STAT 232 Final Project: A Review of Mendelian Randomization

Wei-Hsiang Sun

Abstract

In this paper, we give an overview of Mendelian randomization (MR), including the two frameworks of Mendelian Randomization , the assumptions and commonly used models. Additionally, we also do a survey on estimators used when violations of assumptions occur. Furthermore, we conduct a simulation-based comparison of the two frameworks under different scenarios. We find that under some cases, the one sample individual data estimators would outperform the two sample summary statistics estimators.

1 Introduction

Mendelian randomization (MR) is an approach that is commonly used to help infer the causal effect of some exposure or risk factor on an outcome of interest. It is especially used in observational studies where unmeasured confounding is prevalent and randomized controlled trials (RCTs) are infeasible (Davey Smith and Ebrahim (2003)).

Recent examples using Mendelian randomization to infer causal effects includes, the effect of microbiome and blood metabolites by Liu et al. (2022); the relationship between iron overload and type 2 diabetes by Wang et al. (2021b); the causal effect of Vitamin D levels on risk of type 1 diabetes by Manousaki et al. (2021); and the causality between diverse cardiometabolic conditions, such as obesity, and Covid-19 illness by Leong et al. (2021). Additionally, applications of Mendelian randomization are not limited to public health or epidemiology, MR is also prevalent in socioeconomics and other fields where causality is important whereas unconfoundedness takes place. Such examples are presented by Richmond and Smith (2022).

MR uses genetic variants, also known as single nucleotide polymorphisms (SNPs) [1] as instrumental variables. Instrumental variables (IVs) are variables that satisfies three assumptions, they are (I) related to the risk factor (relevance), (II) independent of the unmeasured confounders (independence) (III) effects the outcome only through the risk factor (exclusion restriction). We will further formalize the three assumptions for instrumental variables in the later section. In an RCT, the treatment assignment is randomized in a way such that the IV assumptions were met. In parallel with RCT, in MR context, instead of randomization of treatment assignment, we have the SNPs that randomly occur. Sanderson et al. (2022) has also drew a comparison between RCT and MR.

Many MR literature reviews have been conducted, recent examples includes, Sanderson et al. (2022), Richmond and Smith (2022), Bowden and Holmes (2019), Emdin et al. (2017), Zheng et al.

[1] A single nucleotide polymorphism (SNP) is a variation at specific location on a gene. A SNP can be viewed as a variable assuming one of the three values 0, 1 or 2. The values represents the number of variations at the specific gene location. In other words, a person can have no variation or at most 2 variations at a specific location in a gene.

(2017), and so forth. Most reviews have went through the assumptions, models and estimators commonly used in MR, in particular, estimators used when violations of IV assumptions or other core assumptions take place. However, few elaborate further on how each estimator is constructed . Furthermore, there are two frameworks in MR, the one-sample individual data framework and the two-sample summary statistics framework. The two frameworks would be elaborated in section 3. Most literatures focus on the two sample summary statistics framework, to the best of our knowledge, none have done comparisons between estimators from the two frameworks. Therefore, in this review, we also attempt to compare methods from each framework through simulations.

The paper is organized as follows. We give a basic model setup for MR in section 2. Then give a high level view of the two frameworks in MR in section 3. Then we review estimators commonly used in MR in section 4, 5. Then we look at a simulation based cross comparison between the two frameworks in section 6. Lastly, discussions about directions to work on in the future are described in section 7.

2 Model Setup Without Violation of IV assumption

We review the three instrumental variable (IV) assumptions, then describe a data generating process commonly used in MR assuming the three IV assumption holds.

2.1 IV assumptions

Let G_j be the jth SNP, X be the risk factor or exposure of interest, Y be the outcome of interest, U be the unmeasured confounder, γ_j be the constant effect of the jth SNP on exposure X. Then we say that G_j is a valid instrument if

(I) **Relevance** The instrument G_j is related to the exposure X, that is, $\gamma_j \neq 0$.

(II) **Independence** The instrument G_j is independent of the unmeasured confounder U.

(III) **Exclusion Restriction** The instrument G_j only effects the outcome through exposure X.

The assumption is also illustrated in the DAG given in Figure. 1.

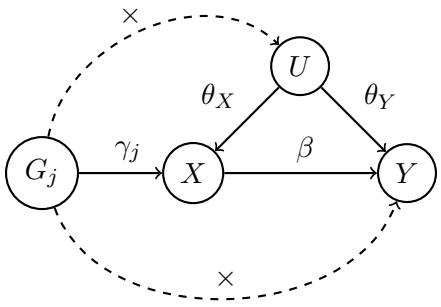

Figure 1: DAG to illustrate the 3 instrumental variable assumptions.

2.2 Linear data generating process

Supposing the IV assumptions were true, and assuming the relationship among variables linear, we can form a data generating process described by two structural equations. The model setup follows that of Ye et al. (2021b). Let the p independent SNPs be bounded and mutually independent random variables, denoted by $G_1, \ldots, G_p$. Let X be the exposure and Y be the continuous outcome. Consider the structural equation Model

$$X = \sum_{j=1}^{p} \gamma_j G_j + \theta_X U + E_X, \tag{1}$$

$$Y = \beta X + \theta_Y U + E_Y \tag{2}$$

$\theta_X, \theta_Y, \beta, \gamma_1, \ldots, \gamma_p$ are unknown parameters. U is an unmeasured confounder independent of $G_1, \ldots, G_p$, E_X and E_Y are mutually independent random noises that are also independent of $(G_1, \ldots, G_p, U)$, and U, E_X and E_x have finite 4th order moments. We are interested in the inference of the causal parameter β.

3 Two frameworks within MR

We now elaborate the two frameworks in MR, the one-sample individual data framework and the two-sample summary data framework. The two frameworks are different in terms of data availability, causing the estimators applicable to each framework to be different.

Data from one-sample individual data framework, all comes from one sample, in the data generating process 1, 2, the data available for individual i would be the p SNPs data $G_{i1}, \ldots, G_{ip}$, the exposure data X_i and the outcome data Y_i.

Due to public availability of Genome-wide association studies (GWAs) summary data and difficulty to obtain one-sample individual data, MR analysis based on two-sample summary data instead became a more popular alternative to one-sample individual data (Richmond and Smith (2022)). The two sample summary data utilizes two sets of samples, one sample is used to estimate the marginal effect of each SNP $G_1, \ldots, G_p$ on exposure X, denote the effect estimates $\hat{\gamma}_1, \ldots, \hat{\gamma}_p$, and the corresponding standard errors $\sigma_{X_1} \ldots \sigma_{X_p}$; another sample is used to estimate the marginal effect of each SNP $G_1, \ldots, G_p$ on outcome Y, denote the estimates $\hat{\Gamma}_1, \ldots, \hat{\Gamma}_p$, and the corresponding standard errors $\sigma_{Y_1}, \ldots, \sigma_{Y_p}$. The two-sample summary statistics data can be further elaborated using table 1.

	Original datasets	Public available summary data
1. Exposure dataset	$\{(X_i, G_{ij})\}_{i=1}^{n_X}$	$\{(\hat{\gamma}_j, \sigma_{X_j})\}_{j=1}^{p}$
2. Outcome dataset	$\{(Y_i, G_{ij})\}_{i=1}^{n_Y}$	$\{(\hat{\Gamma}_j, \sigma_{Y_j})\}_{j=1}^{p}$

Table 1: Table illustrating two-sample summary data framework. Let n_X denote the sample size used to estimate the marginal SNP-exposure effect, and n_Y denote the sample size used to estimate the marginal SNP-outcome effect. $\hat{\gamma}_j$ is obtained by the regression $X \sim G_j$, with standard error σ_{X_j}. $\hat{\Gamma}_j$ is obtained by the regression $Y \sim G_j$, with standard error σ_{Y_j}. By convention, σ_{X_j} and σ_{Y_j} are assumed to be known.

4 Overview of MR Estimators

4.1 Methods for Individual Data

Suppose one-sample individual data is available, it is common to directly apply instrumental variable related methods from econometrics to obtain the causal estimates. Andrews et al. (2019) gave an overview of such methods, which includes the k-class estimators. Alternatively, we can also consider generalized method of moments (GMM) estimator (Newey (2007)). One could also consider applying two-sample summary statistics method to individual data. However, this would lead to additional bias when weak instruments are prevalent (Barry et al. (2021)).

4.2 Two sample summary data estimator: IVW

With two-sample summary data available. In addition to the three IV assumptions, we impute distributional assumptions on the data. The goal is to obtain the maximum likelihood estimator for the causal effect β.

Assumption 1.

$$\begin{bmatrix} \hat{\Gamma}_j \\ \hat{\gamma}_j \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \Gamma_j \\ \gamma_j \end{bmatrix}, \begin{bmatrix} \sigma^2_{Y_j} & 0 \\ 0 & \sigma^2_{X_j} \end{bmatrix} \right) \tag{3}$$

The $\{\hat{\gamma}_j, \hat{\Gamma}_j | j = 1, \ldots, p\}$ are mutually independent.

We directly cite the assumption used by Zhao et al. (2020). The approximation of the summary statistics to normality is plausible since $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ are obtained from data of large sample size. Elements in $\{\hat{\gamma}_j | j = 1, \ldots, p\}$ are mutually independent of elements in $\{\hat{\Gamma}_j | j = 1, \ldots, p\}$ since the two sets of estimates are obtained from two datasets with non-overlapping samples. The elements within $\{\hat{\gamma}_j | j = 1, \ldots, p\}$ are mutually independent since the SNPs are often decorrelated through data pruning or clumping (Ye et al. (2021b)). The same goes for elements in the set $\{\hat{\Gamma}_j | j = 1, \ldots, p\}$.

Assumption 2 (No measurement error)**.** The no measurement error (NOME) assumption assumes the standard error σ_{X_j} for the SNPs-exposure effect $\hat{\gamma}_j$ equals to zero for all $j = 1, \ldots, p$.

This assumption will likely hold since the sample size used to obtain $\hat{\gamma}_j$ is often sufficiently large. However, many later developed estimators drop this assumption, such as the debiased IVW estimator proposed by Ye et al. (2021b).

Model 1.

$$\Gamma_j = \beta \gamma_j \tag{4}$$

Assuming the data generating process follows equation 1 and 2, then model 4 follows directly by plugging equation 1 into 2.

With distribution and model assumptions for two-sample summary statistics data, the complete data likelihood can be obtained, and the corresponding maximum likelihood estimator is shown in equation 5.

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{p} w_j \hat{\beta}_j}{\sum_{j=1}^{p} w_j} \tag{5}$$

where $\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$, and $w_j = \left(\frac{\sigma^2_{Y_j}}{\hat{\gamma}_j^2} + \frac{\hat{\Gamma}_j \sigma^2_{X_j}}{\hat{\gamma}_j}\right)^{-1}$. The IVW estimator is a weighted average of the ratio $\hat{\Gamma}_j/\hat{\gamma}_j$, which corresponds to the effect of the jth SNP on outcome divided by the effect of the jth SNP on exposure. The estimator gives larger weights to SNPs with larger SNPs-exposure effect size $\hat{\gamma}_j$, and smaller SNPs-outcome effect standard error σ_{Y_j}. Note that, as shown in Zhao et al. (2020), after profiling out the $\gamma_1, \ldots, \gamma_p$ from the likelihood function, the solution to the profile likelihood corresponds to running the regression of $\hat{\Gamma}_j$ on $\hat{\gamma}_j$ excluding the intercept, which can be formulated as 6. Note that both sides of equation 6 are often divided by σ_{Y_j} to make the variance of the error term homoscedastic.

$$\hat{\Gamma}_j = \beta \hat{\gamma}_j + \epsilon_j \tag{6}$$

4.3 Connection to noncompliance in experiments

Recall that given one instrument G_j, the complier average causal effect (CACE) of a treatment X on Y can be viewed as the ratio

$$CACE = \frac{ITT_Y}{ITT_X} \tag{7}$$

Where intention to treat on Y, ITT_Y, can be obtained by regressing Y on G_j, and similarly ITT_X can be obtained by regressing X on G_j. This corresponds to the ratio $\hat{\Gamma}_j/\hat{\gamma}_j$ in the IVW estimator. So the IVW estimator can be roughly viewed as a weighted average of CACE's. However, in the experiment setting, the population is well defined as the participant in the experiment. While in MR context, we assume a infinite population framework. Additionally, in MR we have continuous treatment (or exposure), while in experiments, we often define discrete treatments and thus are able to divide the population into latent strata. Thus IVW estimator can only be viewed analogously as weighted average of CACE's but does not equate the CACE in randomized experiments. Instead of potential outcomes, a more adequate framework to look at MR would be through the lens of causal inference defined by structural equations (Pearl et al. (2000)), in which Pearl et al. (2000) also drew a connection between potential outcomes and causality defined by structural equations.

4.4 Importance of using multiple instruments

The use of multiple instruments increases statistical efficiency of the IVW estimator. By the delta method, the first order Taylor expansion of the standard error of the IVW estimator is shown by Burgess et al. (2013) to follow equation 8.

$$se(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_{j=1}^{p} \hat{\gamma}_j^2 \sigma_{Y_j}^{-2}}} \tag{8}$$

where multiple instruments reduces the variance due to increase in the denominator of the approximated standard error 8.

Another reason to include multiple instruments, is that consistency of causal estimators may depend on the number of instruments included in estimation, the debiased IVW estimator proposed by Ye et al. (2021b) is an example of such.

5 Methods Dealing with Violation of Key Assumptions

Estimation and inference in MR relies heavily on assumptions in order for the causal effect estimator to attain ideal properties, such consistency (Bowden et al. (2017)). However, it is unlikely for SNPs to always abide by the assumptions imputed in the previous sections. In this section, we review three cases where assumptions used in the construction of the IVW estimator are violated, horizontal pleiotropy, weak instruments and linkage disequilibrium. We go over the descriptions and impacts of the violations as well as estimators that mend the impacts. In this section, we focus on two-sample summary data estimators, but also briefly touch on one-sample individual data estimators.

5.1 Horizontal Pleiotropy

Horizontal pleiotropy refers to the direct effect of a SNP on the outcome not through the exposure of interest. Horizontal pleiotropy can be viewed as a violation of exclusion restriction in the instrumental variable context. Applying the IVW estimate directly without adjustments causes bias. In this subsection, we describe some estimators that deal with horizontal pleiotropy.

Zhao et al. (2020) proposed the MR-RAPS estimator that takes horizontal pleiotropy into account. Zhao et al. (2020) considered incorporating the direct effect of SNPs into the linear structural equation model 1 and 2. Let α_j be the direct random effect of SNP j on the outcome. Then the data generating process is assumed to follow equation 9, 10. Figure 5.1 is a graphical illustration of the corresponding data generating process.

$$X = \sum_{j=1}^{p} \gamma_j G_j + \theta_X U + E_X, \tag{9}$$

$$Y = \beta X + \theta_Y U + \sum_{j=1}^{p} \alpha_j G_j + E_Y \tag{10}$$

Zhao et al. (2020) then proposed two models that follows directly from the assumed data generating process. Model 2 and 3.

Model 2 (Systematic pleiotropy)**.** $\alpha_j = \Gamma_j - \beta\gamma_j \sim N(0, \tau_0^2)$ for $j \in [p]$ and some small τ^2.

Model 3 (Systematic and idiosyncratic pleiotropy)**.** Assume α_j, $j \in [p]$ are from a contamiated normal distribution: most α_j are distributed as $N(0, \tau_0^2)$ but some $|\alpha_j|$ may be much larger.

Both models can be obtained by plugging equation 9 into 10 and assuming the direct random effect of SNP j on the outcome, α_j, follow a normal distribution. Note that with the random effect α_j having mean zero, the horizontal pleiotropy is also called balanced horizontal pleiotropy. The difference between model 2 and 3 is that, model 3 allows large pleiotropic effects. Zhao et al. (2020) proposed robustified adjusted profile scores to allow likelihood estimates of parameters under model 2 to be consistent and robust to large α_j values in model 3.

Burgess and Thompson (2017) proposed the MR-Egger estimator to account for horizontal pleiotropy. The MR-Egger regression is motivated from meta analysis. Recall that the IVW estimator is equivalent to running the regression of $\hat{\Gamma}_j$ on $\hat{\gamma}_j$ without intercept 6. The MR-Egger estimator can then be obtained by the regression 11, which is regression 6 including the intercept. Unlike the MR-RAPS that only accounts for balanced pleiotropy, both balanced and directional

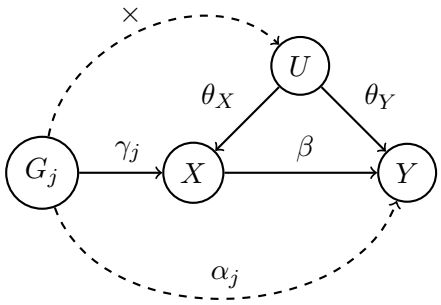

Figure 2: DAG to illustrate horizontal pleiotropy

pleiotropy is considered by the MR-Egger estimator. The intercept μ_α can be interpreted as the average pleiotropic effect of the SNPs included in the analysis. The MR-Egger estimator is consistent when the sample covariance $\text{cov}(\alpha_j, \gamma_j) = 0$ for all j. That is, the pleiotropy term is uncorrelated with the SNP-exposure effect.

$$\hat{\Gamma}_j = \mu_\alpha + \beta\hat{\gamma}_j + \epsilon_j \tag{11}$$

More horizontal-pleiotropy robust estimators are compared by Slob and Burgess (2020). Since the IVW estimator can be viewed as a weighted average of the ratio of $\hat{\Gamma}_j/\hat{\gamma}_j$, for $j = 1, \ldots, p$, as shown in the IVW formula 5. The IVW breaks down if any of the SNPs is an invalid IV. A straightforward method to filter the effect of an invalid SNPs, is to use the median of the sampling distribution of the $\hat{\Gamma}_j/\hat{\gamma}_j$ ratios as estimates, which allows the estimate to be consistent even when 50% of the SNPs are invalid. Bowden et al. (2016) further developed a scheme to adjust the percentile to allow the median estimator to be consistent under weaker conditions than the 50% breakdown point. In a similar fashion, Hartwig et al. (2017) instead consider the mode of the ratios $\hat{\Gamma}_j/\hat{\gamma}_j$ by constructing a sampling distribution of the $\hat{\Gamma}_j/\hat{\gamma}_j$ ratio. Such estimate would be consistent when the largest number of similar (identical in infinite samples) individual-instrument causal effect estimates comes from valid instruments.

5.2 Weak Instruments

The instrumental variable assumption requires instruments G_j's $j = 1, \ldots, p$ to be related to the exposure X. However, do the strength of correlation between the instruments and the exposure matter? Turns out it does. Citing Zhao et al. (2020), the average instrumental strength κ can described by equation 12. The SNPs with larger SNP-exposure effect size $\hat{\gamma}_j$ and smaller standard error $\sigma^2_{X_j}$ has stronger association with the exposure. Zhao et al. (2020) has shown that the IVW estimator can be approximated as a function of the average instrumental strength κ as shown by the formula 13. The bias of the IVW estimator goes up with decreasing instrument strength. Ye et al. (2021b) showed that even by screening IV's with a p-value threshold, the presence of weak iv may not necessarily allow consistency for the inverse variance weighted estimator.

Conventionally, to avoid inclusion of weak instruments, a screening procedure based on a p-value threshold for $\hat{\gamma}_j$ would be used. However, under the two-sample summary data framework, without a third independent dataset to conduct screening, if the SNPs-exposure dataset is used not only

to estimate SNPs-expsoure effect, but also to screen out weak SNPs, it may lead to overestimation of the $\hat{\gamma}_j$, and thus induce bias in the estimate for β. The phenomenon is otherwise known as the winner's curse. (Mounier and Kutalik (2021),Richmond and Smith (2022))

$$\kappa = \frac{1}{p}\sum_{j=1}^{p}\frac{\gamma_j^2}{\sigma_{X_j}^2} \tag{12}$$

$$\hat{\beta}_{IVW} \approx \frac{\beta}{1+1/\kappa} \tag{13}$$

Ye et al. (2021b) proposed the debiased IVW estimator that does not require weak instrument screening and is robust to many weak instruments. The debiased IVW estimator 14 introduces a bias reduction multiple to the IVW estimator. The multiple allows the IVW estimator to achieve consistency under weaker conditions compared to the standard IVW estimator. To be more specific, a less strict requirement is imposed on instrumental strength κ and the number of instruments p to achieve consistency. They also proposed an algorithm to find the screening threshold which increases efficiency of the debiased estimator. Ye et al. (2021b) also made extensions to cases where there are both balanced horizontal pleiotropy and weak instruments.

$$\hat{\beta}_{dIVW} = \hat{\beta}_{IVW}\frac{\sum_{j=1}^{p}\hat{w}_j}{\sum_{j=1}^{p}(\hat{w}_j - \hat{v}_j)} = \frac{\sum_{j=1}^{p}\hat{\Gamma}_j\hat{\gamma}_j\hat{\sigma}_j^{-2}}{\sum_{j=1}^{p}(\hat{\gamma}_j^2 - \hat{\sigma}_{X_j}^2)\hat{\sigma}_{Y_j}^{-2}} \tag{14}$$

where $\hat{w}_j = \hat{\gamma}_j^2/\hat{\sigma}_{Y_j}^2$, $\hat{v}_j = \hat{\sigma}_{X_j}^2/\hat{\sigma}_{Y_j}^2$.

Xu et al. (2021) proposed the pIVW estimator which has smaller bias and variance than debiased IVW under some regularity conditions. Xu et al. (2021) noticed that the debiased IVW 14 estimator may yield extreme estimates, which stems from small values in the denominator. They modelled the denominator and numerator of the dIVW estimator in 14 as a bivariate normal distribution, and constructed a penalized likelihood function to penalize denominators from zero. That is let $\hat{\mu}_1$ be the numerator, $\hat{\mu}_2$ be the denominator of the debiased IVW estimator

$$\hat{\mu}_1 = \sum_{j=1}^{p}\hat{\Gamma}_j\hat{\gamma}_j\hat{\sigma}_j^{-2}$$
$$\hat{\mu}_2 = \sum_{j=1}^{p}(\hat{\gamma}_j^2 - \hat{\sigma}_{X_j}^2)\hat{\sigma}_{Y_j}^{-2}$$

They assumed $(\hat{\mu}_1, \hat{\mu}_2)$ follow a bivariate normal distribution, then the penalized likelihood function can be written as

$$l_p(\mu_1, \mu_2) = \log f(\mu_1, \mu_2) + \lambda \log|\mu_2|$$

where f is the bivariate normal distribution for $(\hat{\mu}_1, \hat{\mu}_2)$. Let $\tilde{\mu}_1$, $\tilde{\mu}_2$ be the estimators obtained by maximizing l_p, then the pIVW estimator

$$\hat{\beta}_{pIVW} = \frac{\tilde{\mu}_1}{\tilde{\mu}_2}$$

Note that the penalized likelihood prevent the μ_2 to be close to zero.

A few methods are proposed to deal with weak instruments in one-sample individual data setting. Ye et al. (2021a) proposed GENIUS-MAWII, where an continuous updating estimator is used for estimation. Liu et al. (2020) proposed the MR MiSTERI estimator. The estimation process starts from a consistent three-stage estimator that can be used as preliminary estimator. The preliminary estimator is then used to constructed one-step-update estimator. Both methods deal with presence of multiple weak instruments in individual data.

5.3 Linkage Disequilibrium

Linkage disequilibrium (LD) is the case where the SNPs are correlated. Wang et al. (2021a) has shown empirically, that the type I error rate would inflate under linkage disequillibrium for the standard IVW estimator.

The convention would be to conduct data clumping or pruning to decorrelate the SNPs. However, after pruning, less valid instruments would be used, and thus increasing the standard error of methods that relies on pruning (Cheng et al. (2020)).

Without pruning, the general strategy to account for linkage disequillibrium (LD) is to incorporate the LD information into the distributional assumptions of the data. In particular, the LD information is captured by the correlation between SNPs, and can be presented as a matrix, called the LD matrix $\hat{\mathbf{R}}$. The element $\hat{\mathbf{R}}_{ij}$ is the correlation between the ith SNP and the jth SNP. Such matrix can also be obtained through GWAs. By notation of Cheng et al. (2020), the likelihood function for two-sample summary statistics data then is assumed to having the model form 15, 16.

$$\hat{\gamma}|\gamma, \hat{\mathbf{R}}, {\sigma_{\mathbf{X}}}^2 \sim N({\sigma_{\mathbf{X}}}^2 \hat{\mathbf{R}} {\sigma_{\mathbf{X}}}^{-1}\gamma, {\sigma_{\mathbf{X}}}^2 \hat{\mathbf{R}} \sigma_{\mathbf{X}}) \tag{15}$$

$$\hat{\mathbf{\Gamma}}|\mathbf{\Gamma}, \hat{\mathbf{R}}, {\sigma_{\mathbf{Y}}}^2 \sim N({\sigma_{\mathbf{Y}}}^2 \hat{\mathbf{R}} {\sigma_{\mathbf{Y}}}^{-1}\mathbf{\Gamma}, {\sigma_{\mathbf{Y}}}^2 \hat{\mathbf{R}} \sigma_{\mathbf{Y}}) \tag{16}$$

where $\sigma_{\mathbf{X}}$ is the diagonal matrix with the standard error σ_{X_j} on it's diagonal. $\sigma_{\mathbf{Y}}$ is the diagonal matrix with the standard error σ_{Y_j} on it's diagonal.

Cheng et al. (2020) proposed the MR-LDP estimator, that is based on the model 15, 16. The type I error rate of MR-LDP is controlled empirically in the presence of correlated instruments. By assuming the true SNPs-Exposure effect $\gamma = [\gamma_1, \ldots, \gamma_p]^T$ and horizontal pleiotropy random effect $\alpha = [\alpha_1, \ldots, \alpha_p]^T$ follow a gaussian distribution. The complete data likelihood can be written as

$$\begin{aligned} &P(\hat{\mathbf{\Gamma}}, \hat{\gamma}, \gamma, \alpha | {\sigma_{\mathbf{X_j}}}^2, {\sigma_{\mathbf{Y_j}}}^2, \hat{\mathbf{R}}; \beta, \sigma_\gamma^2, \sigma_\alpha^2) \\ &= P(\hat{\mathbf{\Gamma}}|\gamma, \alpha, \hat{\mathbf{R}}, {\sigma_{\mathbf{Y_j}}}^2; \beta) P(\hat{\gamma}|\gamma, \hat{\mathbf{R}}, {\sigma_{\mathbf{Y_j}}}^2) P(\alpha|\sigma_\alpha^2) P(\gamma|{\sigma_{\mathbf{X}}}^2) \end{aligned}$$

By EM algorithm, the MR-LDP estimator can be obtained.

Wang et al. (2021a) proposed a Robust Bayesian Mendelian Randomization model (RBMR). Under a similar framework as Cheng et al. (2020), the horizontal pleiotropy effect parameter α is instead modelled to follow a multivariate generalized t-distribution as opposed to a Gaussian distribution to allow estimation to be robust to outliers in horizontal pleiotropy.

Additional methods include MR-Corr by Cheng et al. (2022) to account for correlated horizontal pleiotropy which refers to the case where the SNPs-Exposure effect is correlated with the pleiotropic effect, in other words, $\text{cov}(\alpha_j, \gamma_j) \neq 0$. Wang et al. (2021c) also proposed PLDMR, that uses one-sample individual data, to handle pleiotopy and linkage disequillibrium by using the SNPs matrix $\mathbf{G}$ instead of the LD matrix $\hat{\mathbf{R}}$ to model the complete data likelihood.

6 Simulation study: the two frameworks in MR

In this simulation, we attempt to make a comparison between one-sample individual data methods and two-sample summary statistics methods. We consider four cases, no violation of IV assumptions, balanced pleiotropy, balanced pleiotropy with weak instruments, and directional pleiotropy with weak instruments.

$$X = \sum_{j=1}^{p} \gamma_j G_j + \theta_X U + E_X, \tag{17}$$

$$Y = \beta X + \theta_Y U + \sum_{j=1}^{p} \alpha_j G_j + E_Y \tag{18}$$

The data generating process is described by equation 17 and 18. The four cases differ by the choice of parameters, which is described below. In terms of the parameter common for each case, we choose the causal parameter $\beta = 0.4$, the sample size $n = 1000$, and two candidates for the number of instruments $p \in \{50, 100\}$.

We run a total of 100 replications for each case, for each replication, the jth SNP data for individual i, G_{ij}, are generated randomly with 25% $G_{ij} = 0$, 50% $G_{ij} = 1$, and 25% $G_{ij} = 2$. The independent error terms E_X, E_Y, and the unmeasured confounding U are all drawn from $N(0, 1)$. Whereas the effect of U on X and on Y are respectively $\theta_X = 1$ and $\theta_Y = 2$.

Case 1 $\gamma_1, \ldots, \gamma_p$ are chosen from the interval [0.5,1], which is bounded away from zero.

Case 2 $\gamma_1, \ldots, \gamma_p$ are chosen from the interval [0.5,1]. The balanced pleiotropy $\alpha_1, \ldots, \alpha_p \sim N(0, 1)$.

Case 3 $\gamma_1, \ldots, \gamma_p$ are chosen from N(0,0.01) and fixed for all replications. The balanced pleiotropy $\alpha_1, \ldots, \alpha_p \sim N(0, 1)$.

Case 4 $\gamma_1, \ldots, \gamma_p$ are chosen from N(0,0.01) and fixed for all replications. The balanced pleiotropy $\alpha_1, \ldots, \alpha_p \sim N(1, 1)$.

The estimators that uses individual data are the two stage least squares (tsls) estimator with and without intercept, the limited information maximum likelihood (liml) estimator with and without intercept. The MR-GENIUS estimator by Ye et al. (2021a). We did not consider the MR-MISTERI estimator by Liu et al. (2020) given the computational speed corresponding to the number of instruments we used in the simulation.

The estimators that uses two-sample summary statistics data are MR-RAPS by Zhao et al. (2020), MR-Egger by Bowden et al. (2016), the standard IVW estimator, the MR-Median estimator and the debiased IVW estimator by Ye et al. (2021b). Note that in this simulation, we applied the two sample summary statistics methods directly on one sample summary data obtained from data generating process 17 and 18. Minelli et al. (2021) has shown based on simulation, that the estimates should not be affected much.

To measure performance, for each case, we calculate the mean of the causal estimates over 100 replications, the bias proportion, and the Monte Carlo standard error. The bias proportion is defined as

$$Bias = \frac{\bar{\hat{\beta}} - \beta}{\beta}$$

where $\bar{\hat{\beta}}$ is the mean of estimates over replications. The results are shown in table 2 and 3.

By table 2. The majority of the estimators did well when there is no violation, with only the MR-GENIUS estimator being more biased. When balance pleiotropy is present, most methods have small bias proportion since the pleiotropy is balanced. With the number of SNPs increasing, the two sample summary statistics method performance improves and often performs slightly better in terms of bias proportion compared to individual data methods.

By table 3. When balanced horizontal pleiotropy and weak instruments are both present, the LIML estimator performs poorly whereas the TSLS estimator with intercept and the MR-GENIUS estimator performed quite well, though they still don't quite match the performance of most two-sample summary statistics methods. Interestingly, the debiased IVW estimator did not perform well, presumably due to the average instrument strength being too weak and the number of SNPs too small, such that consistency cannot be achieved. And even though debiased IVW is less biased when number of instruments increase, the standard error is large, reflecting the unstable property of dIVW mentioned in Xu et al. (2021). When directional pleiotropy and weak instruments are both present, the one sample individual data methods stood out, especially the MR-GENIUS estimator and the TSLS estimator with intercept, in particular, the MR-GENIUS performs better when the number of instruments are small, which is consistent with the requirements for consistency derived in Ye et al. (2021a). One explanation for bad performance for two-sample summary statistics estimators is that Barry et al. (2021) has shown that when weak instruments are present, applying two-sample summary data methods to one-sample individual data adds additional bias, so this may partially explain the relative large bias which occurred in case 3 and 4.

7 Discussion

In this paper, we surveyed MR methods, especially ones related to violations of core assumptions, in particular, horizontal pleiotropy, weak instruments and linkage disequillibrium. We also did simulations to compare methods used in the two frameworks of MR, the one-sample individual data framework and the two-sample summary data framework. It turned out that when limited instruments are available, or when instrument strength is too weak for consistency of two-sample summary data estimators to be achieved, the two-staged least squares estimator with intercept and the MR-GENIUS estimator can be used to obtain causal effect.

More work can be done in the future simulationwise and comparisonwise. In terms of simulation, we did not create parameters based on the average instrumental strength. Instead, we bound the SNPs-exposure effect $\hat{\gamma}_j$'s away from zero to avoid weak instruments in case 1 and case 2; similarly, we allow the $\hat{\gamma}_j$'s to assume values close to zero to introduce weak instruments for case 3 and 4. Additonally, when weak instruments are present, we did not do instrumental strength screening on the SNPs, then compare the performance of estimators after SNPs screening. Also, we assumed that the IV "independence" assumption holds for all cases. Inspection of the estimators' sensitivity to violation of independence is also a direction to explore in the future.

In terms of comparison, more metrics can be considered, including coverage probability and the empirical power for each estimators. Furthermore, we can only obtain one-sample data from the individual data to make a comparison between frameworks. Which violates the assumption in the two-sample summary statistics framework that $\hat{\Gamma}_j \perp\!\!\!\perp \hat{\gamma}_j$. The way to conduct fairer comparisons between the one-sample individual data and the two-sample summary data framework is also a direction to work on.

				No violation			
			p = 50			p = 100	
	Method	mean	bias	se	mean	bias	se
case 1	tsls	0.4000	0.0000	0.0000	0.4000	0.0001	0.0000
	liml	0.3999	-0.0002	0.0000	0.3999	-0.0002	0.0000
	tsls_intercept	0.4045	0.0113	0.0001	0.4056	0.0141	0.0001
	liml_intercept	0.3984	-0.0041	0.0001	0.3989	-0.0026	0.0001
	mr.genius	0.5229	0.3073	4.0057	0.3523	0.1193	3.7610
	mr.raps	0.3999	-0.0002	0.0000	0.3999	-0.0002	0.0000
	mr.egger	0.4038	0.0096	0.0001	0.4005	0.0012	0.0000
	ivw	0.3999	-0.0002	0.0000	0.3999	-0.0002	0.0000
	mr.median	0.4000	0.0001	0.0000	0.3999	-0.0002	0.0000
	divw	0.4001	0.0003	0.0000	0.4001	0.0003	0.0000
				Balanced pleiotropy			
			p = 50			p = 100	
	Method	mean	bias	se	mean	bias	se
case 2	tsls	0.39900	-0.00251	0.00005	0.39974	-0.00064	0.00003
	liml	0.39893	-0.00267	0.00005	0.39971	-0.00072	0.00003
	tsls_intercept	0.39744	-0.00640	0.00503	0.41013	0.02533	0.00545
	liml_intercept	0.39118	-0.02205	0.00523	0.40428	0.01069	0.00565
	mr.genius	0.59624	0.49061	9.13922	0.32710	-0.18225	16.24030
	mr.raps	0.39905	-0.00237	0.00005	0.39977	-0.00057	0.00003
	mr.egger	0.39787	-0.00531	0.00292	0.39891	-0.00273	0.00102
	ivw	0.39907	-0.00231	0.00005	0.39980	-0.00051	0.00003
	mr.median	0.39898	-0.00256	0.00005	0.39981	-0.00049	0.00003
	divw	0.39927	-0.00182	0.00005	0.39999	-0.00002	0.00003

Table 2: Simulation results for case 1 and case 2

		Balanced pleiotropy + Weak Instrument					
		p = 50			p = 100		
	Method	mean	bias	se	mean	bias	se
	mean	bias	se	mean	bias	se	
case 3	tsls	0.8792	1.1979	0.3421	0.8313	1.0783	0.2696
	liml	6.1492	14.3730	2220.5316	30.6904	75.7261	68391.3542
	tsls_intercept	0.8810	1.2025	0.3148	0.8547	1.1367	0.2913
	liml_intercept	46.0262	114.0654	89859.5902	-0.9669	3.4174	2931.6954
	mr.genius	1.0684	1.6710	12.1140	1.0479	1.6197	16.2546
	mr.raps	-6.9571	-18.3928	30451.5493	0.6407	0.6017	11.5322
	mr.egger	0.7434	0.8584	2.0616	0.9033	1.2582	1.0918
	ivw	0.6164	0.5411	4.9313	0.3555	-0.1112	4.9282
	mr.median	0.5969	0.4923	5.1578	0.4112	0.0280	4.4779
	divw	8.5252	20.3130	3270.5995	0.2972	-0.2569	47.6319
		Directional pleiotropy + weak instruments					
		p = 50			p = 100		
	Method	mean	bias	se	mean	bias	se
case 4	tsls	-12.3552	-31.8881	139.8651	15.3794	37.4484	187.4155
	liml	-788.4446	-1972.1115	24016185.1605	-11344.7657	-28362.9143	12644576141
	tsls_intercept	0.6683	0.6709	2.1553	0.9514	1.3786	1.0669
	liml_intercept	-13.2891	-34.2228	405602.5696	77.8299	-193.5748	389651.1256
	mr.genius	0.2727	-0.3181	9.4918	0.6184	0.5461	13.6991
	mr.raps	-1280.3995	-3201.9989	90046192.4292	-4549.8109	-11375.5271	1206249730.9738
	mr.egger	-101.0927	-253.7318	56301.0594	180.0228	449.0569	185827.9766
	ivw	-282.7847	-707.9618	93355.9632	546.9756	1366.4391	302801.6535
	mr.median	-258.7733	-647.9333	107861.1135	491.5477	1227.8692	380457.6784
	divw	360.9491	901.3729	40829121.4890	1153.7324	2883.3310	34303750.7985

Table 3: Simulation results for case 3 and case 4

References

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753.

Barry, C., Liu, J., Richmond, R., Rutter, M. K., Lawlor, D. A., Dudbridge, F., and Bowden, J. (2021). Exploiting collider bias to apply two-sample summary data mendelian randomization methods to one-sample individual level data. *PLoS genetics*, 17(8):e1009703.

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. *Statistics in medicine*, 36(11):1783–1802.

Bowden, J. and Holmes, M. V. (2019). Meta-analysis and mendelian randomization: A review. *Research synthesis methods*, 10(4):486–496.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665.

Burgess, S. and Thompson, S. G. (2017). Interpreting findings from mendelian randomization using the mr-egger method. *European journal of epidemiology*, 32(5):377–389.

Cheng, Q., Qiu, T., Chai, X., Sun, B., Xia, Y., Shi, X., and Liu, J. (2022). Mr-corr2: a two-sample mendelian randomization method that accounts for correlated horizontal pleiotropy using correlated instrumental variants. *Bioinformatics*, 38(2):303–310.

Cheng, Q., Yang, Y., Shi, X., Yeung, K.-F., Yang, C., Peng, H., and Liu, J. (2020). Mr-ldp: a two-sample mendelian randomization for gwas summary statistics accounting for linkage disequilibrium and horizontal pleiotropy. *NAR genomics and bioinformatics*, 2(2):lqaa028.

Davey Smith, G. and Ebrahim, S. (2003). 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22.

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian randomization. *Jama*, 318(19):1925–1926.

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology*, 46(6):1985–1998.

Leong, A., Cole, J. B., Brenner, L. N., Meigs, J. B., Florez, J. C., and Mercader, J. M. (2021). Cardiometabolic risk factors for covid-19 susceptibility and severity: a mendelian randomization analysis. *PLoS medicine*, 18(3):e1003553.

Liu, X., Tong, X., Zou, Y., Lin, X., Zhao, H., Tian, L., Jie, Z., Wang, Q., Zhang, Z., Lu, H., et al. (2022). Mendelian randomization analyses support causal relationships between blood metabolites and the gut microbiome. *Nature genetics*, pages 1–10.

Liu, Z., Ye, T., Sun, B., Schooling, M., and Tchetgen, E. T. (2020). On mendelian randomization mixed-scale treatment effect robust identification (mr misteri) and estimation for causal inference. *arXiv preprint arXiv:2009.14484*.

Manousaki, D., Harroud, A., Mitchell, R. E., Ross, S., Forgetta, V., Timpson, N. J., Smith, G. D., Polychronakos, C., and Richards, J. B. (2021). Vitamin d levels and risk of type 1 diabetes: A mendelian randomization study. *PLoS medicine*, 18(2):e1003536.

Minelli, C., Del Greco M, F., van der Plaat, D. A., Bowden, J., Sheehan, N. A., and Thompson, J. (2021). The use of two-sample methods for mendelian randomization analyses on single large datasets. *International journal of epidemiology*, 50(5):1651–1659.

Mounier, N. and Kutalik, Z. (2021). Correction for sample overlap, winner's curse and weak instrument bias in two-sample mendelian randomization. *bioRxiv*.

Newey, W. (2007). course materials for 14.385 nonlinear econometric analysis.

Pearl, J. et al. (2000). Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19:2.

Richmond, R. C. and Smith, G. D. (2022). Mendelian randomization: concepts and scope. *Cold Spring Harbor perspectives in medicine*, 12(1):a040501.

Sanderson, E., Glymour, M. M., Holmes, M. V., Kang, H., Morrison, J., Munafò, M. R., Palmer, T., Schooling, C. M., Wallace, C., Zhao, Q., et al. (2022). Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):1–21.

Slob, E. A. and Burgess, S. (2020). A comparison of robust mendelian randomization methods using summary data. *Genetic epidemiology*, 44(4):313–329.

Wang, A., Liu, W., and Liu, Z. (2021a). A two-sample robust bayesian mendelian randomization method accounting for linkage disequilibrium and idiosyncratic pleiotropy with applications to the covid-19 outcomes. *Genetic Epidemiology*.

Wang, X., Fang, X., Zheng, W., Zhou, J., Song, Z., Xu, M., Min, J., and Wang, F. (2021b). Genetic support of a causal relationship between iron status and type 2 diabetes: a mendelian randomization study. *The Journal of Clinical Endocrinology & Metabolism*, 106(11):e4641–e4651.

Wang, Y., Li, T., Fu, L., Yang, S., and Hu, Y.-Q. (2021c). A novel method for mendelian randomization analyses with pleiotropy and linkage disequilibrium in genetic variants from individual data. *Frontiers in Genetics*, 12.

Xu, S., Fung, W. K., and Liu, Z. (2021). pivw: A novel mendelian randomization method accounting for weak instruments and horizontal pleiotropy with applications to the covid-19 outcomes. *medRxiv*.

Ye, T., Liu, Z., Sun, B., and Tchetgen, E. T. (2021a). Genius-mawii: For robust mendelian randomization with many weak invalid instruments. *arXiv preprint arXiv:2107.06238*.

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

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

Zheng, J., Baird, D., Borges, M.-C., Bowden, J., Hemani, G., Haycock, P., Evans, D. M., and Smith, G. D. (2017). Recent developments in mendelian randomization studies. *Current epidemiology reports*, 4(4):330–345.