotropy and with the violation of the InSIDE assumption to

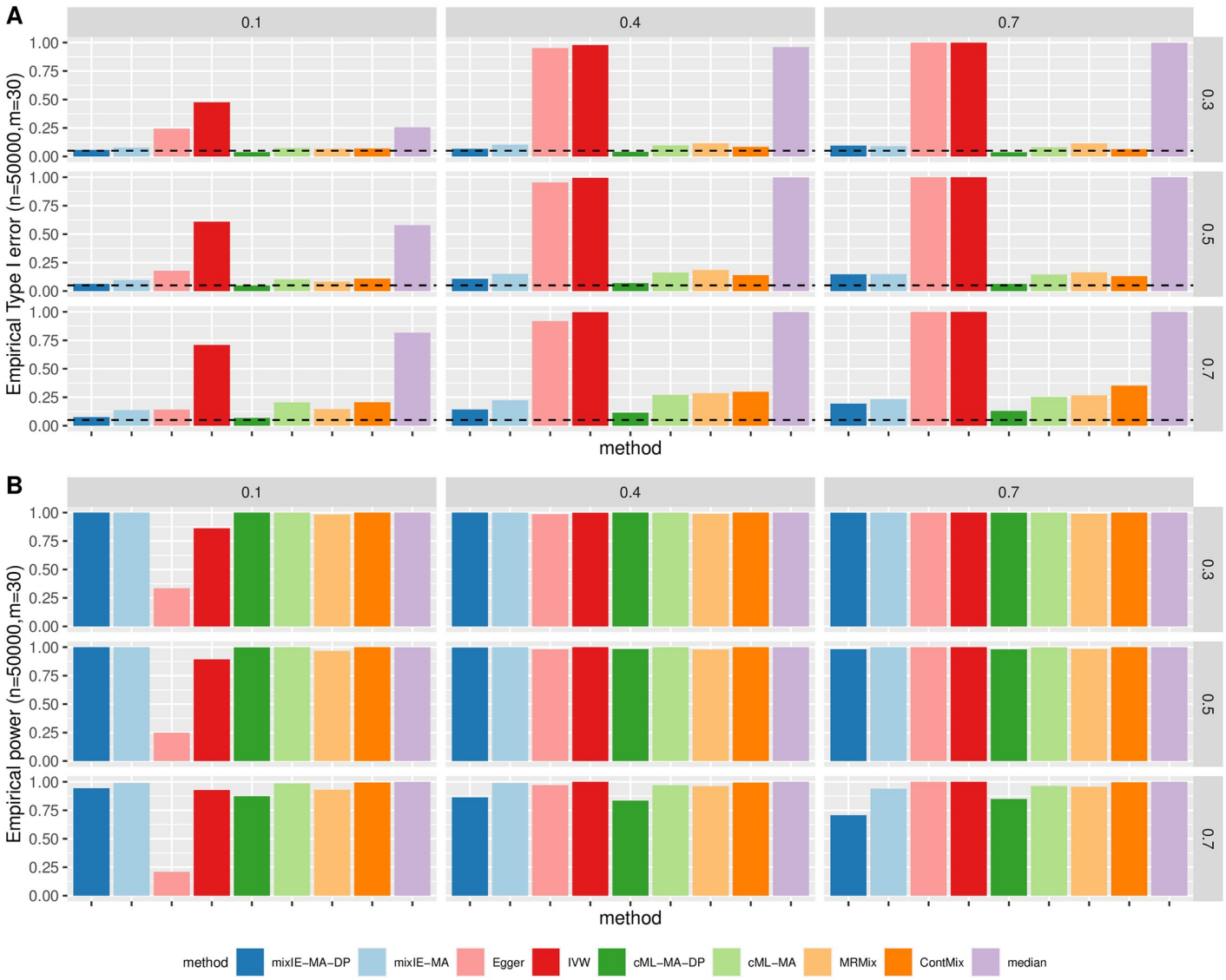


**Fig 6. Simulation results with directional pleiotropy and InSIDE violated.** A: Empirical type-I error; B: Power with sample size $n$ = 50 000 and $m$ = 30. Each column corresponds to $b$ = 0.1, 0.4, 0.7 and each row corresponds to 30%, 50%, 70% invalid IVs.

https://doi.org/10.1371/journal.pgen.1009922.g006

different degrees for $m$ = 30 IVs. Other results are shown in Section B.1.3 in S1 Text. First, when the correlated pleiotropy ($\phi_{ug}$) was relatively small compared to the directional pleiotropy (e.g. when $b$ = 0.1), our proposed method mixIE-MA-DP was still able to control the type 1 error and achieve high power. Also as shown in Fig 7, mixIE gave unbiased estimates when $b$ = 0.1. In contrast, Egger regression yielded inflated type 1 error and highly biased estimates, so did IVW and weighted-median. Other methods such as MRMix and MR-ContMix also had inflated type 1 error when the proportion of invalid IVs was high. However, as the degree of InSIDE assumption violation increased (i.e., the effect size of correlated pleiotropy increased), as expected, the performance of our proposed method went down as with larger inflated type 1 error and more biased estimates. On the other hand, it still performed more robustly than Egger regression, which almost completely broke down. We also point out that, unlike when the InSIDE assumption held as shown before, the data perturbation version mixIE-MA-DP

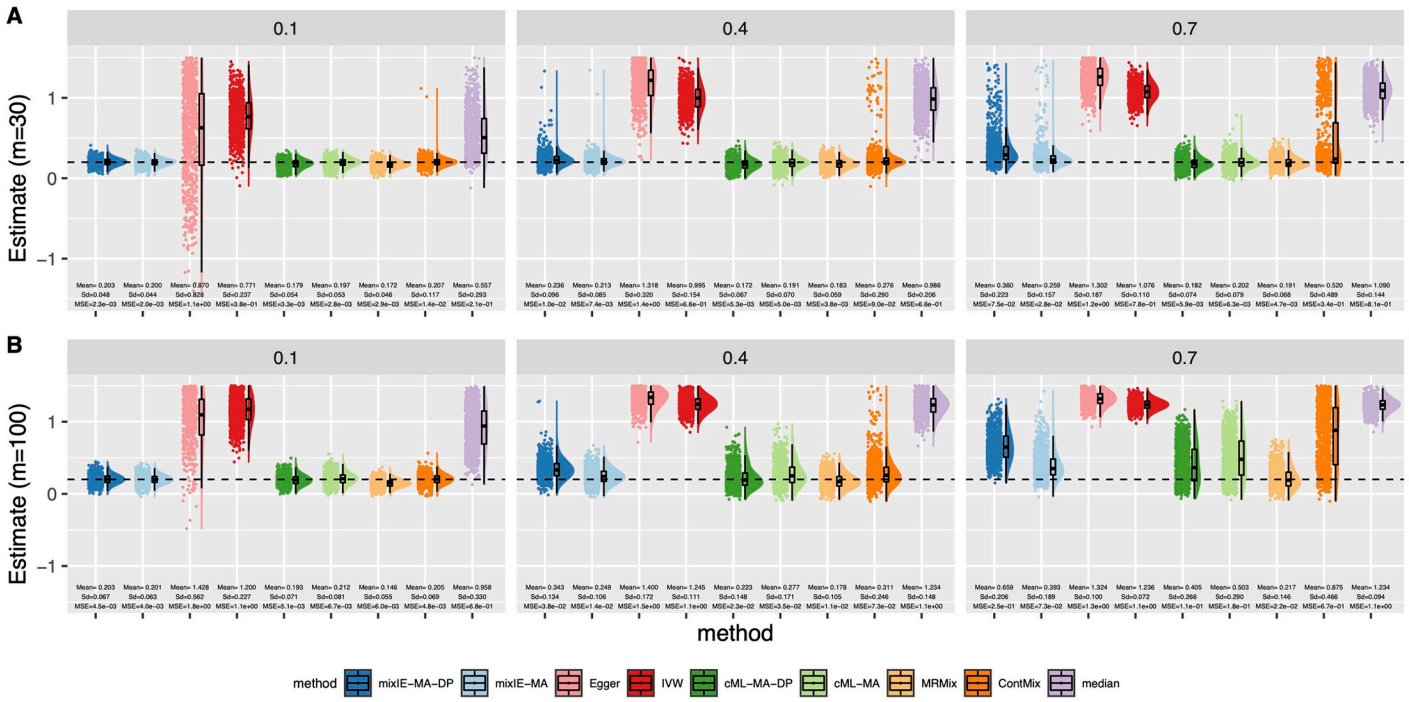


**Fig 7. Simulation results with directional pleiotropy and InSIDE violated.** Empirical distributions of the estimates of the causal effect $\theta$ by the methods with $n$ = 50 000, $\theta$ = 0.2 and 70% invalid IVs. A: $m$ = 30. B: $m$ = 100. Each column corresponds to $b$ = 0.1, 0.4, 0.7.

https://doi.org/10.1371/journal.pgen.1009922.g007

performed worse than mixIE-MA when the effect size of correlated pleiotropy was large. This is probably due to the fact that the proposed data perturbation scheme depends on the InSIDE assumption. In terms of estimation and inference, cML-MA-DP performed most robustly among all methods when the proportion of invalid IVs was small (e.g. 30%) regardless of the degrees of the violation of the InSIDE assumption—it could control the type 1 error well while yielding unbiased estimates. In terms of estimation, MRMix performed robustly as well—it could give (almost) unbiased estimates in most of the scenarios.

We conclude that when the InSIDE assumption is violated moderately and/or the proportion of invalid IVs is small, our proposed methods still have an edge over most of the popular MR methods, especially compared with Egger regression and IVW. As the degree of the violation increases, it still performs more robustly than Egger regression, IVW and weighted-median.

In summary, with the motivation of balancing and combining IVW and Egger regression, in most of the scenarios, our proposed methods were able to boost statistical power dramatically over Egger regression while controlling the type 1 error satisfactorily. The proposed methods were also able to control the type 1 error much better than IVW (except when all IVs were invalid with balanced pleiotropy and InSIDE satisfied) and improved the power in many scenarios. In general, as we mentioned before, under the InSIDE assumption, when mixIE-MA gives a high estimated proportion of invalid IVs (e.g. $>$ 70%), we suggest that mixIE-MA-DP could give more reliable results and we could even go with Egger regression when (almost) all IVs are invalid, which might be conservative; on the other hand, when mixIE-MA gives a small estimated proportion of invalid IVs (e.g. $<$ 20%), the original method and its data perturbation version mixIE-MA-DP are expected to give good and similar results.

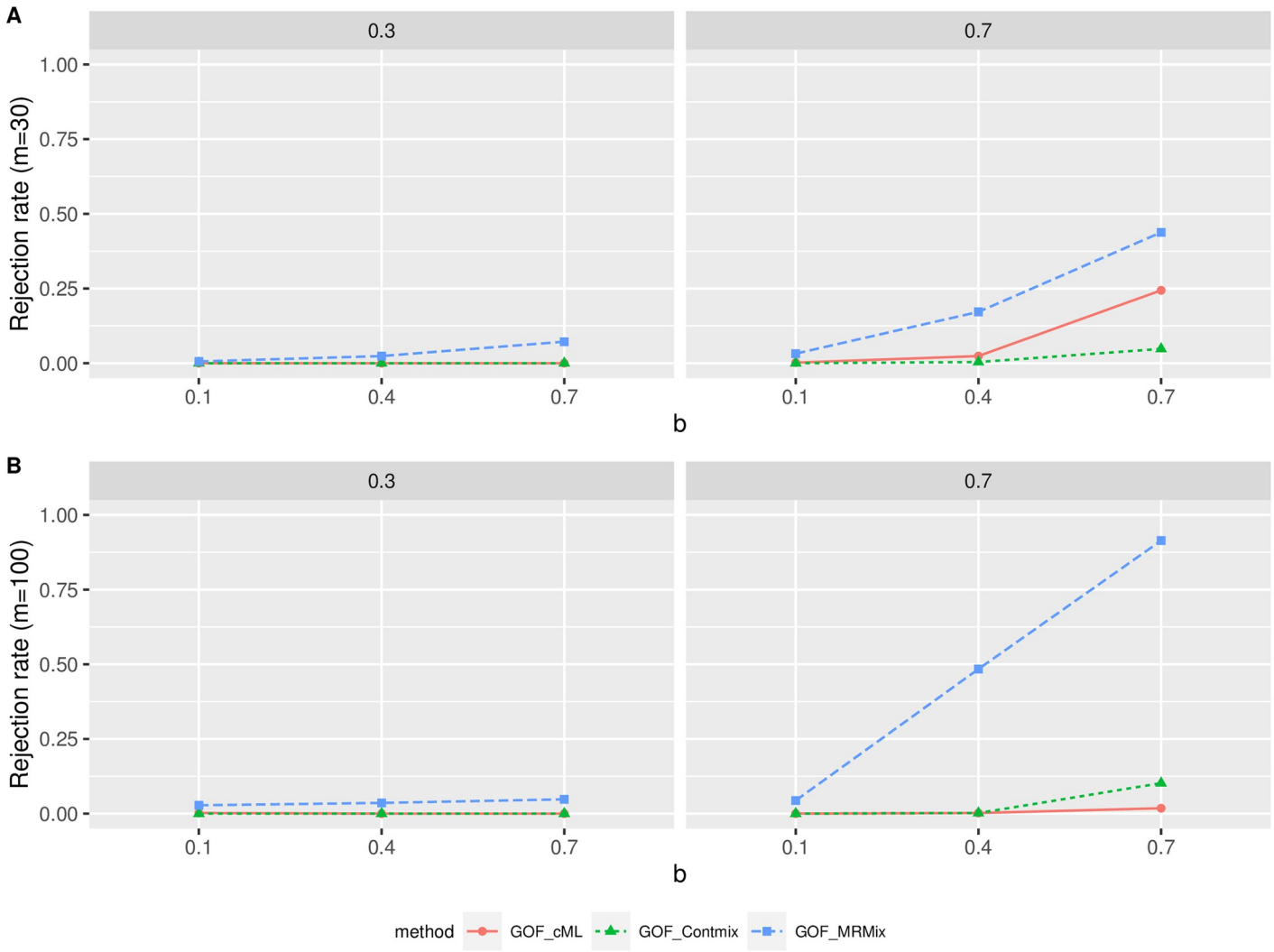


**Fig 8. Simulation results for GOF testing with directional pleiotropy and InSIDE violated (while the plurality assumption for other three methods holding).** $\theta$ = 0.2 and $n$ = 50 000. The y-axis gives the rejection rate that the results from two methods were consistent, while the x-axis gives the increasing degree of InSIDE being violated. A: $m$ = 30. B: $m$ = 100. The two columns correspond to 30% and 70% invalid IVs respectively.

https://doi.org/10.1371/journal.pgen.1009922.g008

**Model checking.** We applied the proposed GOF testing procedure under the main simulation scenario (c) with directional pleiotropy and InSIDE violated, under which the plurality of valid IV assumption held. We used $n$ = 50 000 and $\theta$ = 0.2, and varying proportions of invalid IVs and degrees of violation of the InSIDE assumption. We compared mixIE-MA with three other methods, cML-MA, MRMix and MR-ContMix, and the corresponding tests are referred as GOF-cML, GOF-MRMix and GOF-ContMix respectively. Fig 8 shows the rejection rates of the three tests for 30 or 100 IVs. First, when the degree of violation of InSIDE assumption was small ($b$ = 0.1) and/or the proportion of invalid IVs was small (30%), mixIE and cML, MR-ContMix had good consistency, while there was some inconsistency between mixIE and MRMix. This was perhaps because the estimates of MRMix were biased towards the null. Second, when the proportion of invalid IVs was high (70%), the rejection rate of GOF-MRMix increased as the degree of the InSIDE violation (i.e. $b$) increased, since the estimates of mixIE became more biased while MRMix still yielded stable estimates as shown in Fig 7. Meanwhile,

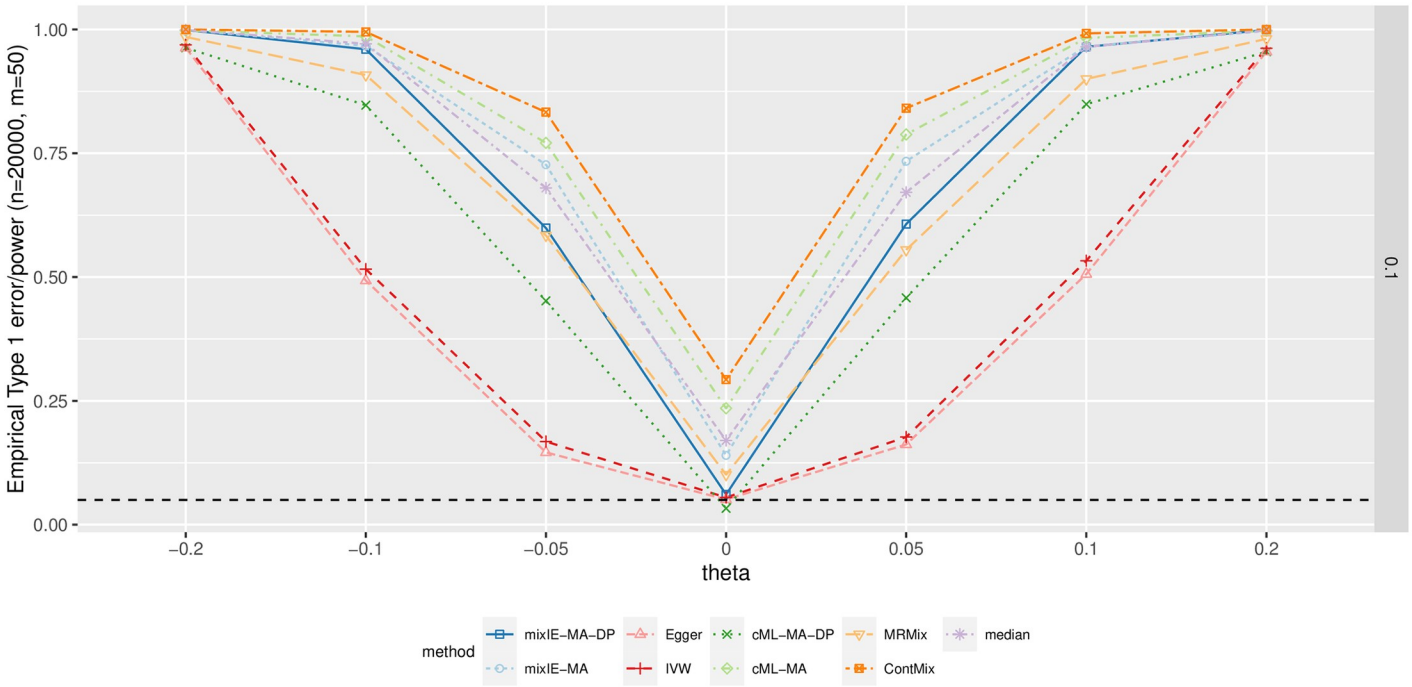


**Fig 9. Simulation results with many invalid IVs having weak pleiotropic effects.** Empirical type-I error (for $\theta = 0$) and power (for $\theta \neq 0$) curves with sample size $n = 20000$ and $h_y = 0.1$.

https://doi.org/10.1371/journal.pgen.1009922.g009

GOF-cML and GOF-ContMix had low rejection rates when $m = 100$ because the other two methods both gave biased estimates as mixIE with the same bias direction and to a similar degree (Fig 7B). In short, the proposed GOF testing was able to capture inconsistency among different methods, which could be due to the violation of the InSIDE assumption if other assumptions of mixIE and the other method held. It also depended on the performance of the other method being compared with mixIE.

**Simulations with weak invalid IVs.** Fig 9 shows the empirical type 1 error and power when $h_y = 0.1$ for different methods; the complete results are given in Section B.2 in S1 Text. In agreement with [14], cML-MA-DP, Egger regression, IVW and our proposed method mixIE-MA-DP could control the type 1 error across different scenarios, while MR-ContMix gave the highest type 1 error. mixIE-MA-DP was much more powerful than both Egger regression and IVW. It was also slightly more powerful than cML-MA-DP, especially with weaker direct effects of invalid IVs when $h_y = 0.1$, in which case it was more challenging for cML-MA-DP to identify invalid IVs. Also, as shown in Supplementary, mixIE-MA-DP was unbiased while cML-MA-DP had slight under-estimation biases towards the null. We also note that the power for Egger regression shown here was larger than the one shown in [14] because we did not re-orient SNPs for Egger regression for the purpose of fair comparison.

**Computational time.** We did simulations to compare the computational time for each method. The details are given in Section E in S1 Text. In summary, our proposed methods run reasonably fast. The computational time of mixIE-MA is comparable to that of cML-MA and faster than weighted-median and MRMix, but slower than IVW, Egger regression and MR-ContMix. As expected, with data perturbation it takes longer to run mixIE-MA-DP, though it is still quite feasible: with 200 perturbations and 50 starting points within each perturbation, it only took less than a minute with 10 to 100 IVs on a Macbook laptop.

## Primary real data example

**Main analysis.** We compared our proposed methods with other methods to identify causal effects of 12 risk factors for cardiometabolic diseases on coronary artery disease (CAD), stroke and type 2 diabetes (T2D), as well as asthma, which largely served as a negative control. The sample sizes for the traits are summarized in Table 1. Following [14], we classified the 48 pairs into 4 categories: 19 pairs considered causal or likely causal as supported by the literature, 17 pairs correlated but unknown to be causal or with conflicting evidence, 10 pairs unrelated, and 2 pairs considered non-causal.

In the main analysis, we re-oriented SNPs such that all IVs were positively associated with the exposure as recommended for Egger regression [15]; the results when we did not re-orient the SNPs are shown in Section C.2.3 in S1 Text as a sensitivity analysis. In Fig 10 we compare mixIE-MA-DP and mixIE-MA with Egger regression, IVW, cML-MA-DP and cML-MA. Table 2 compares the total numbers of the pairs identified to be significant by different methods at the significance level 0.001 (with a Bonferroni adjustment $0.05/48 \approx 0.001$). First, mixIE-MA and mixIE-MA-DP identified more causal risk factor-disease pairs than Egger regression and IVW. Egger regression only identified 6 known or likely causal pairs, while our proposed methods identified 13 pairs. As an example, for the causal BF-T2D pair, both mixIE-MA and mixIE-MA-DP gave some significant results while both Egger regression and IVW yielded only marginally significant ones. Both mixIE-MA and cML-MA identified the same single invalid IV and gave similar estimates. In contrast, IVW would be affected by the invalid IV and thus gave a smaller estimate with an only marginally significant p-value. See Section C.1.1 in S1 Text for more details.

Second, as shown in Table AG in S1 Text, among the 48 risk factor-disease pairs, most of the pairs had less than 20% IVs identified to be invalid by mixIE-MA. In agreement with the simulations, mixIE-MA-DP and mixIE-MA gave similar results in general when the (estimated) proportion of invalid IVs was small. However, they might give different results for some of the pairs with a high proportion of invalid IVs. For example, for (causal) FG-T2D

**Table 1. Genome wide association studies for 4 common diseases and 12 risk factors.**

| Abbreviation | Trait | Sample size | Number of variants used in MR[1] | Reference |
|---|---|---|---|---|
| TG | triglycerides | 188577 | 122–128 | [28] |
| LDL | low-density lipoproteins | 188577 | 173–184 | [28] |
| HDL | high-density lipoproteins | 188577 | 188–197 | [28] |
| Height | body height | 253288 | 977–986 | [29] |
| BMI | body mass index | 322154 | 88–90 | [30] |
| BF | body fat percentage | 100716 | 9–10 | [31] |
| BW | birth weight | 153781 | 54–65 | [32] |
| DBP | diastolic blood pressure | 757601 | 1108–1345 | [33] |
| SBP | systolic blood pressure | 757601 | 1106–1324 | [33] |
| FG | fasting glucose | 46186 | 17 | [34] |
| Smoke | ever regular smoker | 1232091 | 114–129 | [35] |
| Alcohol | drinks per week | 941280 | 44–54 | [35] |
| CAD | coronary artery disease | 547261 | | [36] |
| Stroke | any stroke | 446696 | | [37] |
| T2D | type 2 diabetes | 69033 | | [38] |
| Asthma | asthma | 142486 | | [39] |

[1] This is the range of the number of variants used in the analysis for the risk factor (exposure), which would be slightly different based on the disease (outcome). Specific numbers are given in Table AG in S1 Text.

https://doi.org/10.1371/journal.pgen.1009922.t001

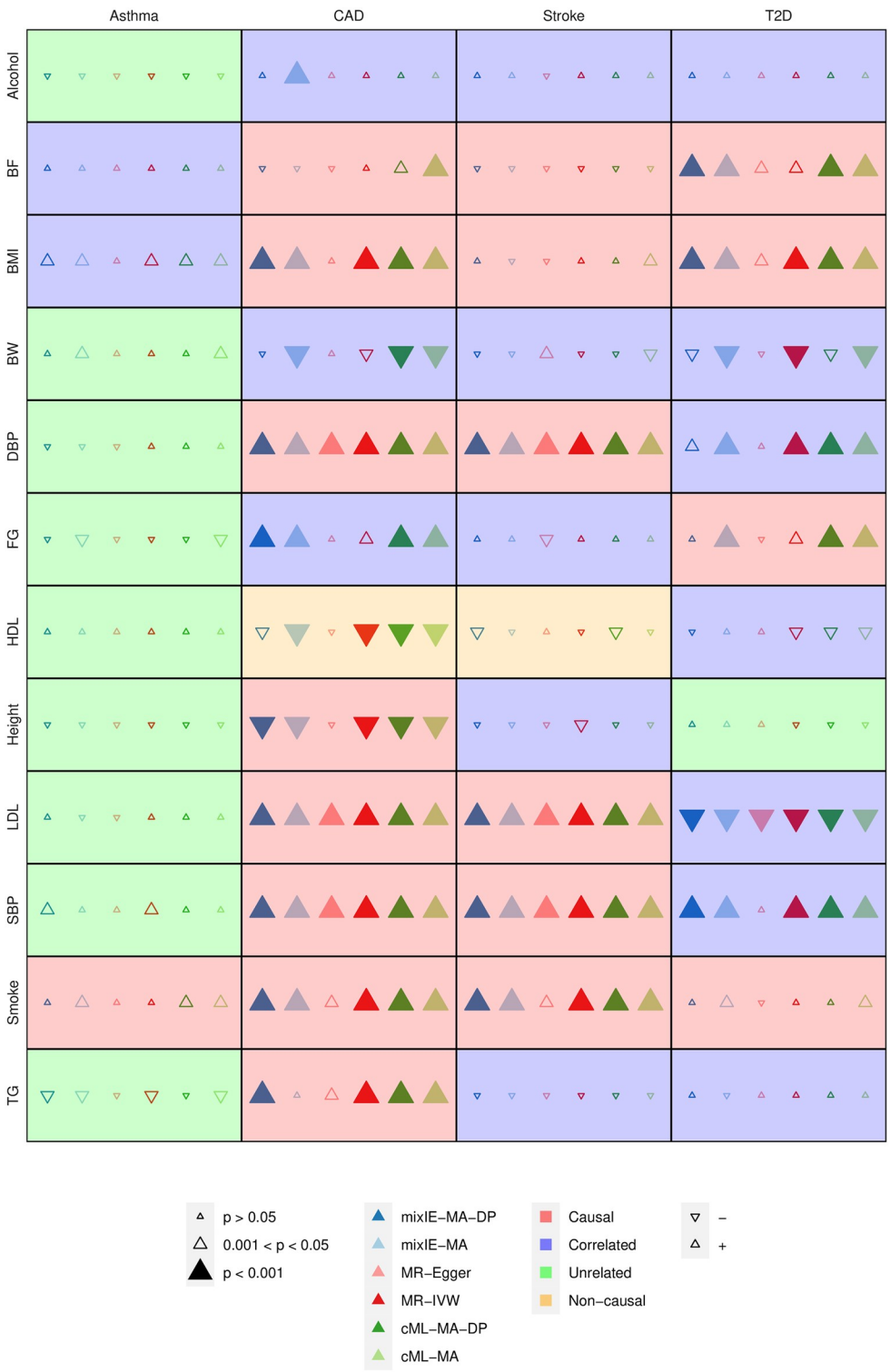


**Fig 10. Results of various methods to detect causal relationships among 48 risk factor-disease pairs.**

https://doi.org/10.1371/journal.pgen.1009922.g010

**Table 2. Numbers of significant pairs among 48 risk factor-disease pairs at the significance cutoff of p-value < 0.001.**

| | Causal | Correlated | Unrelated | Non-causal |
|---|---|---|---|---|
| mixIE-MA | 13 | 7 | 0 | 1 |
| mixIE-MA-DP | 13 | 3 | 0 | 0 |
| Egger | 6 | 1 | 0 | 0 |
| IVW | 12 | 4 | 0 | 1 |
| cML-MA | 15 | 6 | 0 | 1 |
| cML-MA-DP | 14 | 5 | 0 | 1 |
| MR-ContMix | 11 | 5 | 0 | 1 |
| Weighted-Median | 12 | 2 | 0 | 1 |

https://doi.org/10.1371/journal.pgen.1009922.t002

pair, mixIE-MA identified about 40% of 17 SNPs to be invalid IVs and a significant causal effect (p-value < 0.001), while mixIE-MA-DP did not give a significant result. Another example is for the (causal) TG-CAD pair, mixIE-MA identified about 50% of 128 SNPs to be invalid IVs and had a p-value > 0.05, but mixIE-MA-DP gave a p-value < 0.001. See Section C.1.2 in S1 Text for more details.

Third, it is notable that mixIE-MA-DP was one of the only two methods that did not give a false positive for the HDL-CAD pair. The other one was Egger regression, which however is in general low-powered.

For model checking, due to the heavy computation burden of GOF-MRMix, we only applied GOF-cML and GOF-ContMix to compare our proposed mixIE with cML and MR-ContMix. It turned out that, at the 95% confidence level, there was only one pair, SBP-CAD, for which the (causal) estimate of mixIE was inconsistent with those from cML and MR-ContMix, though the three estimates were 0.031, 0.037 and 0.039 with only small differences, and all three were statistically significant (with 1324 SNPs/IVs). Nevertheless, cautions should be taken in interpreting the results for this pair.

Some sensitivity analysis results are included in Section C.2 in S1 Text.

### Secondary real data example

Now we consider an example of 63 trait pairs that are not genetically correlated and thus are unlikely to be causally related. It can serve as a negative control to examine whether an MR method can control type I error satisfactorily. The main challenge with this example arises due to that some invalid IVs with only weak pleiotropic/direct effects are difficult to identify. The trait pairs include 13 traits: fasting proinsulin (FP), height, homeostasis model assessment of beta-cell function (HOMA), LDL, rheumatoid arthritis (RA), schizophrenia (SCZ), T2D, age at smoking, anorexia nervosa, childhood IQ, ever/never smoked, former current smoker, and infant head circumference. We applied the mixIE methods to these 63 genetically uncorrelated trait pairs. Following [14], we show the Q-Q plots of our proposed methods for the 53 pairs, which excluded 10 pairs with only 2 IVs. As shown in Fig 11, consistent with our simulation results, while mixIE-MA seemed to have inflated type I error, its data perturbation version mixIE-MA-DP appeared to be able to control the type I error well. The detailed results are given in Table AI in S1 Text. In contrast, as shown in the Supplementary of [14], MRMix and MR-ContMix yielded inflated type I errors too.

## Discussion

We have proposed mixIE, a new method to combine two of the most popular MR approaches, namely IVW and Egger regression, aiming to maintain each one's strengths while overcoming

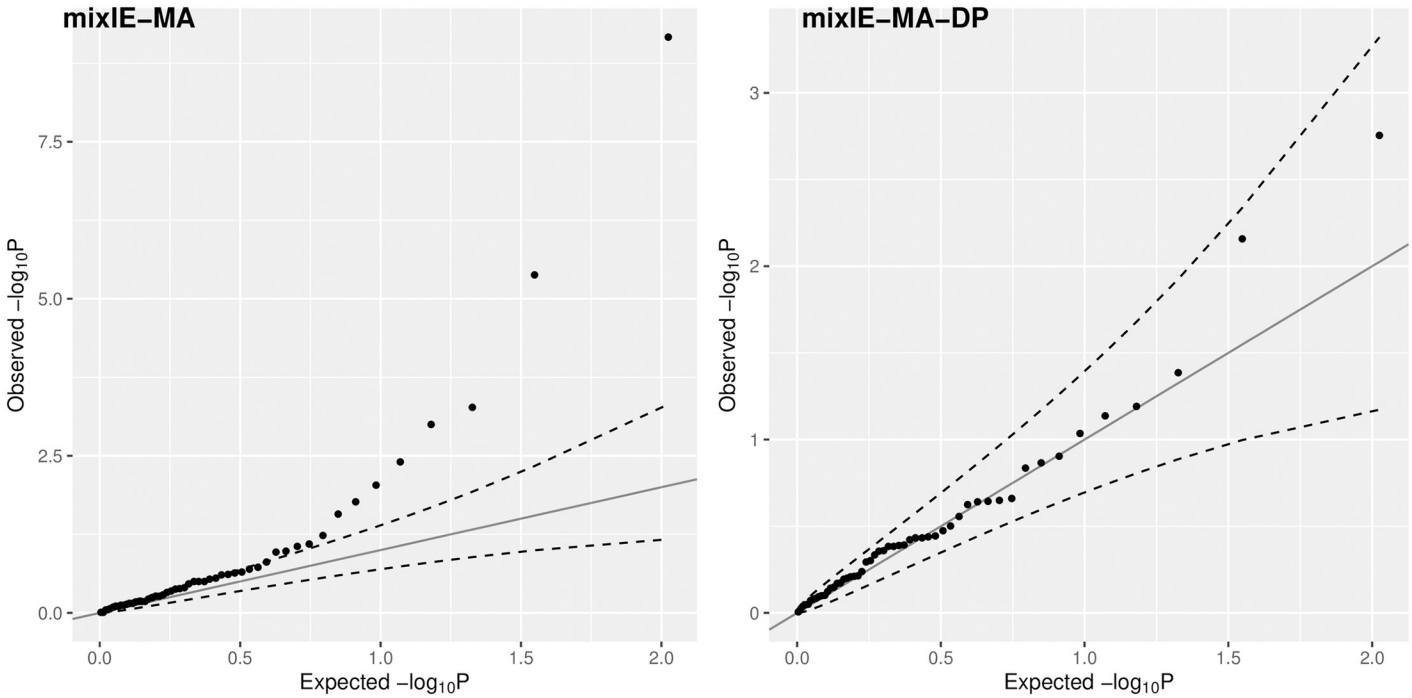


**Fig 11. Q-Q plots for 53 (likely) null trait-pairs in the secondary real data examples.** Left panel: mixIE-MA; right panel: mixIE-MA-DP.

https://doi.org/10.1371/journal.pgen.1009922.g011

their main limitations. We have found that model averaging performed better than the usual asymptotics-based statistical inference for finite samples, and thus proposed a model averaging approach as the default to implement mixIE, denoted mixIE-MA. We have also proposed a data perturbation-based version, mixIE-MA-DP, which can be more robust in accounting for model selection uncertainties, especially in more challenging situations (e.g. with many invalid IVs of small effects) for model selection with small sample sizes. We note that our proposed data perturbation scheme is novel in that it is applicable to GWAS summary data while dealing with the presence of both valid and invalid IVs at the same time. As shown in simulations and real data analyses, mixIE improved the power of Egger regression while controlling type I error rates well in most of the scenarios. It can handle directional pleiotropic effects by identifying invalid IVs while IVW cannot (with severely inflated type I error rates). Even in cases with balanced pleiotropy, mixIE could be often much more powerful than IVW (with a random-effect model). It also had some edge over its strong competitor cML in some scenarios with many weak invalid IVs. We have further demonstrated the usefulness of data perturbation through simulations and real data examples where it could better identify invalid IVs in some challenging scenarios. We also proposed a model checking procedure to compare our method with other methods, which do not require the InSIDE assumption, by evaluating the consistency of their estimates.

There are a few limitations of our proposed mixIE. First, like Egger regression, strictly speaking, our proposed method requires the InSIDE assumption to hold, and thus can be problematic with correlated pleiotropy (when the InSIDE assumption is violated). However, as shown in our simulations, even when the InSIDE assumption did not hold, mixIE still could often control type 1 error relatively close to a nominal level with decent power unless the degree of violation is dramatic; the reason is due to its (often much larger) dependence on the IVW estimate with valid IVs (than that on the Egger regression estimate). Relatedly, it is

unclear yet how commonly correlated pleiotropy is present and what its typical effect size would be in real data. Although our proposed goodness-of-fit testing is able to detect inconsistency between the results from mixIE and another method, suggesting possible violation of the InSIDE assumption, it can be due to other reasons (i.e. other assumptions being violated). Second, inherited from Egger regression, mixIE may be sensitive to the coding/orientation of IVs. But as shown in the real data application, our proposed method was more robust than Egger regression, again due to its dependence on the IVW estimate with (detected) valid IVs. Until a better approach is available, we would follow the same guideline to SNP reorientation for Egger regression. Third, as a mixture model, mixIE may suffer from the weak identifiability issue, especially when the proportion of invalid IVs is close to 1. We alleviate this problem by a model averaging approach with both IVW and Egger regression models added in the list of the candidate models, as well as by data perturbation to better identify the set of (weak) invalid IVs. Fourth, as in typical MR, we assume and choose SNPs to be independent throughout this article. Extensions to the use of correlated SNPs may gain power in other applications, including transcriptome-wide association studies [40–44]. Furthermore, like most of summary-data based MR, we select and apply the instrument variables from and to the same exposure GWAS dataset, which could lead to biased inference because of selection bias [45]. Finally, more applications to real data, including comparisons with other MR methods, are warranted [3, 14, 27].

## Supporting information

**S1 Text. Supplementary file describing standard error estimation, additional simulation results, additional real data analysis results and computational time.**
(PDF)

## Acknowledgments

WP thanks Haoran Xue and Xiaotong Shen for helpful discussions.

## Author Contributions

**Conceptualization:** Wei Pan.

**Data curation:** Zhaotong Lin.

**Formal analysis:** Zhaotong Lin, Yangqing Deng, Wei Pan.

**Funding acquisition:** Wei Pan.

**Investigation:** Zhaotong Lin, Yangqing Deng, Wei Pan.

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

**Project administration:** Wei Pan.

**Resources:** Wei Pan.

**Software:** Zhaotong Lin, Yangqing Deng.

**Supervision:** Wei Pan.

**Validation:** Zhaotong Lin, Yangqing Deng.

**Visualization:** Zhaotong Lin.

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

**Writing – review & editing:** 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.
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. 2020; p. 1–11. https://doi.org/10.1007/s40484-020-0216-3 PMID: 34386270
4. 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
5. 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
6. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic epidemiology. 2016; 40 (4):304–314. https://doi.org/10.1002/gepi.21965 PMID: 27061298
7. Verbanck M, Chen Cy, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature genetics. 2018; 50(5):693–698. https://doi.org/10.1038/s41588-018-0099-7 PMID: 29686387
8. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. International journal of epidemiology. 2017; 46(6):1985–1998. https://doi.org/10.1093/ije/dyx102 PMID: 29040600
9. Zhu X, Li X, Xu R, Wang T. An iterative approach to detect pleiotropy and perform Mendelian Randomization analysis using GWAS summary statistics. Bioinformatics. 2021; 37(10):1390–1400. https://doi.org/10.1093/bioinformatics/btaa985 PMID: 33226062
10. Celeux G, Govaert G. A classification EM algorithm for clustering and two stochastic versions. Computational statistics & Data analysis. 1992; 14(3):315–332. https://doi.org/10.1016/0167-9473(92)90042-E
11. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977; 39(1):1–22.
12. Burgess S, Foley CN, Allara E, Staley JR, Howson JM. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nature communications. 2020; 11(1):1–11. https://doi.org/10.1038/s41467-019-14156-4 PMID: 31953392
13. 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
14. 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
15. 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
16. Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1982; 44(2):226–233.
17. Schwarz G, et al. Estimating the dimension of a model. Annals of statistics. 1978; 6(2):461–464. https://doi.org/10.1214/aos/1176344136
18. Buckland ST, Burnham KP, Augustin NH. Model Selection: An Integral Part of Inference. Biometrics. 1997; 53(2):603–618. https://doi.org/10.2307/2533961
19. Shen X, Ye J. Adaptive model selection. Journal of the American Statistical Association. 2002; 97 (457):210–221. https://doi.org/10.1198/016214502753479356
20. Breiman L. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. Journal of the American Statistical Association. 1992; 87(419):738–754. https://doi.org/10.1080/01621459.1992.10475276
21. Shao J, Tu D. The jackknife and bootstrap. Springer Science & Business Media; 2012.
22. 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

23. Slob EA, Groenen PJ, Thurik AR, Rietveld CA. A note on the use of Egger regression in Mendelian randomization studies. International journal of epidemiology. 2017; 46(6):2094–2097. https://doi.org/10.1093/ije/dyx191 PMID: 29025040

24. 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

25. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. International journal of epidemiology. 2016; 45(6):1866–1886. https://doi.org/10.1093/ije/dyw314 PMID: 28108528

26. 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

27. Slob EA, Burgess S. A comparison of robust Mendelian randomization methods using summary data. Genetic epidemiology. 2020; 44(4):313–329. https://doi.org/10.1002/gepi.22295 PMID: 32249995

28. 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

29. 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

30. 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

31. 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

32. 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

33. 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

34. 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

35. 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

36. 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

37. 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

38. 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

39. Demenais F, Margaritte-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

40. Liu L, Zeng P, Xue F, Yuan Z, Zhou X. Multi-trait transcriptome-wide association studies with probabilistic Mendelian randomization. The American Journal of Human Genetics. 2021; 108(2):240–256. https://doi.org/10.1016/j.ajhg.2020.12.006 PMID: 33434493

41. 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

42. Knutson Katherine A, Pan W. Integrating brain imaging endophenotypes with GWAS for Alzheimer's disease. Quantitative Biology. 2021; 9(2):185.

43. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nature communications. 2018; 9(1):1–12. https://doi.org/10.1038/s41467-017-02317-2 PMID: 29335400

44. Gleason KJ, Yang F, Chen LS. A robust two-sample transcriptome-wide Mendelian randomization method integrating GWAS with multi-tissue eQTL summary statistics. Genetic Epidemiology. 2021; 45 (4):353–371. https://doi.org/10.1002/gepi.22380 PMID: 33834509

45. Wang K, Han S. Effect of selection bias on two sample summary data based Mendelian randomization. Scientific reports. 2021; 11(1):1–8. https://doi.org/10.1038/s41598-021-87219-6 PMID: 33828182