A Comprehensive Evaluation of Methods for Mendelian Randomization Using Realistic Simulations of Genome-wide Association Studies

Guanghao Qi[1] and Nilanjan Chatterjee[1,2]*

[1]Department of Biostatistics, Bloomberg School of Public Health and [2]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

*** Correspondence to Nilanjan Chatterjee (nilanjan@jhu.edu)**

Abstract

Mendelian randomization (MR) has provided major opportunities for understanding the causal relationship among complex traits. Previous studies have often evaluated MR methods based on simulations that do not adequately reflect the data-generating mechanism in GWAS and thus there is often discrepancies in performance of MR methods in simulation studies and in real datasets. We use a simulation framework that generates data on full GWAS for two traits under realistic model for effect-size distribution coherent with heritability, co-heritability and polygenicity typically observed for complex traits. We select instruments based on SNPs which reach genome-wide significance in the underlying study of one of the traits the causal effect of which to be tested on the other. Results show that the weighted mode method and MRMix are the only two methods which maintain correct type I error rate in a diverse set of scenarios. Between the two methods, MRMix tends to be more powerful for larger GWAS while the opposite being true for smaller sample sizes. Among other methods, random effect based IVW analysis, MR analyses based on robust loss function (MR-Robust) and robust profile-scores (MR-RAPS) tend to perform best in terms of maintaining low MSE when the InSIDE assumption is satisfied. However, when the InSIDE assumption is violated, all methods except weighted mode and MRMix can produce large bias and correspondingly large mean squared errors (MSE). In conclusion, the results show that relative performance of different methods depends heavily on sample sizes of underlying GWAS, proportion of valid instruments and validity of the InSIDE assumption.

Introduction

Epidemiological associations are often biased by unobserved confounders. Mendelian randomization (MR) – a form of instrumental variable approach that uses genetic variants as instruments – has provided major opportunities for understanding the causal relationship across complex traits [1–3]. Two-sample MR analysis [4,5] is particularly popular as these methods can be applied even if the genome-wide association study for the different traits are conducted on non-overlapping samples. In recent years, the growing sample size of GWAS and volume of publicly available summary-level datasets have facilitated the popularization of MR. MR analyses have led to improved understanding of epidemiological associations, discovery of new drug targets [3,6], and new insights into biological mechanisms through applications to large-scale omics data [7–9].

The validity of early MR methods relied on a crucial assumption that the genetic variants have no effects on the outcome that are not mediated by the exposure. This assumption can be violated in the presence of “horizontal pleiotropy”. Recent studies have found that pleiotropy is a wide-spread phenomenon [10–15], leading to concerns over the accuracy of Mendelian randomization analysis. To deal with this challenge, many methods have been proposed that take advantage of the multitude of genetic instruments to reduce the bias due to horizontal pleiotropy. Different methods deal with different kinds of pleiotropy and often rely on different assumptions. For example, the median and mode-based method take (weighted) median or mode of the ratio estimates [16,17]. The former method requires that more than 50% of the instruments are valid and the latter requires a plurality. Egger regression was proposed to specifically deal with directional pleiotropy and requires the InSIDE assumption [18]. Another type of methods is based on outlier detection and re-estimation of causal effects after removing outliers [13]. Very recently, we proposed the method MRMix which uses an underlying mixture model to implicitly distinguish valid and invalid instruments and estimate the causal effect accordingly. Most recently, a variation of mixture-model based method has been proposed, called contamination mixture [19], which models the ratio estimates using normal-mixture with pre-specified variance parameters.

Choosing a method for MR analysis can be challenging. While a number of previous studies have conducted various simulation studies to evaluate MR methods under alternative modelling assumptions, conclusions may be limited because these studies often do not incorporate realistic model for genetic architecture of complex traits as implied by recent studies of heritability/co-heritability [10,20–24] and effect-size distribution [25–28]. Further, many simulation studies also directly simulate data on the instruments ignoring the process that instruments in reality are selected to be SNPs that reach genome-wide significance in an underlying genome-wide association study - as a result of which there should be a close relationship between sample size, number of available instruments, their average effect-sizes and precision of their estimated effects. Previous MR studies have used a fixed number of IVs and a fixed sample size [4,13,19] or vary one of them without the other [16–18,29,30], and generate the effects of genetic instruments from a fixed distribution without varying with the sample size or the number of IVs. In addition, genetic effects on exposure and outcomes are often simulated using uniform distributions while clearly many studies have shown they more likely to follow a spike-and-slab type distributions [27,28,31]. The magnitude of genetic effects simulated is also unrealistic in some studies such as only 25 SNPs explaining as much as 94% of variance of the exposure [18] or selected IVs explaining larger variance of the outcome than the exposure [32]. Because of these issues, performance of MR methods, in absolute and relative terms, can be discrepant between simulation studies and real GWAS datasets.

In this paper, we use a simulation framework that closely mirror real genome-wide association studies. In particular, we simulate data on genome-wide set of SNPs and select instruments based on SNPs which reach genome-wide significance in the underlying study of the exposure. We simulate genetic effects constrained by realistic values for heritability/co-heritability and models for effect-size distribution [10,20–28]. We compare performance of a variety of existing methods under different sample sizes of underlying GWAS and correspondingly number of IVs and vary the proportion of valid instruments and mechanisms of pleiotropy. Results from these simulation studies provide comprehensive and realistic insights into strengths and limitations of existing methods.

Methods

We begin by introducing a few notations. Let X denote the exposure, Y denote the outcome and U denote a potential confounder. Let G_j denote the genotype of SNP j. Throughout most MR literature [16–18,30,32], the simulations are conducted using the following model

$$U = \sum_{j=1}^{M} \phi_j G_j + \epsilon^U \quad (1)$$

$$X = \sum_{j=1}^{M} \gamma_j G_j + \theta_{Ux} U + \epsilon^X \quad (2)$$

$$Y = \sum_{j=1}^{M} \alpha_j G_j + \theta X + \theta_{Uy} U + \epsilon^Y \quad (3)$$

Here $\phi_j, \gamma_j, \alpha_j$ denote the direct effect of SNP j on U, X and Y, respectively. We also adopt this model in our simulations. But unlike most previous studies that simulate only the selected instruments, we generate data for all common variants in the genome. We are interested in estimating the causal effect (θ) of X on Y. The effects of the confounder U on X and Y are denoted by θ_{Ux} and θ_{Uy}, respectively. The error terms ϵ^U, ϵ^X and ϵ^Y are independent and normally distributed with mean 0. For convenience, we chose the variance of the error terms so that U, X and Y all have unit variance. We generate genotypes G_j by first simulating $\tilde{G}_j$s independently from Binomial(2, 0.3) and then standardizing by $G_j = \frac{\tilde{G}_j - 2\times 0.3}{\sqrt{2\times 0.3\times(1-0.3)}}$ to make them have mean 0 and variance 1. They represent SNPs with minor allele frequency 0.3 after standardization. We generated data from model (1)-(3) using 200,000 independent SNPs as representative of all underlying common variants. We generate $\phi_j, \gamma_j, \alpha_j$ from mixture normal distributions which have been shown to be appropriate for modeling effect-size distribution for complex traits in GWAS [25–28]. Under the above model, when the confounder U has heritable component, the InSIDE assumption[18] is violated as direct and indirect effect of some SNPs on the outcome are correlated due to mediation by common factor U .

Balanced Horizontal Pleiotropy with InSIDE Assumption Satisfied

We first simulate settings where SNPs with direct effect on X can also have direct effect on Y, thus allowing horizontal pleiotropy, but we allow the InSIDE assumption to be satisfied by setting $\phi_j = 0$ for all SNPs. We generate γ_j and α_j across SNPs from the following distribution:

$$\begin{pmatrix}\gamma\\\alpha\end{pmatrix} \sim \pi_1\begin{pmatrix}N(0,\sigma_x^2)\\\delta_0\end{pmatrix} + \pi_2\begin{pmatrix}N(0,\sigma_x^2)\\N(0,\sigma_y^2)\end{pmatrix} + \pi_3\begin{pmatrix}\delta_0\\N(0,\sigma_y^2)\end{pmatrix} + \pi_4\begin{pmatrix}\delta_0\\\delta_0\end{pmatrix} \qquad (M1)$$

In the above, the first component is the set of valid IVs, i.e. the SNPs which have no horizontal pleiotropic effect on Y. The second component is the set of SNPs with horizontal pleiotropic effect on Y. However, because here we assume γ_j and α_j are independent, in this setting the InSIDE assumption is satisfied. The third component is the set of SNPs that are only associated with Y and the fourth component is the set of SNPs that have no association with either trait. Mixture proportions are denoted by $\pi_1, \pi_2, \pi_3, \pi_4$ and variance parameters are denoted by σ_x^2 and σ_y^2 (see **Table 1** and **Supplementary Table 1** for details on values used for the parameters).

Balanced Horizontal Pleiotropy with InSIDE Assumption violated

Next we allow the InSIDE assumption to be violated by allowing a fraction of SNPs to have an effect on the confounder U. Here we generate $\phi_j, \gamma_j, \alpha_j$ from the tri-variate normal mixture:

$$\begin{pmatrix}\gamma\\\phi\\\alpha\end{pmatrix} \sim \pi_1\begin{pmatrix}N(0,\sigma_x^2)\\\delta_0\\\delta_0\end{pmatrix} + \pi_2\begin{pmatrix}N(0,\tilde{\sigma}_x^2)\\N(0,\sigma_u^2)\\N(0,\tilde{\sigma}_y^2)\end{pmatrix} + \pi_3\begin{pmatrix}\delta_0\\\delta_0\\N(0,\sigma_y^2)\end{pmatrix} + \pi_4\begin{pmatrix}\delta_0\\\delta_0\\\delta_0\end{pmatrix} \qquad (M2)$$

In the above, the first component corresponds to valid instruments which have only direct effect on X. The second component allows a set of SNPs that have effects on U and thus creating horizontal pleiotropic effect with the InSIDE assumption violated. We also allow the same set of SNPs to have direct effects on X and Y, but the effect sizes are of smaller magnitude.

Directional Pleiotropy

In the above two settings, we assumed direct effects of the SNPs on the outcome Y have mean zero. Next to simulate directional pleiotropy, we generate α_j from a distribution with non-zero mean (μ_y). When the InSIDE assumption holds, we simulate from

$$\begin{pmatrix}\gamma\\\alpha\end{pmatrix} \sim \pi_1\begin{pmatrix}N(0,\sigma_x^2)\\\delta_0\end{pmatrix} + \pi_2\begin{pmatrix}N(0,\sigma_x^2)\\N(\mu_y,\sigma_y^2)\end{pmatrix} + \pi_3\begin{pmatrix}\delta_0\\N(\mu_y,\sigma_y^2)\end{pmatrix} + \pi_4\begin{pmatrix}\delta_0\\\delta_0\end{pmatrix}, \quad \phi = 0; \qquad (M3)$$

when InSIDE does not hold, we simulate from

$$\begin{pmatrix}\gamma\\\phi\\\alpha\end{pmatrix} \sim \pi_1\begin{pmatrix}N(0,\sigma_x^2)\\\delta_0\\\delta_0\end{pmatrix} + \pi_2\begin{pmatrix}N(0,\tilde{\sigma}_x^2)\\N(0,\sigma_u^2)\\N(\mu_y,\tilde{\sigma}_y^2)\end{pmatrix} + \pi_3\begin{pmatrix}\delta_0\\\delta_0\\N(\mu_y,\sigma_y^2)\end{pmatrix} + \pi_4\begin{pmatrix}\delta_0\\\delta_0\\\delta_0\end{pmatrix}. \qquad (M4)$$

Simulating Data on Genome-wide Association Studies

We simulate individual level data for independent genome-wide association studies for X and Y following the above model when sample size is not too large ($N \leq 100k$). We first conduct association analysis for each of the 200k SNPs with X using data from the underlying GWAS. SNPs which reach genome-wide significance (p-value $< 5\times10^{-8}$) are then selected as instruments and then we analyze association of each of these SNPs with Y using the underlying GWAS. However, for very large sample size, generation and analysis of individual level data can become computationally prohibitive and we simulate summary-level association statistics directly (addressed as *summary-level simulations*). We observe that the total effects of SNPs on X and Y are implied by model (1)-(3):

$$\beta_{jx} = \gamma_j + \theta_{Ux}\phi_j \quad (4)$$

$$\beta_{jy} = \alpha_j + \theta\beta_{jx} + \theta_{Uy}\phi_j \quad (5)$$

Thus, we simulate γ_j, ϕ_j and α_js as before and then directly generate summary statistics as $\hat{\beta}_{jx} = \beta_{jx} + N\left(0, \frac{1}{n_x}\right), \hat{\beta}_{jy} = \beta_{jy} + N\left(0, \frac{1}{n_y}\right)$ where the estimation error terms are generated with mean zero normal distribution with variances inversely proportional to n_x and n_y, the sample size of the study associated with X and Y, respectively. We explore sample sizes (N) varying from 50k to 1000k with $n_x = N$ and $n_y = N/2$, as well as other combinations of n_x and n_y without fixing the ratio. Both individual and summary-level simulations are repeated 200 times.

Choosing Parameters Values Reflecting Realistic Genetic Architecture

We found previous studies for evaluation of MR methods have often not followed realistic model for genetic architecture of complex traits. In particular, a very recent study that evaluated a large number of methods for MR analysis following the basic setup described in (1)-(3), used highly unrealistic parameter settings[32]. The study, for example, assumes 10% of the variance of X can be explained by the chosen instruments regardless of the number of instruments being 10, 30, 100 or 500. Results from existing GWAS reveal that complex traits tend to be extremely polygenic and only a very large number of variants can explain 10% of the variance of a trait. For example, latest GWAS has shown that more than 1000 SNPs are needed to explain 10% variance of a trait like BMI [33].

Further, the same study generated the direct effects of the chosen instruments on the outcome (α_j) to be so large that the instruments could explain more variability of Y than that of X. For example, under the scenario with balanced pleiotropy they considered, the proportion of variance explained by direct effects of the instruments on Y ranges from 20-35% for 30 SNPs and can go up to 90% for 500 IVs (**Supplementary Table 2**). Finally, the study fixes the sample size at 10,000 for both the exposure and the outcome – much smaller than the size of recent genome-wide association studies that have led to dozens or even hundreds of IVs for various traits. The study varies the number of instruments independently of the fixed sample size. In reality, sample size directly determines the number of instruments available and precision of their effects.

We chose parameter values in our model so that they reflect realistic genetic architecture of complex traits and results from our simulated GWAS track that are typically observed in empirical studies. In **Table 1**, we show the parameter values chosen for different simulation settings and corresponding values of heritability of the two traits and their co-heritability due to horizontal and vertical pleiotropy. Further in **Figure 1**, we show how under simulation settings as the sample size for GWAS of X increases, the number of available IVs and the amount of variance they explain for the two traits increase. These patterns closely correspond to that observed in GWAS of many traits, such as BMI. Further see **Supplementary Table 1** for the exact choice of parameters.

Existing Robust MR Methods

We compare all nine methods investigated in the recent study indicated above [32]. In addition, we include the inverse-variance weighted method with multiplicative random effects (IVW-r) in comparison [34]. The methods can be classified into the following categories:

A. Location parameter of ratio estimates

- IVW-r: The IVW estimator with multiplicative random effects is a simple extension of the standard IVW. IVW-r computes the estimate of causal effect using the same weights as fixed-effect IVW, but incorporates an over-dispersion parameter into the variance to account for pleiotropy [34].

- Weighted median: The weighted median method takes the median of the ratio estimates after assigning to them probabilistic weights that are inversely proportional to their variances. The underlying assumption of that method is that >50% of the weight comes from valid IVs [16].
- Weighted mode: The weighted mode estimator takes the mode of the smoothed empirical density function of the ratio estimates, using the same weights as the weighted median approach. The method requires the ZEro Modal Pleiotropy Assumption (ZEMPA) [17].
- MR-Egger: Egger regression fits the regression model $\hat{\beta}_y = \theta\hat{\beta}_x + \theta_0$, where $\hat{\theta}$ is the estimated causal effect and $\hat{\theta}_0$ is the estimated directional pleiotropy. Since IVW estimator is in effect the slope of a regression model through the origin, Egger regression is a direct extension of the method by allowing an intercept term accounting for directional pleiotropy. The method requires the Instrument Strength Independent of Direct Effect (InSIDE) assumption [18].

B. Robust regression

- MR-Robust: IVW is performed by fitting the regression using MM-estimation, which consists of an initial S-estimate followed by an M-estimate of regression[35], combined with Tukey's bi-weight loss function [29].
- MR-Lasso: The IVW regression is augmented by adding SNP-specific intercept terms, which represents SNP-specific pleiotropy effects, and penalizing the intercept terms with L1 loss function [29].
- MR-RAPS: Profile likelihood can be used for MR when there is no horizontal pleiotropy. The MR Adjusted Profile Score method incorporates random effect and robust loss functions into the profile score to account for systematic and idiosyncratic pleiotropy[36].

C. Outlier detection and removal

- MR-PRESSO: The MR Pleiotropy Residual Sum and Outlier (MR-PRESSO) method uses the leave-one-out sum of squared residuals to detect global horizontal pleiotropy. It also detects and removes outliers that are in horizontal pleiotropy and conducts a distortion test of the influence of the invalid IVs. The MR-PRESSO outlier test requires that at least 50% of the variants are valid instruments and relies on the InSIDE assumption [13]. Due to long computational time, we only implements MR-PRESSO up to sample size $N = 200k$.

D. Mixture model approach

- MRMix: MRMix uses a mixture model for effect-size distribution assuming existence of a fraction of the genetic markers that are valid instruments. Causal effects are estimated based on a novel spike-detection algorithm: it fits the mixture model $\hat{\beta}_y - \theta\hat{\beta}_x \sim \pi_0 N(0, \sigma_0^2) + (1 - \pi_0)N(0, \sigma^2)$ and searches for the θ that maximizes the probability concentration at the null component $N(0, \sigma_0^2)$ corresponding to valid IVs. This approach requires ZEMPA and tends to be robust and efficient under large sample size [37].
- Contamination mixture (addressed as Con-mix for simplicity): The contamination mixture approach also uses a mixture model to characterize a cluster of valid IVs and a cluster of invalid IVs. Unlike MRMix which models the genetic effect size, Con-mix models the ratio estimates using normal mixture with a pre-specified variance of the pleiotropic effects[19].

See **Supplementary Table 3** for the software and tuning parameters used to implement the methods above.

Summary of Simulation Results

We calculate the type I error rate of all 10 methods at nominal significance threshold $p < 0.05$, as well as the power of the methods that have well controlled or only moderately inflated type I error. We also calculate the mean squared error (MSE) as

$$MSE = \frac{1}{200}\sum_{r=1}^{200}\left(\hat{\theta}_r - \theta^*\right)^2,$$

where θ^* is the true value of the causal effect. The MSE measures the accuracy of the point estimate combining bias and variance. We also use the mean and standard deviation of causal estimates across simulations to compare bias and efficiency separately. To investigate bias of the underlying standard error (SE) estimates in some of the MR methods, we compare the empirical standard deviation of the causal effect estimates and the average estimated SE calculated across simulations. For the contamination mixture method [19], we calculate the "standard error" as length of the 95% confidence set divided by 2×1.96, since the method generates confidence sets based on the likelihood ratio test and does not report a standard error.

Results

We present main results under the simulation scheme that generate summary-level data directly as it allows exploration of GWAS of very large sample size ($N > 100k$). Under smaller sample size where we did simulation with both individual and summary-level data, we see the results are very comparable across the two schemes (see **Supplementary Figures 1 and 2**).

Under balanced pleiotropy and InSIDE assumption, the weighted mode estimator and MRMix controls type I error at the nominal level across different scenarios (**Figure 2**). The type I error rate of weighted mode usually falls far below the nominal value while that of MRMix generally remains fairly close to the nominal value. Among other methods, IVW-r, MR-Egger, MR-Robust and MR-RAPS are the most robust as they generally maintain the nominal type I error except when the number of invalid IVs are large or/and sample size is small. The least robust methods are MR-Lasso and Con-mix which often have extremely high type I error reaching even up to 100% for the latter method when $N = 1000k$ and 70% of the instruments were invalid. When we compare the methods in terms of MSE (**Figure 3**), we find that IVW-r, weighted median, MR-Robust, MR-Lasso, MR-PRESSO and MR-RAPS perform comparably to each other and have an advantage over the others. Among weighted mode and MRMix, which are the only two methods maintaining nominal type-I error, we find MSE for weighted mode could be much smaller than MRMix for smaller sample size (e.g $N = 50k$) but the latter method has a clear advantage when sample size becomes larger ($N \geq 200k$).

Among the methods that have reasonably well controlled or only moderately inflated type I error rates, power tracks closely with the MSE (**Supplementary Figure 3**). IVW-r, MR-Robust, MR-RAPS have the highest power; MRMix has similar power to these three methods when $N \geq 200k$ and lower power at smaller sample size. Weighted mode and Egger regression have lower power than the other methods throughout the scenarios.

We also compared the different methods in terms of bias and variance separately. Under balanced pleiotropy and InSIDE assumptions, all MR methods generally show some bias when the sample size is small and gradually the bias diminishes as the sample size increases

(**Supplementary Figure 4**). The observed biases are generally towards the null except for Con-mix which was often biased away from the null. When sample size is very large ($N = 1000K$) and number of invalid IVs are correspondingly large, it appears that the Con-mix method completely fails and can create very large bias. When there is no causal effect, all methods give average estimates of causal effect close to 0 across wide range of sample sizes except for Egger regression with $N \leq 100k$ and Con-mix with $N = 1000k$ and large number of invalid IVs. Comparing the empirical versus estimated standard errors of the methods we observe that the MR-Lasso and Con-mix produce severely underestimated standard errors (**Supplementary Figure 5**), while weighted median and MR-PRESSO has underestimated standard error when 70% of the instruments are valid. This is likely to be the reason for type I error inflation. Throughout the settings IVW-r, MR-Egger, MR-Robust and MR-RAPS give accurate standard error estimation; weighted mode has overestimated standard error which is likely to be the reason for its overly conservative type I error rate. The standard error estimate of MRMix tends to be too large when $N \leq 100k$, but converges to the truth when $N \geq 200k$.

When the InSIDE assumption is violated, we find all methods except weighted mode and MRMix could have extremely inflated type I error (**Figure 4**). As before while MRMix maintains type I error close to the nominal level, weighted mode is very conservative. Among other methods, Con-mix and MR-Egger are less biased, but even these methods have unacceptably high type I error in a variety of scenarios. When the methods are compared in terms of MSE (**Figure 5**), for smaller sample sizes ($N \leq 100k$), the methods IVW-r, weighted median, MR-Robust, MR-Lasso and MR-PRESSO seem to be the best even though they may have appreciable bias. For larger sample size ($N \geq 200k$), MRMix generally has the smallest MSE; and had substantially higher power than weighted mode, which also controls the type I error (**Supplementary Figure 6**). When we inspected bias and variance separately, it is evident that when the InSIDE assumption is violated all methods, except weighted mode and MRMix, can have large bias, even when there is no causal effect, and this bias does not disappear with increasing sample size (**Supplementary Figure 7**). The patterns of bias in standard error estimation for the different methods are similar as what we described before when the InSIDE assumption is satisfied (**Supplementary Figure 8**).

When we simulate unbalanced pleiotropy without and with violation of the InSIDE assumption, the patterns are fairly similar with corresponding scenarios for balanced pleiotropy setting except that the type I error for a number of methods increased somewhat in certain settings

(**Supplementary Figures 9-14**). In particular, the MRMix method now shows modestly inflated type I error when sample size is small and the number of invalid IVs is large (e.g. 70%), but the bias disappears with increasing sample size (**Supplementary Figures 9 and 11**). The weighted mode method maintains type I error in a conservative manner in all settings. All the other methods have extremely inflated type I errors in a range of scenarios, especially when sample size is large, the number of invalid IVs are large or/and the InSIDE assumption is violated.

To further understand the role of sample size in MR analysis, we conducted additional simulations where the sample sizes of GWAS of X and Y are allowed to vary freely without being held fixed at 2:1 proportion. It appears that the sample size of X plays a bigger role but the sample size of Y can also play an important role on relative performance of the methods (**Supplementary Figures 15 and 16**). For example, when sample size for GWAS of X is relatively large ($n_x = 200K$), but that for Y is small ($n_y = 25K$), MR-Lasso had the smallest MSE and is closely followed by a number of other methods including IVW-r, weighted median, MR-Robust, MR-RAPS and MR-PRESSO. In this setting, MR-Egger and Con-mix have much higher MSE and those of weighted mode and MRMix are in between. In contrast, when the sample size for GWAS of Y increased, MRMix emerged clearly as the best performing method for intermediate sample size $n_y = 50k$ to $100k$ and MR-Lasso again has the smallest MSE for even larger sample size $n_y = 200k$ to $1000k$ and Con-mix follows closely. We further observe that when sample size for GWAS of X is much larger than that of Y ($n_x = 1000k$ and $n_y = 25k$), the MRMix method clearly has the smallest and Con-mix has the largest MSE among all methods. In the other extreme, when the sample size for GWAS of Y is very large and that for X is small, the Con-mix also performed poorly. In this setting, the methods IVW-r, weighted median, MR-Robust, MR-Lasso, MR-RAPS and MR-PRESSO had much smaller MSE and performed comparably to each other. The performance of weighted mode and MRMix is intermediate.

Discussion

In this paper, we evaluate a variety of methods for polygenic MR analysis using a simulation framework which generate data closely resembling patterns observed in empirical genome-wide association studies. Results reveal varying performance of the MR methods under different scenarios. When the sample size is large (e.g. $n_x > 200k, n_y > 100k$), MRMix appears to be best or close to be best, whether or not InSIDE assumption is satisfied, in terms of its ability to control type I error rate and bias and yet maintaining relatively high power and low MSE. When the sample size is smaller (e.g. $n_x \leq 100k, n_y \leq 50k$), no method appears to be performing uniformly well across all scenarios. When the InSIDE assumption holds, IVW-r, MR-Robust and MR-RAPS lead to the smallest mean-squared errors and they usually have either well controlled or modestly inflated type I error rates. The weighted median method also performs well when the proportion of valid IVs is not too high (e.g. $\leq 30\%$) but suffers from more severe type I error when this proportion increases. When the InSIDE assumption is violated, only weighted mode and MRMix have well controlled type I error and all the other methods can have severely inflation in type I error. Between these two methods, weighted mode tends to be more efficient and powerful for moderate sample size (e.g. $n_x \leq 80k, n_y \leq 40k$).

We observe that the type I error of a number of methods are significantly affected due to not only bias in point estimation but also that of the underlying standard error estimators. In particular, we found that the type I error of the weighted mode can often be substantially lower than the desired nominal level due to conservativeness of underlying standard error estimator (**Supplementary Figure 5 and 8**). Further for a number of other estimators, which did not have bias in point estimation at least when the InSIDE assumption is satisfied, have inflated type I error due to anticonservative standard error estimation. It is possible that in the future the type I error or/and power for some of these procedures can be improved through implementation of more robust standard error estimation procedures.

Both similarities and differences exist between our simulation results and the results reported in a recent study also comparing most of the same MR methods [32]. Both studies found that the weighted mode estimator has well controlled type I error rate; among the outlier-robust methods (MR-PRESSO, MR-Robust, MR-Lasso), MR-Lasso has the lowest MSE but MR-Robust has better controlled type I error rates. The biggest difference is observed for mixture model based methods. In our study, MRMix is shown to perform well under large sample size, especially when the InSIDE assumption is violated. The study by Slob and Burgess restricted their simulation studies with sample size $n_x = n_y = 10{,}000$. Under such small sample size, MRMix can be unstable because the performance of the method depends on the ability of the underlying mixture model to cluster valid and non-valid IVs based on underlying variance components. If the sample size is small and estimates of effect sizes for the IVs have large variability, then the two variance components are not well separable and the resulting estimates can have large uncertainty. When sample size increases, variance component associated with valid IVs are expected to be smaller than those with the non-valid IVs and then the method allows robust estimation of causal effects.

Another major distinction of our study with that of Slob and Burgess is that that these authors allow extremely large direct effects for a relatively small number of IVs on the outcome Y, while we use a more realistic model that allows much larger number of invalid IVs, which individually has direct effects on Y of smaller magnitudes (see **Table 1** and **Supplementary Table 2).** In our own simulation study, we find the alternative mixture model-based method, Con-mix can have smaller MSE than that of MRMix specially for smaller sample sizes, but the former method can have much higher type I error across a variety of scenarios with or without the InSIDE assumption violated. We also observed a numerical breakdown of Con-mix for very large GWAS (e.g $n_x = 1000k$) for which the cause is not well understood.

The simulation framework we propose can be broadly useful for future evaluation of emerging MR methods. We simulate data on genome-wide panel of SNPs and apply a p-value threshold to select IVs as is done in real studies. This procedure naturally reflects the relationship among sample size, number of IVs and instrument strength. We also use realistic distributions to generate genetic effect sizes based on recent work on heritability and effect size distributions [10,20–28]. We studied the performance of MR methods in a wide range of sample sizes and scenarios of violations of standard assumptions of MR analysis. We propose a framework for directly simulating summary-level data implied by the model for individual data for reducing the computational burden associated with simulating vary large GWAS.

In summary, we conducted large-scale and realistic simulation studies to compare 10 methods for Mendelian randomization analysis. Our results show that while for GWAS with very large size the mixture model based method MRMix emerges as the most robust method, for medium to smaller sample sized studies there is no single method that performs uniformly well across all scenarios. Thus in real data analysis it is prudent to apply a few alternative methods with complimentary features and strengths and assess sensitivity of findings across all these methods.

Software code availability

The code for the simulation studies is available on GitHub: https://github.com/gqi/MR_comparison_simulations

Acknowledgement

The research was supported by a RO1 grant from the National Human Genome Research Institute (1 R01 HG010480-01).

References

1. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease. *Int J Epidemiol* **32**, 1-22 (2003).
2. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89-R98 (2014).
3. Zheng, J. et al. Recent developments in Mendelian randomization studies. *Current Epidemiology Reports* **4**, 330-345 (2017).
4. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized Data. *Genet. Epidemiol.* **37**, 658-665 (2013).
5. Harbord, R. M. et al. Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies. *Stat Med* **32**, 1246-1258 (2013).
6. Pingault, J.-B. et al. Using genetic data to strengthen causal inference in observational research. *Nat Rev Genet* **19**, 566-580 (2018).
7. Porcu, E. et al. Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *bioRxiv* 377267 (2019).
8. Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L. & Smith, G. D. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *BioRxiv* 563379 (2019).
9. Sanna, S. et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat Genet* 1 (2019).
10. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-1236 (2015).
11. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**, 709 (2016).
12. Sivakumaran, S. et al. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607-618 (2011).

13. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* **50**, 693 (2018).
14. Visscher, P. M. & Yang, J. A plethora of pleiotropy across complex traits. *Nat Genet* **48**, 707-708 (2016).
15. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).
16. Bowden, J., Davey, S. G., Haycock, P. C. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304-314 (2016).
17. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol* **46**, 1985-1998 (2017).
18. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-525 (2015).
19. Burgess, S., Foley, C. N., Allara, E., Staley, J. R. & Howson, J. M. M. A robust and efficient method for Mendelian randomization with hundreds of genetic variants: unravelling mechanisms linking HDL-cholesterol and coronary heart disease. *bioRxiv* 566851 (2019).
20. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-291 (2015).
21. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2542 (2012).
22. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics* **91**, 1011-1021 (2012).
23. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat Genet* **51**, 277 (2019).
24. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565 (2010).
25. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18**, 275-294 (2016).
26. Zeng, J. et al. Signatures of negative selection in the genetic architecture of human

complex traits. *Nat Genet* **50**, 746 (2018).

27. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet* **50**, 1318 (2018).
28. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications* **9**, 4361 (2018).
29. Burgess, S., Bowden, J., Dudbridge, F. & Thompson, S. G. Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. *arXiv preprint arXiv:1606.03729* (2016).
30. Burgess, S., Zuber, V., Gkatzionis, A. & Foley, C. N. Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. *Int J Epidemiol* dyy080 (2018).
31. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284 (2015).
32. Slob, E. A. W. & Burgess, S. A Comparison Of Robust Mendelian Randomization Methods Using Summary Data. *BioRxiv* 577940 (2019).
33. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in~ 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
34. Bowden, J. et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med* **36**, 1783-1802 (2017).
35. Koller, M. & Stahel, W. A. Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis* **55**, 2504-2515 (2011).
36. Zhao, Q., Wang, J., Bowden, J. & Small, D. S. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *arXiv preprint arXiv:1801.09652* (2018).
37. Qi, G. & Chatterjee, N. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications* **10**, 1941 (2019).

Figures and Tables

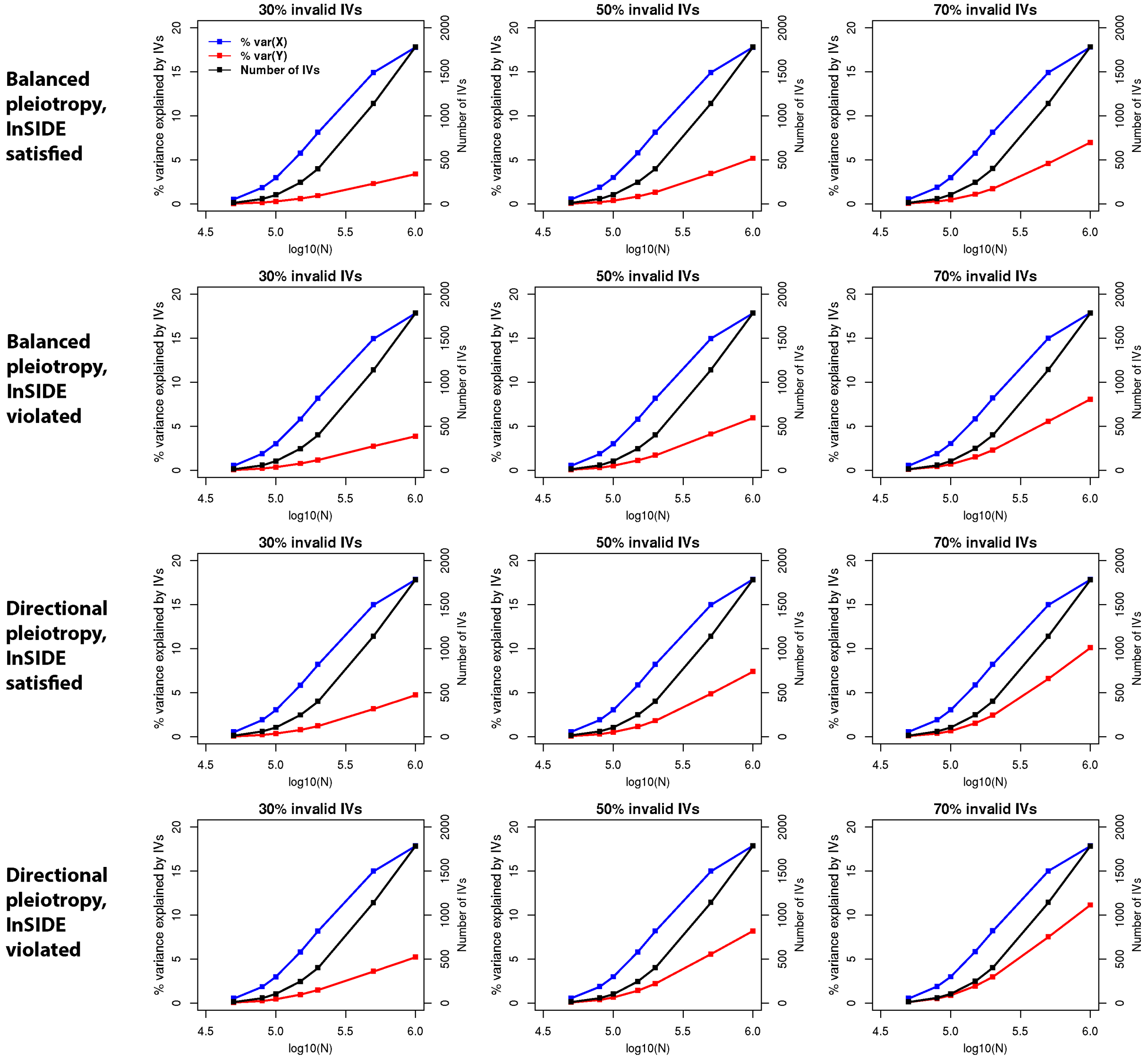

Figure 1. Relationship among sample size, average number of IVs and variance of traits explained by the IVs under different scenarios of simulation studies. The true causal effect from X to Y is 0.2. Sample size of the study associated with X is N; sample size of the study associated with Y is $N/2$. IVs are defined as the SNPs which reach genome-wide significance (z-test $p < 5\times10^{-8}$) in the study associated with X. Averages are calculated over 200 simulations.

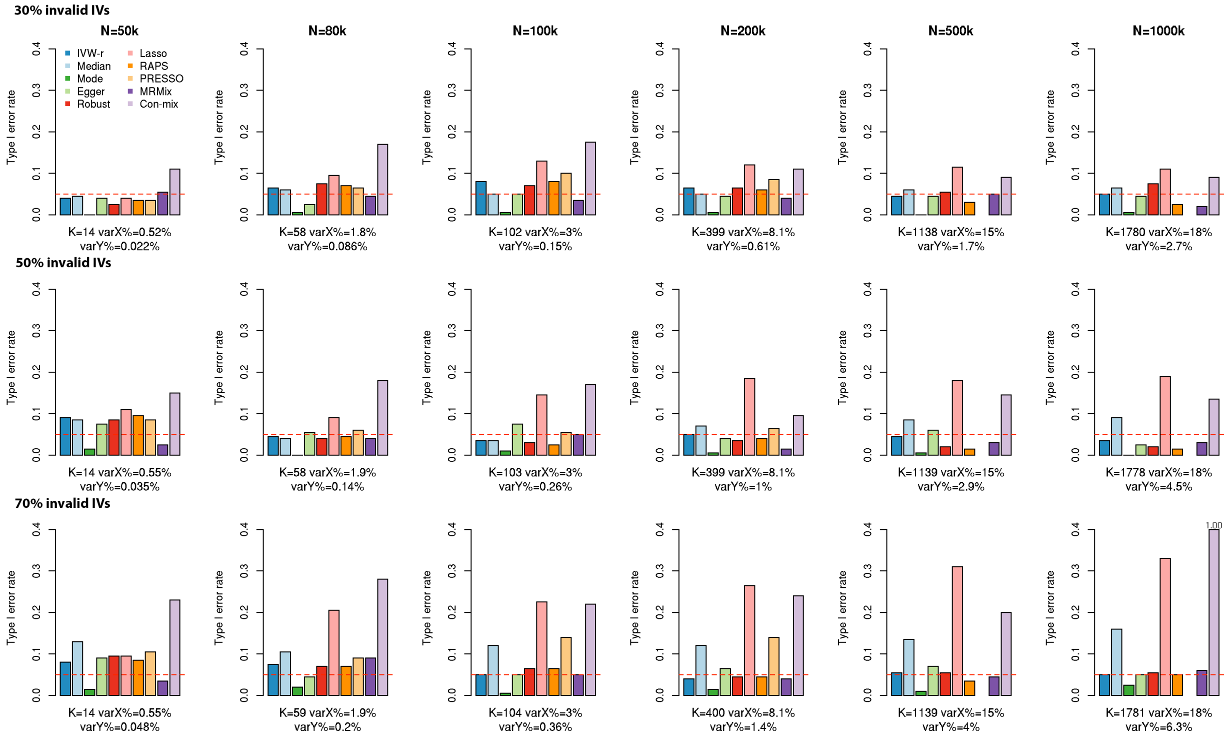

Figure 2. Type I error rates for alternative MR methods in simulations with balanced pleiotropy and InSIDE assumption satisfied. There is no true causal effect from X to Y. Empirical type I error rates are reported over 200 simulations. Sample size of the study associated with X is N; sample size of the study associated with Y is $N/2$. K: the average number of IVs, defined as the SNPs which reach genome-wide significance (z-test $p < 5\times10^{-8}$) in the study associated with X; varX%: average percentage of variance of X explained by IVs; varY%: average percentage of variance of Y explained by IVs. The red dashed line is the nominal significance threshold 0.05.

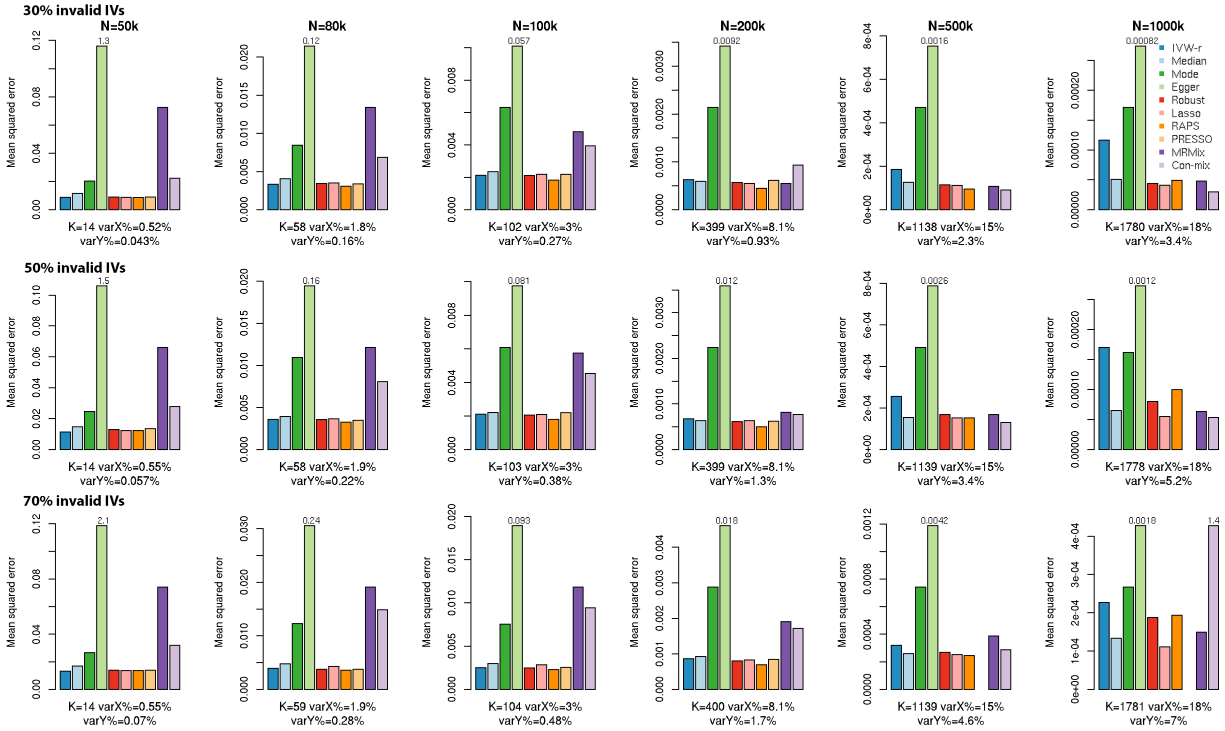

Figure 3. Mean squared error for estimation of causal effect under alternative MR methods in simulations with balanced pleiotropy and InSIDE assumption satisfied. The true causal effect from X to Y is 0.2. Mean squared errors are reported over 200 simulations. Sample size of the study associated with X is N; sample size of the study associated with Y is $N/2$. K: the average number of IVs, defined as the SNPs which reach genome-wide significance (z-test $p < 5\times10^{-8}$) in the study associated with X; varX%: average percentage of variance of X explained by IVs; varY%: average percentage of variance of Y explained by IVs. Bars higher than the upper limit of the panel are truncated and marked with the true value.

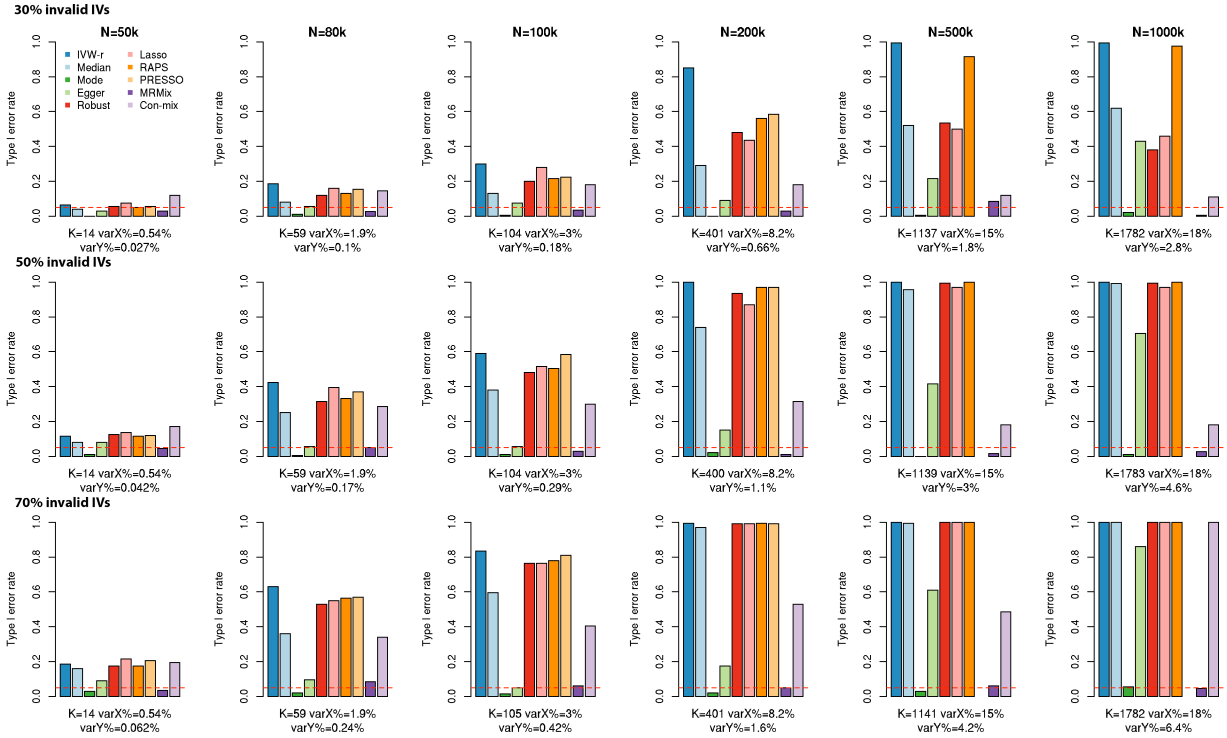

Figure 4. Type I error rate in in simulations with balanced pleiotropy and InSIDE assumption violated. There is no true causal effect from X to Y. Type I error rates are reported over 200 simulations. Sample size of the study associated with X is N; sample size of the study associated with Y is $N/2$. K: the average number of IVs, defined as the SNPs which reach genome-wide significance (z-test $p < 5\times10^{-8}$) in the study associated with X; varX%: average percentage of variance of X explained by IVs; varY%: average percentage of variance of Y explained by IVs. The red dashed line is the nominal significance threshold 0.05.

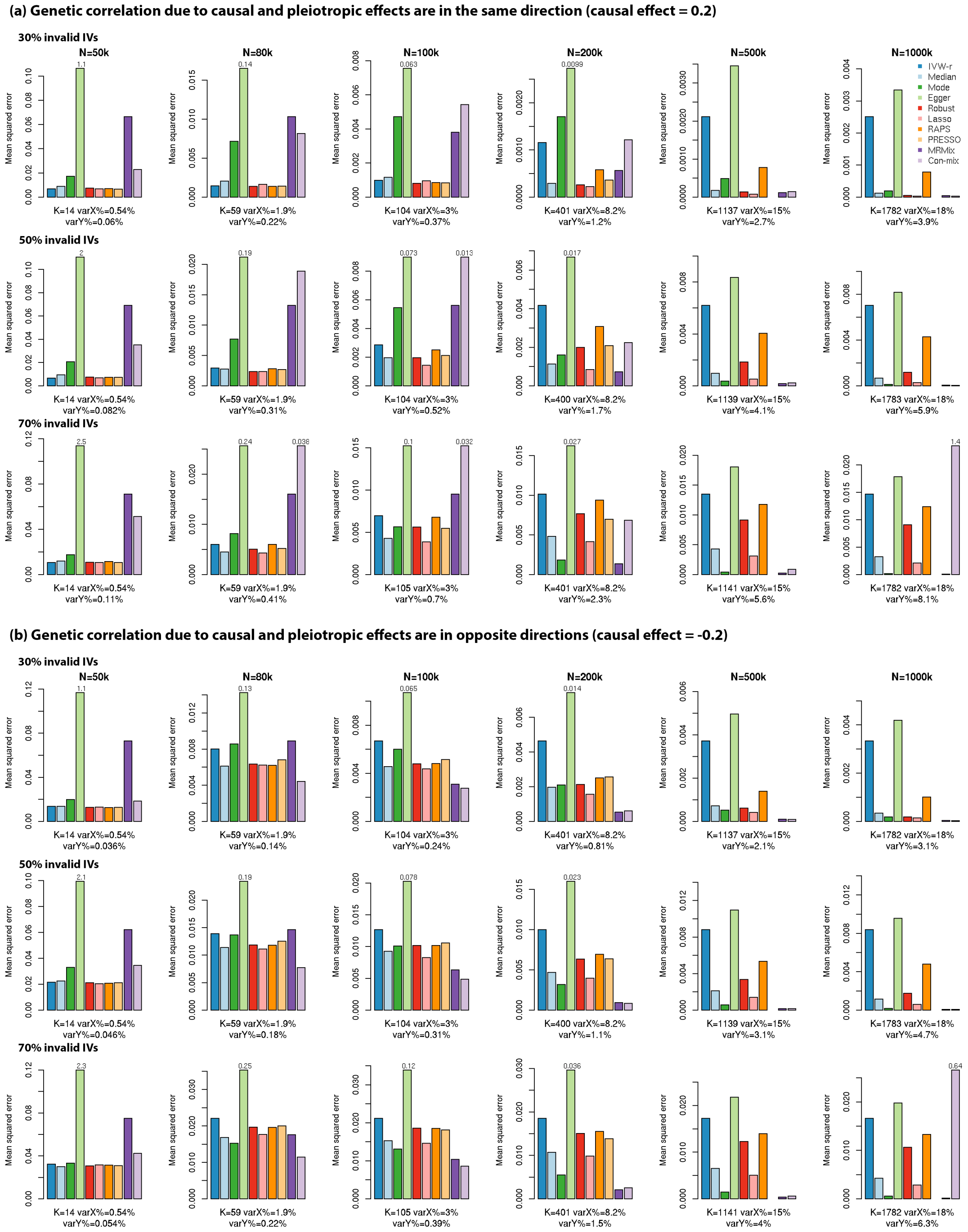

Figure 5. Mean squared error in in simulations with balanced pleiotropy and InSIDE assumption violated. Mean squared errors are reported over 200 simulations. Sample size of the study associated with X is N; sample size of the study associated with Y is $N/2$. K: the average number of IVs, defined as the SNPs which reach genome-wide significance (z-test $p < 5\times10^{-8}$) in the study associated with X; varX%: average percentage of variance of X explained by IVs; varY%: average percentage of variance of Y explained by IVs. Bars higher than the upper limit of the panel are truncated and marked with the true value.

Table 1. Choice of parameters in simulation studies where the true causal effect of X on Y is 0.2 and 50% of the potential IVs are valid.

Setting	# of causal variants	Heritability of X	Heritability of Y [b]	Genetic correlation [c] due to vertical/horizontal pleiotropy
Balanced pleiotropy, InSIDE satisfied [Model (M1), Methods]	200,000 independent SNPs representing the common variants in the genome. 1) 2,000 SNPs have direct effect on X but not Y (addressed as **potential valid IVs**). 2) 2,000 SNPs have direct effects on both X and Y (addressed as **SNPs in horizontal pleiotropy**). 3) 2,000 SNPs have direct effects on Y but not X. 4) The remaining 194,000 SNPs have no effects on either X or Y.	**Total 20%**: 10% due to potential valid IVs; 10% due to SNPs in horizontal pleiotropy.	**Total 20.8%**: 0.8% due to causal effect of X; 10% due to SNPs in horizontal pleiotropy; 10% due to SNPs that have effects on Y but not X.	0.196 / 0
Balanced pleiotropy, InSIDE violated [Model (M2), Methods]		**Total 20%**: 10% due to potential valid IVs; 8.2% due to direct effects of SNPs in horizontal pleiotropy; 1.8% due to confounder U [a].	**Total 21.52%**: 0.8% due to causal effect of X; 2.52% due to confounder U and its covariance with X; 8.2% due to direct effects from SNPs in horizontal pleiotropy that are not mediated by U; 10% due to SNPs that have effects only on Y.	0.193 / 0.087
Directional pleiotropy, InSIDE satisfied [Model (M3), Methods]		**Total 20%**: 10% due to potential valid IVs; 10% due to SNPs in horizontal pleiotropy.	**Total 30.8%**: 0.8% due to causal effect of X; 15% due to SNPs in horizontal pleiotropy (10% balanced + 5% directional effect); 15% due to SNPs that have effects only on Y.	0.161 / 0
Directional pleiotropy, InSIDE violated [Model (M4), Methods]		**Total 20%**: 10% due to potential valid IVs; 8.2% due to direct effects of SNPs in horizontal pleiotropy; 1.8% due to confounder U.	**Total 31.52%**: 0.8% due to causal effect of X; 2.52% due to confounder U and its covariance with X; 13.2% due to direct effects from SNPs in horizontal pleiotropy that are not mediated by U (8.2% balanced + 5% directional effect); 15% due to SNPs that have effects only on Y.	0.159 / 0.072

See Supplementary Table 1 for the exact choice of parameters in all settings.

[a] U is the heritable confounder.

[b] When directional pleiotropy exists, the direct SNP effects on Y follow distribution $N\left(\mu_y, \tilde{\sigma}_y^2\right)$, which can be decomposed into $N\left(0, \tilde{\sigma}_y^2\right) + \mu_y$. "Balanced effects" refer those from the first component $N\left(0, \tilde{\sigma}_y^2\right)$ and "directional effects" refer to those from the second component μ_y.

[c] Genetic correlation between two traits X and Y are defined as $h_{xy}/\sqrt{h_x^2 h_y^2}$, where h_{xy} is the covariance of the genetic components of X and Y, h_x^2 and h_y^2 are the heritability of X and Y respectively.