Constrained Instruments and their Application to Mendelian Randomization with Pleiotropy

Lai Jiang[1,2], Karim Oualkacha[3], Vanessa Didelez[4], Antonio Ciampi[1,2], Pedro Rosa[5,6], Andrea L. Benedet[6], Sulantha Mathotaarachchi [6], J. Brent Richards[7], and Celia M.T. Greenwood[1,2]

[1]*Lady Davis Institute, Jewish General Hospital, Montreal*
[2]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal*
[3]*Université du Québec À Montréal, Montréal*
[4] *Leibinz Institute for Prevention Research and Epidemiology - BIPS & Department of Mathematics, University of Bremen*
[5]*Department of Neurology & Neurosurgery, McGill University*
[6]*Translational Neuroimaging Laboratory, McGill University Research Centre for Studies in Aging, Douglas Hospital, McGill University*
[7]*Department of Medicine, McGill University, Montreal*

November 30, 2017

Abstract

In Mendelian randomization (MR), genetic variants are used to construct instrumental variables, which enable inference about the causal relationship between a phenotype of interest and a response or disease outcome. However, standard MR inference requires several assumptions, including the assumption that the genetic variants only influence the response through the phenotype of interest. Pleiotropy occurs when a genetic variant has an effect on more than one phenotype; therefore, a pleiotropic genetic variant may be an invalid instrumental variable. Hence, a naive method for constructing instrumental variables may lead to biased estimation of the causality between the phenotype and the response. Here, we present a set of intuitive methods (Constrained Instrumental Variable methods [*CIV*]) to construct valid instrumental variables and perform adjusted causal effect estimation when pleiotropy exists, focusing particularly on the situation where pleiotropic phenotypes have been measured. Our approach includes an automatic and valid selection of genetic variants when building the instrumental variables. We also provide details of the features of many existing methods, together with a comparison of their performance in a large series of simulations. *CIV*

methods performed consistently better than many comparators across four different pleiotropic violations of the MR assumptions. We analyzed data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) Mueller et al. (2005) to disentangle causal relationships of several biomarkers with AD progression. The results showed that *CIV* methods can provide causal effect estimates, as well as selection of valid instruments while accounting for pleiotropy.

1 Introduction

1.1 Mendelian Randomization

Mendelian randomization is a method for estimating the causal effect of a modifiable exposure ($\mathbf{X}$) on a disease ($\mathbf{Y}$) by using measured genetic variation ($\mathbf{G}$) as an instrument to eliminate bias from unmeasured confounding factors ($\mathbf{U}$). The Mendelian inheritance patterns of genetic data, from parents to children, can be viewed as comparable to a randomized controlled trial. The reason is, if the choice of mate is not associated with genotype, the genotypes distribution (of the offspring) should be unrelated to any confounding factors. From a statistical perspective, Mendelian randomization is an application of instrumental variable methods using genetic information, as instruments (Smith and Ebrahim, 2004; Didelez and Sheehan, 2007; Lawlor et al., 2008; Wehby et al., 2008) for the exposure of interest $\mathbf{X}$, as illustrated in Figure 1. In order to obtain valid results from instrumental variable analysis, several important assumptions about the relationships between the genotype instruments, $\mathbf{G}$ (usually single nucleotide polymorphisms (SNPs)), and the other variables must hold. When working with a structural equation modeling (SEM) set-up, the assumptions for Mendelian randomization are:

(A1) $\mathbf{G}$ is associated with the exposure $\mathbf{X}$ (i.e. $\mathbf{G} \not\perp \mathbf{X}$, or $\mathbf{G}$ and $\mathbf{X}$ are not independent).

(A2) $\mathbf{G}$ and $\mathbf{Y}$ are independent conditional on exposure $\mathbf{X}$ and unmeasured confounding factors $\mathbf{U}$ (i.e. $\mathbf{G} \perp \mathbf{Y}|\mathbf{X}, \mathbf{U}$).

(A3) $\mathbf{G}$ and confounders $\mathbf{U}$ must be independent (i.e. $\mathbf{G} \perp \mathbf{U}$).

If linear models are assumed for the dependencies among the $\mathbf{G}, \mathbf{X}, \mathbf{Y}$, then "independent" in the assumption can be relaxed to "uncorrelated", and "associated" can be replaced with "correlated". In the following we assume linear relationship for causal dependencies.

These assumptions may be violated in some contexts (Didelez and Sheehan, 2007; Lawlor et al., 2008). For example, linkage disequilibrium, which refers to the association of alleles at different loci in the population, may lead to violations of condition (A2). If the genetic variant of interest, G_1, is in linkage disequilibrium with another genetic variant, G_2, which has a direct or indirect influence on the disease $\mathbf{Y}$, then (A2) is not satisfied for G_1. It is often believed that genotypes will not be associated with the socioeconomic and behavioral characteristics that confound $\mathbf{X}$ and $\mathbf{Y}$, however, careful assessment of possible violations of

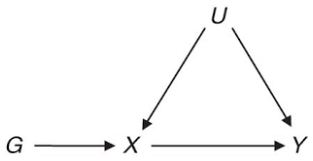


Figure 1: A directed acyclic graph (DAG) representing a situation where Mendelian randomization using genetic variants **G** as instruments can be useful for inferring whether a phenotype **X** is causally related to an outcome **Y**. **U** represents unmeasured confounding factors.

(A3) is still necessary in some situations (Lawlor et al., 2008). Often most the problematic assumption is (A2), since genetic variants often have effects on cell functioning that are not well understood and could plausibly act through many mechanisms.

1.2 Challenges Arising from Pleiotropy

Pleiotropy–when more than one phenotype is influenced by the same group of genotypes–may violate assumption (A2) if all these phenotypes are themselves on the causal pathway for the response **Y**. Two kinds of pleiotropy can be defined (Solovieff et al., 2013): biological pleiotropy (Stearns, 2010; Wagner and Zhang, 2011) refers to associations involving multiple phenotypes sharing the common genetic pathways. For example, a gene variant in *PTPN22* is known to be associated with immune related disorders such as Type 1 diabetes (Todd et al., 2007) and Crohn's disease (Barrett et al., 2008). This variant has been shown to interfere with the function of various T cells (Rieck et al., 2007) and affect the removal of autoreactive B cells (Menard et al., 2011). Hence, the impact of T cell levels on the risk of diseases will be confounded by the alternative causal pathway through reactive B cells, and vice versa. In contrast, mediated pleiotropy refers to direct causal impacts between phenotypes, such that the genotypes of interest have direct/indirect causal impact on both phenotypes. For example Thorgeirsson et al. (2008) reported a common variant in the nicotinic acetylcholine receptor gene cluster that affects both nicotine dependence (ND) and the smoking quantity (SQ). Both of these phenotypes (ND and SQ) are associated with the risk of lung cancer (Hung et al., 2008; Lamin et al., 2014) and yet smoking quantity can affect nicotine dependence. On the other hand, it is also believed that the most important factor for smoking persistence is nicotine dependence. The relationship between ND and SQ (Figure 2) depends on the duration of smoking and magnitude of craving (Donny et al., 2008). Therefore, in order to estimate the magnitude of the effect of nicotine dependence alone on lung cancer, one would need to appropriately account for the effect of smoking quantity. Methods to clarify such complex relationship are essential.

The motivation for this work is derived from two straightforward questions:

(B1) In many applications, researchers may have access to large data collections including many phenotypes. Although some are of primary interest for their causal effects, others could be considered of secondary interest – yet may be influenced by some of

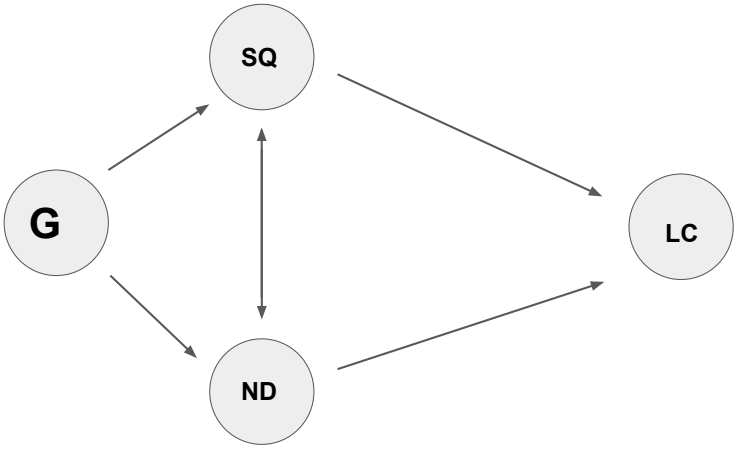


Figure 2: Diagram representing pleiotropy in smoking studies. **G**: genotypes. ND: nicotine dependence. SQ: smoking quantity. LC: lung cancer (risk). A bidirectional arrow between SQ and ND reflects multiple dependencies. For simplicity we omit possible confounding factors here.

the same genetic variants. What is the best way to use data from these additional phenotypes and perform inference on causal effects?

(B2) Does the solution for (B1) select genetic variants that *only* affect the phenotype of primary interest?

Many different statistical methods have been proposed to address challenges arising from pleiotropy. In section 2 we introduce notation and review several of the popular methods. Then in section 3, we propose a novel instrumental variable approach for Mendelian randomization in the presence of potential pleiotropic phenotypes. First, we describe our instruments, which are based on a weighted combination of the original genetic information and maximize the association between **G** and **X** under a set of constraints. We then show that approximate sparse solutions for these weights can be obtained, and furthermore that we can select approximately valid instruments as a result. In section 4, we compare all these methods against each other in simulated data, and in section 5 we analyze an Alzheimer's disease dataset ADNI to demonstrate the performance of our *CIV* method as well as other instrumental variable methods.

2 Context of the Problem and Review of Existing Methods

For each individual $i \in \{1, ..., n\}$, let Y_i be the response of interest and let $\mathbf{G_i} \in \mathbb{R}^{\mathbf{p}}$ represents the genotypes that have been collected for ith observation, where p is the number of all genotypes available. Let $\mathbf{X_i} \in \mathbb{R}, \mathbf{Z_i} \in \mathbb{R}^k$ be the phenotypes that have been measured for this individual; $\mathbf{X_i}$ is the phenotype of interest and $\mathbf{Z_i}$ are phenotypes that may be affected by

elements of $\mathbf{G_i}$. We denote $\mathbf{Y} = (Y_1, .., Y_n)^\top, \mathbf{X} = (\mathbf{X_1}, ..., \mathbf{X_n})^\top \in \mathbb{R}^{n\times 1}, \mathbf{Z} = (\mathbf{Z_1}, ..., \mathbf{Z_n})^\top \in \mathbb{R}^{n\times k}$ and $\mathbf{G} = (\mathbf{G_1}, ..., \mathbf{G_n})^\top \in \mathbb{R}^{n\times p}$ to be n-dimensional vectors while each individual is assumed to be observed in an i.i.d. fashion. We assume linear structural equation models:

$$\mathbf{Z} = \mathbf{G}\alpha_z + \zeta_\mathbf{z}\mathbf{U} + \epsilon_\mathbf{z} \tag{1a}$$
$$\mathbf{X} = \mathbf{G}\alpha_x + \mathbf{Z}\gamma_{zx} + \zeta_\mathbf{x}\mathbf{U} + \epsilon_\mathbf{x} \tag{1b}$$
$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\eta + \zeta_\mathbf{y}\mathbf{U} + \epsilon_\mathbf{y} \tag{1c}$$

Or

$$\mathbf{X} = \mathbf{G}\alpha_x + \zeta_\mathbf{x}\mathbf{U} + \epsilon_\mathbf{x} \tag{2a}$$
$$\mathbf{Z} = \mathbf{G}\alpha_z + \mathbf{X}\gamma_{xz} + \zeta_\mathbf{z}\mathbf{U} + \epsilon_\mathbf{z} \tag{2b}$$
$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\eta + \zeta_\mathbf{y}\mathbf{U} + \epsilon_\mathbf{y} \tag{2c}$$

where α_x and α_z are the association parameters between genotypes $\mathbf{G}$ and phenotypes $\mathbf{X}, \mathbf{Z}$. β is the causal effect parameter of interest, and η is the pleiotropic causal effect of $\mathbf{Z}$ on $\mathbf{Y}$. $\zeta_\mathbf{x}, \zeta_\mathbf{z}, \zeta_\mathbf{y}$ represent the impact (coefficient) of unmeasured confounding factors $\mathbf{U}$ on $\mathbf{X}, \mathbf{Z}$ and $\mathbf{Y}$ respectively. We use γ_{zx} and γ_{xz} to denote the direct causal impact of $\mathbf{Z}$ on $\mathbf{X}$ and $\mathbf{X}$ on $\mathbf{Z}$ respectively. Note that at least one of γ_{zx} and γ_{xz} should be 0. Let $\epsilon_\mathbf{x}, \epsilon_\mathbf{z}, \epsilon_\mathbf{y}$ be independent errors for $\mathbf{X}, \mathbf{Z}$ and $\mathbf{Y}$ respectively.

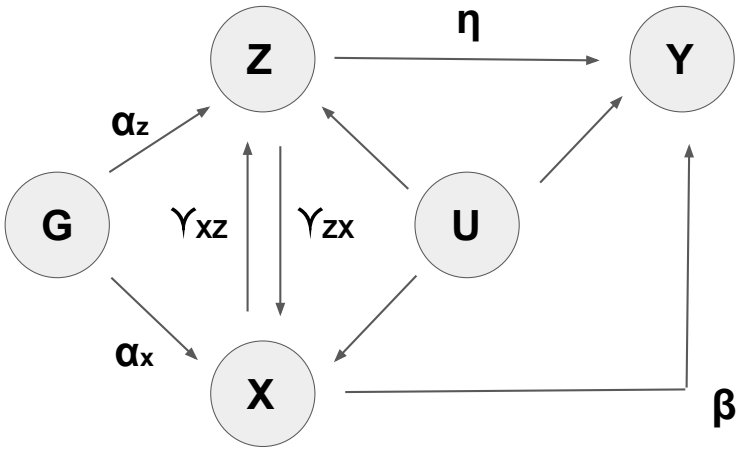


Figure 3: General diagram representing pleiotropic influences in Mendelian randomization studies. $\mathbf{G}$: genotypes. $\mathbf{X}$: phenotype of interest. $\mathbf{Z}$: pleiotropic phenotypes. $\mathbf{Y}$: response of interest. α_x, α_z are the genetic association parameters between $\mathbf{X} \sim \mathbf{G}$ and $\mathbf{Z} \sim \mathbf{G}$. β is the causal effect of interest ($\mathbf{X}$ on $\mathbf{Y}$). η: the pleiotropic pathways of $\mathbf{Z}$ on $\mathbf{Y}$. γ_{xz} and γ_{zx} represent the possible causal effects of $\mathbf{X}$ on $\mathbf{Z}$ and $\mathbf{Z}$ on $\mathbf{X}$ respectively.

Figure 3 lays out the general pleiotropic structure of this model. We assume that genotypes $\mathbf{G}$, phenotypes of interest $\mathbf{X}$, the response $\mathbf{Y}$ and potential pleiotropic phenotypes

Z have all been measured for each individual in the study. The parameter of interest is assumed to be β, the causal effect of **X** on **Y**. Unobserved confounders are indicated by **U**. The relationships between phenotypes of interest **X**, pleiotropic phenotypes **Z**, and outcomes **Y** may vary from one situation to another, i.e. not all of the edges or arrows in Figure 3 need to be present in any particular example. However, if **X** has a direct causal relationship with **Y**, then β must be nonzero.

Three different types of relationship between **X** and **Z** can be assumed:

i **X** and **Z** are conditionally independent given **G** and **U** ($\gamma_{zx} = \gamma_{xz} = 0$).

ii direct causal impact of **X** on **Z** ($\gamma_{xz} \neq 0$ and $\gamma_{zx} = 0$).

iii direct causal impact of **Z** on **X** ($\gamma_{zx} \neq 0$ and $\gamma_{xz} = 0$).

It is worth noting that only (i) and (iii) can be referred to as pleiotropy (by definition) when $\alpha_z \neq 0$. For (ii) even valid instruments (with $\alpha_z = 0$) are still associated with **Z** due to the path $\mathbf{G} \rightarrow \mathbf{X} \rightarrow \mathbf{Z}$ and the total causal effect for **X** would be $\beta + \gamma_{xz}\eta$.

The term "endogenous" variable describes the factors that are explained by the genotype-phenotype relationships and have impact on response **Y**. Common endogenous variables include health-related behaviors and risk-related phenotypes. **X** and **Z** in Figure 3 are both endogenous since they are both determined by genotypes and have impact on response, albeit with different functions. Variables such as age and sex that are not associated with the genotype-phenotype causal pathways of interest are termed "exogenous" variables; normally it is possible to adjust for these variables in a straightforward way in data analysis.

2.1 One Sample and Two Sample Mendelian Randomization

MR can be conducted with one sample of subjects, where individual level data (**G** ,**X**, **Y**) or summary statistics (**G**-**X** associations and **G**-**Y** associations) are used to infer the causal effect $\mathbf{X} \rightarrow \mathbf{Y}$, all in the same data set. An alternative strategy is to use two-sample MR methods with summary statistics, when no joint data is available. The gene-exposure (**G**-**X**) and gene-outcome (**G**-**Y**) associations are taken from different data sources for two-sample MR with summary statistics. Some MR methods for individual level data can also be separated into two steps and used with two-sample data. However, not all MR methods are straightforwardly adapted for two-sample set-ups.

There are three main reasons for using two-sample MR: the first is the separation of **G**-**X** and **G**-**Y** associations in two datasets. In this case only two-sample MR is applicable. Moreover, even if the estimated genetic instrument's effect on **X** is biased in the first data set, this bias should not affect the causal effect estimation obtained in the second data set (Lawlor, 2016). The second reason is to alleviate weak instrument bias, in which instruments are weakly associated with phenotype of interest. As a result, in one-sample analysis, the (unobserved) confounders may explain more variation in the phenotype than the instruments, and the estimates will be biased towards the observational confounded association $\frac{\zeta_y}{\zeta_x}$

(Burgess et al., 2011). In two sample MR analysis the over-fitting of $\mathbf{X}$ is avoided, and then the weak instrument bias is towards the null (Davey Smith and Hemani, 2014). The third reason for using two-sample MR is the possibility to improve the statistical power of causal effect estimates and reduce $\mathbf{G} \rightarrow \mathbf{X}$ bias by using large data collections with many cases of samples.

2.2 Approaches based on 2SLS

If possible, a simple solution to address pleiotropy is to select for analysis only the valid genotypes, i.e. those that influence the phenotype of interest $\mathbf{X}$ and not $\mathbf{Z}$, and use them as instruments with instrumental variable methods such as such as two-stage least squares (2SLS) regression. 2SLS method is a popular technique that is used in the analysis of structural equations. Given valid instruments $\mathbf{G}$, the following two steps define a 2SLS model:

1. In the first stage, we obtain a new variable $\mathbf{X}^*$ as fitted value from ordinary least square regression $\mathbf{X} \sim \mathbf{G}$, where $\mathbf{G}$ are the selected instruments.

2. In the second stage, we substitute $\mathbf{X}$ with $\mathbf{X}^*$ and obtain ordinary least square estimators of β from the regression $\mathbf{Y} \sim \mathbf{X}^*$.

However, this selection of valid instruments may not always be possible without in-depth knowledge of the disease under study. For example, Timpson et al. (2005) uses common *CRP* (C-reactive protein) gene haplotypes, based on 3 SNPs, as instruments to infer the causal effect of the CRP protein (i.e. the phenotype of interest $\mathbf{X}$) on multiple metabolic syndrome phenotypes including body-mass index (BMI) and high density lipoprotein (HDL) levels (i.e. $\mathbf{Y}$. However, Martínez-Calleja et al. (2012) showed that one of these SNPs (rs1130864) is directly related to BMI, and it is also known that there exists a negative association between higher levels of BMI and HDL (Shamai et al., 2011). Hence, these 3 SNPs may not all be valid instruments for causal effect estimation of CRP protein levels on HDL. In many situations, our understanding of SNP effects is not complete enough to select valid instruments based on knowledge.

Another simple solution is to replace $\mathbf{G}$ with residuals after regressing each of these genetic variants on the pleiotropic phenotypes, $\mathbf{Z}$. Specifically, one can replace $\mathbf{G}$ with $\mathbf{G}^* = (\mathbf{I} - \mathbf{P_z})\mathbf{G}$ where $\mathbf{P_z} = \mathbf{Z}^\top(\mathbf{Z}^\top\mathbf{Z})^{-\mathbf{1}}\mathbf{Z}$, which regresses out the exogenous effect of $\mathbf{Z}$. Then two-stage least squares (2SLS) or other instrumental methods can be applied to the new $\mathbf{G}^*, \mathbf{X}, \mathbf{Y}$. We refer to this method as "2SLS_adj". In general, this is not an appropriate solution and it will lead to biased estimation of β since $\mathbf{Z}$ is not an exogenous variable.

A related solution based on the underlying linear structural equation model (Equation 1) turns to multiple linear regression of $\mathbf{Y}$ on $\hat{\mathbf{X}}$ and $\hat{\mathbf{Z}}$ jointly in a two-stage least squares (2SLS) model, where $\hat{\mathbf{X}}$ and $\hat{\mathbf{Z}}$ are the predicted phenotypes using $\mathbf{G}$ as the instruments. We refer to this method as "*2SLS_mul*" since it essentially implements multiple 2SLS regression. The *2SLS_mul* method uses $\mathbf{G}$ to account for the endogeneity of $\mathbf{Z}$ without controlling for it

explicitly. However, using this approach, the resulting estimator of β will be unstable if $\hat{\mathbf{X}}$ and $\hat{\mathbf{Z}}$ are highly correlated (Farrar and Glauber, 1967; Graham, 2003; Grewal et al., 2004).

By nature *2SLS_adj* and *2SLS_mul* methods require individual level data in a single sample. In the two-sample context, when $(\mathbf{G}, \mathbf{X})$ and $(\mathbf{G}, \mathbf{Y})$ are observed separately, the estimators are still available for 2SLS under certain conditions and assumptions (Angrist and Krueger, 1992; Dee and Evans, 2003). In that case, the corresponding estimators are called "two-sample instrumental variable" (TSIV) estimators in empirical studies (Borjas, 2004; Dee and Evans, 2003). In this paper, we only refer to *2SLS* methods as the estimators for individual level data.

2.3 Distribution of Causal Effects Across Multiple Instruments

When there are multiple instruments but not all are valid, causal conclusions can still be drawn by examining the distribution of the estimated causal effects across the instruments. If a group of genotype instruments, such as SNPs, are independent–i.e. located at different sites in the genome–and they all lead to similar estimates of the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ (i.e. β), then this pattern provides strong evidence of a causal relationship between $\mathbf{X}$ and $\mathbf{Y}$. This phenomenon is nicely illustrated with a funnel plot, where the precision of $\hat{\beta}$ (defined as the reciprocal of its variance) for each SNP is plotted against its estimate, $\hat{\beta}$. Asymmetry in a funnel plot might indicate an unbalanced pleiotropy, where variants subject to pleiotropy tend to bias the causal estimate in the same direction. There are tests for symmetry such as Beggs rank correlation test, although they have low statistical power (Begg and Mazumdar, 1994).

Mendelian randomization with *Egger* regression (MR-*Egger*) provides a more specific way to assess whether pleiotropy is present and to obtain an "unbiased" estimate. *Egger* regression is defined as the linear regression of a normalized parameter estimate against its precision (reciprocal of the corresponding standard error) in meta-analysis. *Egger*'s test assesses small study bias by testing the hypothesis that intercept of the *Egger* regression is zero. (Bowden et al., 2015) suggested that bias due to pleiotropy can be considered as analogous to small sample bias (Egger et al., 1997), and therefore that meta-analysis methods could be of use in Mendelian randomization settings. Mendelian randomization with *Egger* regression (MR-*Egger*) uses the slope coefficient from *Egger* regression as a consistent causal effect estimator of β. A key assumption (the InSIDE assumption) here is that the associations of the genetic variants with the phenotypes of interest are independent of the direct effects of the genetic variants on the response – i.e., α_x is independent of α_z for any SNPs in $\mathbf{G}$ (Figure 3). The InSIDE assumption still holds even if some instruments are invalid ($\alpha_z \neq 0$), but it will be violated if genotypes $\mathbf{G}$ influence pleiotropic phenotypes (or confounders) $\mathbf{Z}$ that have impact on both $\mathbf{X}$ and $\mathbf{Y}$ – i.e., $\alpha_x = 0, \alpha_z \neq 0, \gamma_{zx} \neq 0$. MR-*Egger* regression provides no analytical estimate of the standard error of the $\hat{\beta}$, although a confidence interval can be obtained using bootstrap methods. It is also worth noting that MR-*Egger* can be extended to two-sample summary data MR analysis. However, for individual level data analysis, MR-*Egger* is restricted to a single sample because it considers risk factors one at a time. If two distinct samples of $\mathbf{G}, \mathbf{X}, \mathbf{Y}$ are available, MR-*Egger* can be applied separately

to both datasets and evaluates the difference between the two samples.

2.4 Summaries of Multiple Instruments

The genetic information across a set of genotypes can be summarized into a single genetic risk score by calculating a weighted sum, across the set, of the number of risk alleles carried by each person. The *Allele Score* method (Burgess and Thompson, 2013) is one example of this approach. Unweighted scores are simply the total number of risk-increasing alleles carried by an individual; weighted scores will allow the contribution of each risk-increasing allele to be proportional to the estimated additional risk. The weights are ideally derived from external information, such as previously published genetic associations, to reflect the corresponding genetic effect sizes. If no such external information exists, then cross-validation methods can be used to obtain bias-reduced weights for score construction, which then improves the causal effect estimation.

Several other weighted score methods have also been proposed. The Jackknife instrumental variable estimation method (*JIVE*) (Angrist et al., 1995) estimates $\hat{X}$ of each individual in the first stage regression $X \sim G$ without using the corresponding data points. That is, the ith row in the jackknife matrix $\hat{X}$ is estimated without using ith observation. As a result, that *JIVE* estimator is restricted to a single sample.

When the associations between instruments and the endogenous explanatory variable $\mathbf{X}$ are weak, the naive 2SLS estimator for β may be biased. All weighted score methods (including allele scores and *JIVE*), are expected to be able to reduce weak instrument bias in comparison to multiple instrument methods as described in section 2.3. In fact, weighted scores can be strong instruments even when the individual genetic variants are all weak instruments, and hence bias due to weak instruments is expected to be reduced. However, weighted scores usually also reduce the power (sensitivity) of MR studies (Pierce et al., 2010; Palmer et al., 2012) and require all instruments to be valid.

One limitation of this set of methods is that the choice of weights has a considerable impact on the bias of estimates. One popular and usually stable approach is to obtain weights from an external source with a large sample size, although this could introduce some bias due to cohort differences. If relevant external weights are not available, then we have to use "interval" weights derived from the data under analysis. However, there is no analytical standard error estimation of the resulting causal effect when internal weights are used.

In the context of two-sample analysis, the *Allele score* weights obtained from one sample can be used to infer the causal effect on a second distinct sample, given that $\mathbf{G}, \mathbf{X}, \mathbf{Y}$ are observed in both samples. The weights can be generated by (1) averaging the cross-validated weights from the training data; or (2) sampling from a normal distribution around the training weight with a chosen standard deviation (e.g. 0.01). The latter approximates the uncertainty in the estimation of weights from external data (Burgess and Thompson, 2013). In our two-sample analysis we refer to these two methods as *Allele* and *Allele_sim* respectively.

2.5 Reducing Correlations between G and the Environment

A more sophisticated way to reduce bias due to pleiotropy is to set up a stringent condition for the estimate of β that reduces the correlation between $\mathbf{G}$ and any other factors that influence $\mathbf{Y}$ without going through $\mathbf{X}$. Suppose that

$$Y = X\beta + e,$$

where e includes the effects of $\mathbf{U}$, $\mathbf{Z}$ and ϵ_y. Hence, the goal here is to reduce $Cor(\mathbf{G}, e)$, which is induced through the effects of $\mathbf{U}$ and $\mathbf{Z}$. It must be noted that this only makes sense when there is no causal impact of $\mathbf{X}$ on $\mathbf{Z}$. The Limited Information Maximum Likelihood (*LIML*, (Hayashi, 2000)) finds a conservative estimator of β by minimizing

$$\Phi(\beta) = \frac{(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-\mathbf{1}}\mathbf{G}^\top(\mathbf{y} - \mathbf{X}\beta)}{(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)}$$

instead of minimizing $(\mathbf{y} - \mathbf{X}\beta)^\top \mathbf{G}(\mathbf{G}^\top\mathbf{G})^{-\mathbf{1}}\mathbf{G}^\top(\mathbf{y} - \mathbf{X}\beta)$ in the 2SLS. The denominator in the equation of $\Phi(\beta)$ is the variance of regression errors. *LIML* will be less biased than 2SLS when regressors $\mathbf{X}$ and regression error $\mathbf{y} - \mathbf{X}\beta$ are not independent or when there are many weak instruments (Hahn and Inoue, 2002). However, the variance of *LIML* will increase when instruments are weak and the sample size is small, compared to 2SLS (Blomquist et al., 1999).

The idea behind *LIML* was generalized in the Continuously Updating Estimator approach (*CUE*, (Hansen et al., 1996)). This approach seeks "unbiased" estimator for $\hat{\beta}$ that satisfies the empirical analogue of the following (moment) conditions:

$$E[\mathbf{g_i}(\beta)] = E[\mathbf{G_i}(y_i - x_i\beta))] = \mathbf{0}.$$

This is equivalent to minimizing

$$\hat{\mathbf{g}}^\top(\beta)\mathbf{W}\hat{\mathbf{g}}(\beta),$$

where $\hat{\mathbf{g}}(\beta) = n^{-1}\sum_{i=1}^{n}\mathbf{g_i}(\beta) = n^{-1}\sum_{i=1}^{n}\mathbf{G_i}(y_i - x_i\beta)$ is the sample analog of the population moment conditions $E(\mathbf{G_i}(y_i - x_i\beta)) = \mathbf{0}$ in the generalized method of moments (GMM) framework. Here, $\mathbf{G_i}$ is the ith observation of instruments $\mathbf{G}$ and $\mathbf{W}$ is a weighting matrix. *CUE* defines a weighting matrix $\mathbf{W}(\beta)$ as a function $\mathbf{W}(\beta) = (n^{-1}\sum \mathbf{g_i}(\beta)\mathbf{g_i}^\top(\beta))^{-1}$ to give different weights to each moment condition.

All MR estimators based on the generalized method of moments (framework) are restricted to a single sample, since they all rely on the asymptotic properties of generalized methods of moments. Thus, given two distinct samples, *LIML* and *CUE* can only be applied separately to obtain one causal effect estimate for each sample.

Davies et al. (2015) demonstrated that *CUE* is a better choice than *LIML* and 2SLS when there are many weak instruments. In fact Hansen et al. (2012) and Newey and Windmeijer (2005) showed that generalized methods of moments estimators, including *LIML* and *CUE*, are more resistant to bias in this situation. Furthermore, Bound et al. (1995) showed that

the bias of the 2SLS estimator can be approximated by:

$$E(\hat{\beta}_{\text{2SLS}} - \beta) \approx \frac{\sigma_{\epsilon_x \epsilon_y}}{\sigma^2_{\epsilon_x}} \left(\frac{p-2}{\mu^2} \right), \tag{3}$$

where $\sigma^2_{\epsilon_x}$ is the variance of the regression error ϵ_x in the first stage, and $\sigma_{\epsilon_x \epsilon_y}$ is the covariance of the second stage regression error ϵ_y with ϵ_x. The quantity $\mu^2 = \frac{\hat{\mathbf{X}}^\top \hat{\mathbf{X}}}{\sigma^2_{\epsilon_x}}$ denotes the amount of the variation in $\mathbf{X}$ that is jointly explained by the instruments, where $\hat{\mathbf{X}}$ is the fitted value of $\mathbf{X}$ using $\mathbf{G}$. Hence, the bias of the 2SLS estimator is proportional to the number of instruments (p) and inversely proportional to the variation in $\mathbf{X}$ that is explained by $\mathbf{G}$. That is, the bias will increase if we add more instruments that do not explain the variation of $\mathbf{X}$.

2.6 Valid Instrument Selection Methods

The correlation between $\mathbf{G}$ and $\mathbf{e} = (\mathbf{y} - \mathbf{X}\beta)$ may arise from several kinds of violations of assumptions (A2) and (A3). The previously-described methods in sections 2.3, 2.4 and 2.5 do not use the information contained in $\mathbf{Z}$. We have observed that when $\mathbf{X}$ and $\mathbf{Z}$ are highly correlated, both *LIML* and *CUE* fail to provide consistent estimates of β. Therefore, consistent solutions in the presence of pleiotropy have been proposed to embed valid instrument selection within an instrumental variable method. The motivation here is to incorporate both direct and indirect causal effects from $\mathbf{G}$ to $\mathbf{Y}$ using the following model:

$$Y_i = \mathbf{G_i}\boldsymbol{\delta} + \mathbf{X_i}\beta + \epsilon_{y_i} + \zeta_{\mathbf{y}}\mathbf{U}, \tag{4a}$$
$$E(\epsilon_i | \mathbf{G_i}) = 0, \quad i = 1...n, \tag{4b}$$
$$\mathbf{X_i} = \mathbf{G_i}\boldsymbol{\alpha} + \epsilon_{x_i} \tag{4c}$$

where $\boldsymbol{\delta}$ represents the direct effects of the instruments $\mathbf{G}$ on outcome $\mathbf{Y}$. Indirect effects of $\mathbf{G}$ on $\mathbf{Y}$ are captured through $\mathbf{X}$, and β represents the causal effect parameter of interest. $\boldsymbol{\alpha}$ is the association parameter between $\mathbf{G}$ and $\mathbf{X}$. In recent work by Kang et al. (2016), the authors proposed the *some invalid some valid IV estimator* (sisVIVE) to minimize

$$(\beta, \boldsymbol{\delta}) \in \text{argmin} \frac{1}{2} ||\mathbf{P_G}(\mathbf{Y} - \mathbf{G}\boldsymbol{\delta} - \mathbf{X}\beta)||_2^2 + \lambda ||\boldsymbol{\delta}||_1,$$

where $\mathbf{P_G} = \mathbf{G}(\mathbf{G}^\top \mathbf{G})^{-\mathbf{1}} \mathbf{G}^\top$. The first term essentially replaces $\mathbf{X}$ in 2SLS with $(\mathbf{X}, \mathbf{G})$, and thereby the direct causal effect of $\mathbf{G}$ on $\mathbf{Y}$ (via $\mathbf{Z}$) is taken into consideration. The second term $\lambda ||\boldsymbol{\delta}||_1$ enforces an L_1 sparse selection of invalid instruments using a LASSO prior. sisVIVE is robust to certain invalid instruments and their direct causal effect on $\mathbf{Y}$ (without going through $\mathbf{X}$). However, this method's ability to select valid instruments is limited by the assumption that at least 50% of all instruments must be valid.

sisVIVE treats pleiotropic phenotypes as general sources of the indirect causal effect $\mathbf{G}$ on $\mathbf{Y}$, and does not use the information on $\mathbf{Z}$ explicitly. Given variables $\mathbf{Z}$, Kang et al. (2016) suggests either adjusting for them or using them as exogenous variables. We refer to these two methods as "*sisVIVE*_adj" and "*sisVIVE*_exo" and implement them using R package sisVIVE.

If two distinct samples are available, sisVIVE can be extended to a two-sample estimator. First we can apply sisVIVE on a single sample to obtain estimation of $\boldsymbol{\delta}$ and corresponding valid instrument selections. Then this estimation of $\boldsymbol{\delta}$ can be carried over to the MR analysis of a second sample. The corresponding adjustment for pleiotropic phenotypes $\mathbf{Z}$ can also be incorporated within this two-sample estimator based on sisVIVE.

In summary, a high-level comparison of all the methods described above is provided in Table 1, contrasting key features. It is also worth noting that the descriptions above are largely restricted to the case of a univariate variable $\mathbf{X}$, although most methods can be easily extended to multivariate $\mathbf{X}$ (see Discussion). Among all methods new approaches that can incorporate external information from a different sample (in two-sample setup), or construct valid instrumental variables to address pleiotropy, are attracting increased attention.

	Egger	*Allele score*	*2SLS_adj*	*2SLS_mul*	*JIVE*	*LIML*	*CUE*	*sisVIVE*	*CIV*
Summarized Data	✓	✓ ✗ [a]	✗	✗	✗	✗	✗	✗	✗
Analysis of Variance (β)	✗	✗	✓	✓	✗	✓	✓	✗	✗
Selection of Valid IVs	✗	✗	✗	✗	✗	✗	✗	✓	✓
Instrument Construction	✗	✓	✗	✗	✗	✗	✗	✓	✓
Stage I and stage II separation	✗	✓	✓	✓	✗	✗	✗	✗	✓
Sample size requirement [b]	NA	$p < n$	NA	NA	$p < n$	NA	$p < n$	$p < n$	NA
Main Assumption	InSIDE	linearity	**Z** is an exogenous variable	**X** and **Z** independent	Asymptotic distribution assumptions	Linearity [c]	Linearity	At least 50% IVs are valid	Linearity
Motivation	Adapt meta-analysis techniques to test for bias from pleiotropy	Construct one strong IV	control exogenous effect of **Z** in 2SLS	Use **G** to account for exogenous effect of **Z** (without controlling it) in 2SLS	Construct instruments that are independent of small sample disturbances	Reduce correlation between **G** and regression error **e**	Special case of GMMs with better performance using weak IVs	Considers both direct and indirect $G \to Y$ causality	Select valid and strong (as much as possible) instruments given **Z**

Table 1: Properties of selected Mendelian randomization methods. Columns: Selected instrumental variable methods. Rows: Selected properties. ✓: Yes; the corresponding method has this property, or can be used in this scenario. ✗: No.

[a] *Allele score* method with internal weights is not applicable to summarized data analysis.

[b] n: sample size. p: number of genetic variants.

[c] Linearity in the association between **X** and **Y**, and between **G** and **X**.

3 Constrained Instrumental Variable method

The Constrained Instrumental Variable (*CIV*) method proposed here is designed to maximize instrument strength yet provide robustness to pleiotropic effects, specifically for the situation where potentially pleiotropic phenotypes ($\mathbf{Z}$) are measured and available.

3.1 *CIV* when $\mathbf{p < n}$

Specifically, we are interested in finding a weight vector $\mathbf{c} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^r$, s.t.

$$\max_{\mathbf{c}\in\mathbb{R}^p, \mathbf{v}\in\mathbb{R}^r} \mathbf{c}^\top \mathbf{G}^\top \mathbf{X}\mathbf{v} \tag{5}$$

subject to conditions:

$$\mathbf{c}^\top \mathbf{G}^\top \mathbf{G}\mathbf{c} = 1, \tag{6a}$$
$$\mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} = 1, \tag{6b}$$
$$\mathbf{c}^\top \mathbf{G}^\top \mathbf{Z} = \mathbf{0}. \tag{6c}$$

The motivation of this method is to construct a relatively strong instrumental variable $\mathbf{Gc}$ (a linear combination of variables in $\mathbf{G}$) that is uncorrelated with $\mathbf{Z}$. In this way the weak instrument bias and the pleiotropic effect are alleviated in Mendelian randomization with the new instrumental variable. We use Equation (5) to obtain the maximized canonical correlation between $\mathbf{Gc}$ and $\mathbf{X}$. In addition we use Equation (6c) to force the new instrumental variable $\mathbf{Gc}$ to be orthogonal to all possible pleiotropic phenotypes in $\mathbf{Z}$. This maximization problem is well-defined when $p \geq k$ (see Appendix A).

The strength of the *CIV* instruments can be measured with an F-statistic or with a concentration parameter (Stock et al., 2012). The former can be calculated as the F-statistic from a linear regression model against the null that the excluded instruments are irrelevant in the first-stage regression $\mathbf{X} \sim \mathbf{G}$. The latter measures the overall association between X and G without considering the number of instruments used. As a rule of thumb, a single instrumental variable with F-statistics < 10 is usually considered weak instrument. *CIV* is designed to retain instrument strength, however, it may not always yield the strongest global F-statistic since the linear constraint (6c) may force our method to exclude some genotypes, that are associated with both $\mathbf{X}$ and $\mathbf{Z}$, from the construction of CIV.

In the context of one-sample analysis, the solution, $\mathbf{c}$, is used to construct a new instrumental variable $\mathbf{G}^* = \mathbf{Gc}$. The new *CIV* is used directly to infer the causal effect of $\mathbf{X}$ on $\mathbf{Y}$ from $(\mathbf{G}^*, \mathbf{X}, \mathbf{Y})$ using estimation methods for linear structural equation modeling methods such as 2SLS. Alternatively *CIV* can be embedded inside a bootstrap to find a bias-corrected *CIV* weight. We refer to the latter as "CIV_boot".

In the context of model assessment, or when we have two different sets of participants available, we may want to split data into two datasets. Specifically, *CIV* is trained on the

first (training) set, and the solution $\mathbf{c}$ is then applied to the second dataset to construct new instrumental variable $\mathbf{Gc}$ to infer causal effect. Some other methods, such as *Allele score* method can also be implemented in this way since both of the methods rely on the first stage pathway ($\mathbf{X} \sim \mathbf{G}$) to construct instruments, and this process is separated from second stage regression ($\mathbf{Y} \sim \mathbf{X}$) .

One limitation of the *CIV* method lies in the fact that a solution $\mathbf{c}$ only exists when $p < n$. In fact, when $p < n$, there is an unique solution (see Appendix A). This weighted score can then be used directly to estimate the causal effect of $\mathbf{X}$ on $\mathbf{Y}$. However, a solution only exists when $p < n$ since $\mathbf{G}^\top\mathbf{G}$ is not invertible when $p > n$. In addition, we may want to select a subset of valid instruments from among all SNPs (the columns of $\mathbf{G}$) of interest. Therefore, we propose an extension of the previous solution that addresses these two concerns, through imposing a penalty on the problem (5).

3.2 Smoothed CIV

Different choices of penalty functions on the problem (5) lead to different solutions. However, popular LASSO and L_2 penalties would not result in a sparse solution here under any level of regularization because of the linear constraint (6c). An explanation can be understood by examining Figure 4. The LASSO and L_2 contours will touch the linear constraints (straight line in the figure) at two non-sparse solutions of $\mathbf{c}$.

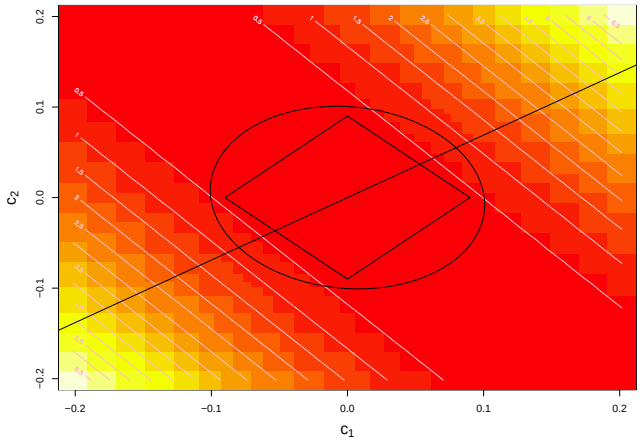


Figure 4: Graph demonstrating the maximization problem with LASSO penalty and L_2 penalty. Rectangle: LASSO penalty contour with the same level of penalization. Circle: L_2 penalty contour with the same level of penalization. Straight line represents the CIV solution space required, and it does not intersect with a sparse solution. Pixels with color from yellow to red: coordinates of $\mathbf{c} = (c_1, c_2)$ with absolution correlation values from high levels to low levels.

Instead we consider an L_0 penalty and maximize the function

$$\max_{\mathbf{c}\in\mathbb{R}^p,\mathbf{v}\in\mathbb{R}^r} \mathbf{c}^\top\mathbf{G}^\top\mathbf{X}\mathbf{v} - \lambda|\mathbf{c}|_0 \tag{7}$$

subject to conditions:

$$\mathbf{c}^\top\mathbf{G}^\top\mathbf{G}\mathbf{c} \leq 1, \tag{8a}$$
$$\mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v} \leq 1, \tag{8b}$$
$$\mathbf{c}^\top\mathbf{G}^\top\mathbf{Z} = \mathbf{0}, \tag{8c}$$

where $|\mathbf{c}|_0$ is the L_0 norm of $\mathbf{c}$ and λ is a regularization parameter.

This problem is equivalent to maximizing a convex function over a convex set. However, it is computationally impractical to exhaustively enumerate all possible sets of $|\mathbf{c}|_0$; this problem with L_0 norm has been proven to be NP-hard (Natarajan, 1995). Therefore, we propose instead to consider smoothed L_0 penalties: $f_\sigma(x) = \exp(-\frac{x^2}{2\sigma^2})$. In the limit when $\sigma \to 0$, $|\mathbf{c}|_0 \approx p - \sum_j f_\sigma(c_j)$, thereby the problem (7) can be approximated by:

$$\max_{\mathbf{c}\in\mathbb{R}^p,\mathbf{v}\in\mathbb{R}^r} \mathbf{c}^\top\mathbf{G}^\top\mathbf{X}\mathbf{v} - \lambda(p - \sum_j f_\sigma(c_j)) \tag{9}$$

subject to conditions (8a), (8b) and (8c). The approach is then to solve problem (9) for a decreasing sequence of σ ($\to 0$) and a given value λ, while resulting in at least approximately sparse solutions. However, there are no theoretical guarantees for the uniqueness of such numerical solutions, and often there are multiple solutions.

In order to implement this smoothed L_0 algorithm and obtain a single solution $\mathbf{c}$ for a given value of λ, we proceed as follows:

1. Initialization: For a given value of λ, start from an initial guess $\tilde{\mathbf{c}}$ and initial L_0 penalty $\sigma_{\max} = \max_j |\tilde{c}_j|$, set $\sigma = \sigma_{\max}$.
2. While $\sigma > \sigma_{\min} = 0.01$ we do
   - i Calculate the gradient of function (7) $\mathbf{d} \in \mathbb{R}^p$, where $d_j = \frac{\lambda\tilde{c}_j}{\sigma^2}\exp(-\frac{\tilde{c}_j^2}{2\sigma^2}) - 2[\tilde{\mathbf{c}}^\top\mathbf{G}^\top\mathbf{M}\mathbf{G}]_j, j \in \{1,...,p\}$ and $\mathbf{M} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top$.
   - ii Set $\mathbf{c} = (\mathbf{I} - \mathbf{A}^-\mathbf{A})(\tilde{\mathbf{c}} - \mu\mathbf{d})$ where $\mathbf{A}^-$ is a general inverse of $\mathbf{A} = \mathbf{Z}^\top\mathbf{G}$ and μ is a step-size parameter in gradient descent algorithm.
   - iii Set $\mathbf{c}^* = \mathbf{c}/\sqrt{\mathbf{c}^\top\mathbf{G}^\top\mathbf{G}\mathbf{c}}$ as the updated solution.
   - iv Repeat (i) (ii) and (iii) (maximum T times) until it converges, i.e. $\sqrt{\sum_{j=1}^{p}(|\mathbf{c}_\mathbf{j}^*| - |\mathbf{c}_j|)^2}/p < 10^{-10}$.

3. Update σ with $\sigma = 0.5\ \sigma_{\text{prev}}$, where σ_{prev} is the previous value of σ used in step 2. If $\sigma > \sigma_{\min}$ repeat all items in step 2. If not, stop the algorithm and record the last iteration of $\mathbf{c}$ as the final solution.

In summary, the maximization problem of Equation (9) is solved by repeatedly taking gradient descent steps (i), and then projecting the possible solution back into constrained set ((ii),(iii)). Note that the step (ii) restricts the solution to be on the constrained set (8c) and step (iii) restricts to the boundary of the constrained set (8a). Note that the unconstrained gradient descent step followed by projection to the feasible set is equivalent to a direct gradient descent step on the feasible set (Cui et al., 2010). The parameters for step-size (μ) and number of iterations (T) should be carefully chosen to achieve balance between computation cost and precision. That is, the states discovered by this algorithm may not achieve the maximized value of Equation (7) even with a large number of iterations, if we use a step size that is too large. The decreasing list of values for σ is chosen to ensure that the approximation accuracy will gradually increase.

The tuning parameter λ affects the prediction performance of our CIV method, as for any penalization method. Higher values of λ lead to stronger penalization on the L_0 norm of $\mathbf{c}$ and an approximately sparser solution for $\mathbf{c}$. This means that more valid instruments will be considered as invalid and omitted from construction of the CIV weighted scores. Smaller values of λ correspond to less regularization and thus lead to less sparse solutions. The ideal value for λ will depend on the actual proportion of invalid instruments among all instruments.

We therefore implement a K-fold cross-validation technique to find an optimal value for λ that minimizes the projected prediction error $||\mathbf{P}_{\mathbf{G}^*}(\mathbf{Y} - \mathbf{X}\beta_*)||$, where $\mathbf{P}_{\mathbf{G}^*} = \mathbf{G}^{*\top}(\mathbf{G}^{*\top}\mathbf{G}^*)^{-\mathbf{1}}\mathbf{G}^*$. We choose to use the projected prediction error as the tuning measurement in order to make $\mathbf{G}^* = \mathbf{G}\mathbf{c}$ as "valid" as possible rather than as "informative" as possible. In the ideal case where $\mathbf{G}^*$ is a valid instrument whose causal impact on $\mathbf{Y}$ only goes through $\mathbf{X}$, then the regression residual $\mathbf{Y} - \mathbf{X}\beta^*$ should be orthogonal to any vectors spanned by the original instruments in $\mathbf{G}$. In other words, only valid instruments $\mathbf{G}$ will yield $\mathbf{P}_{\mathbf{G}^*}(\mathbf{Y} - \mathbf{X}\beta^*) = 0$. In general, we are more interested in the validity of the prediction model rather than the most informative solution of β; the latter may lead to over-estimation of the causal effect of interest.

There may be multiple local solutions of $\mathbf{c}$ to the smoothed problem Equation (9), since this a non-convex optimization problem. A careful reader may recognize that it implies maximization of a convex function over a convex set, yet overall this is not a convex problem! As a result, a local maximum solution of $\mathbf{c}$ may not be the global maximum solution, and numerical optimization techniques may get trapped into a local minimum. Therefore, we start from multiple (e.g. 100) initial points randomly sampled from a multivariate normal distribution $N(0, I_p)$, and let the smoothed L_0 algorithm converges to a set of solutions, possibly arriving at multiple local modes of $\mathbf{c}$. After examining correlations between all pairs of solutions $(\mathbf{c}^{(1)}, \mathbf{c}^{(2)})$, highly correlated solutions (corr ≥ 0.9) are removed. The remaining solutions are combined into a matrix, $\mathbf{c}^*$, of row dimension p. Finally, we construct new instruments $\mathbf{G}^* = \mathbf{G}\mathbf{c}^*$ and refer to this approach as *CIV_smooth*.

3.3 Causal Effect Estimation

The causal effect estimate of exposure on response β is now obtained with valid instruments $\mathbf{G}^*$. Several alternative methods can be used to infer causal effects from $\mathbf{G}^*, \mathbf{X}, \mathbf{Y}$. For simplicity, we choose the two stage least square (2SLS) estimator as the default causal effect estimator. Remember that the 2SLS estimator is defined as $\hat{\beta}_{2SLS} = (\mathbf{X}^\top \mathbf{P_G} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P_G} \mathbf{Y}$; this is equivalent to implementing the least square regression twice, in the first stage $\mathbf{X} \sim \mathbf{G}^*$ and the second stage $\mathbf{Y} \sim \mathbf{X}^*$. The asymptotic variance of the 2SLS estimator β can be estimated with:

$$\hat{\beta}_{\mathbf{2SLS}} = \beta + (\mathbf{X}^\top \mathbf{G} (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{G}^\top \mathbf{G})^{-1} \mathbf{G}^\top (\mathbf{Y} - \mathbf{X}\beta) \tag{10}$$

$$\sqrt{\mathbf{n}}(\hat{\beta}_{\mathbf{2SLS}} - \beta) \to N(0, \mathbf{A}),$$

$$\mathbf{A} = \hat{\sigma}_e^2 [(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x_i} \mathbf{G_i}^\top)(\frac{1}{n} \sum_{i=1}^{n} \mathbf{G_i} \mathbf{G_i}^\top)^{-1} (\frac{1}{n} \sum_{i=1}^{n} \mathbf{x_i} \mathbf{G_i}^\top)^\top]^{-1},$$

$$\hat{\sigma}_e^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x_i}^\top \hat{\beta}_{2SLS})^2,$$

It is worth noting that $\mathbf{Y} - \mathbf{X}\beta$ is the regression error term which is assumed with homoscedastic variance. However, such asymptotic estimation of variance is not available for the CIV methods, since the new instruments G^* are dependent on all observations of $\mathbf{X}$ and $\mathbf{Z}$, and hence $\mathbf{X_i} \mathbf{G_i}^{*\top}$ and $\mathbf{X_j} \mathbf{G_j}^{*\top}$ are not independent. Therefore, the assumption of weak law of large numbers is violated and puts the convergence of $\frac{1}{n} \sum_{i=1}^{n} \mathbf{X_j} \mathbf{G_j}^{*\top} \xrightarrow{P} E[\mathbf{X_j} \mathbf{G_j}^{*\top}]$ into jeopardy.

Bootstrapping can be used to estimate the sample variance of $\hat{\beta}$ and obtain confidence intervals. The resulting empirical confidence intervals should be interpreted with caution and compared with other methods (such as the *Allele score* method) cautiously, especially when weak instruments are present (Moreira et al., 2009).

Both CIV and *CIV_smooth* can be applied to the case when two separate samples are available. For CIV construction, the weight $\mathbf{c}$ can be estimated on the first sample and then applied to the second sample for causal effect estimation. The same approach can be applied to the *CIV_smooth* algorithm. An alternative way to apply *CIV_smooth* to the two-sample MR analysis, similar to *sisVIVE*, is to conduct valid instrument selection using the first sample with *CIV_smooth*, and then to conduct causal effect estimation with that information on a second sample. All these three methods are included in our two-sample analysis below.

4 Simulation

4.1 Simulation Design

Simulations have been conducted under a variety of different kinds of violations of the MR assumptions in order to compare the performance of our CIV methods with other popular methods. In all simulations, we assume that there is a fairly large set of potential genetic instruments $\mathbf{G}$, and that individual level data are available for the phenotype of interest $\mathbf{X}$, potential pleiotropic phenotype $\mathbf{Z}$, and the response $\mathbf{Y}$.

Table 2 provides the broad goals behind four series of simulations designed to address different types of violations of the key assumptions. All simulations assume the presence of pleiotropy, and hence assumption (3) of the MR assumptions is always violated. In Series I, we generate strong but pleiotropic instruments $\mathbf{G}$ for phenotype of interest $\mathbf{X}$; this violates the assumption (A2). In Series II, we investigate the association between $\mathbf{X}$ and $\mathbf{Z}$, by varying the direction and strength of the simulated causal relationships between these two sets of variables. In Series III, we simulate a set of weak, pleiotropic instruments, where the associations between $\mathbf{G}$ and $\mathbf{X}$ are not strong; this violates assumption (A2) and jeopardizes (A1) of the three MR assumptions. Finally, in Series IV, we examine performance when the selection of important SNPs is desirable, such as what one might expect if many variants of interest are simultaneously considered within an MR analysis. Parameter settings in common across scenarios are given in Table 2. It is worth noting that we do not include simulations with $p \geq n$, i.e. when there are more genotypes in $\mathbf{G}$ than observations. The reason is *CIV* and *CIV_smooth* are the only methods that would work under such restrictions and we do not have competitors in this case.

Table 3 summarizes the implementation details of the specific methods used for causal inference in our simulations. Note that we have two variants of *CIV* and *CIV_smooth* methods: *CIV_boot* and *CIV_smooth.sel*. The former is a bootstrapped version of *CIV*; that is, a bootstrap corrected estimate of $\mathbf{c}$ is obtained and used to infer a causal effect from $(\mathbf{G}, \mathbf{X}, \mathbf{Y})$. The latter is a selection method based on *CIV_smooth*: we first obtain *CIV_smooth* estimates $\hat{\mathbf{c}}$. For each converged solution $\hat{\mathbf{c}}$, a feature j is recognized as significant if coefficient $|c_j| \geq \psi * \max_j |c_j|, j = 1, ..., p$. This criteria ψ can be tuned with cross-validation to achieve optimal selection performance in application. In our simulations we choose a small value $\psi = 0.2$ for simplicity. All selected features are then recognized as valid instruments $\mathbf{G}^*$ for MR analyses. *CIV_boot* method is included in all simulations to check the consistency of *CIV*, while *CIV_smooth.sel* method is only applied in simulation Series IV to test the feature selection performance of *CIV_smooth* solutions. The description of all other methods can be found in Section 2. In series I, II and III, all methods in Table 3 are implemented for one-sample analysis, and a few comparable methods are used for two-sample analysis. In series IV, only applicable feature selection methods are considered in simulation.

Simulation Series	**n**	η	p	p_z	**MAF**	α_x	α_z	F_x	**Methods (One Sample)**
I. Standard Pleiotropy	500	1.0	9	2	0.33	1	1	42.86	All methods
II. $\mathbf{X} \to \mathbf{Z}$ or $\mathbf{Z} \to \mathbf{X}$	500	1.0	50	15	0.33	0.1 ; 0.5	0.1 ; 0.5	3.6;10.04 3.57;9.75	All methods
III. Weak Instruments	500	1.0	9 25 100	2 5 22	0.33	0.2,0.5 0.12,0.3 0.06,0.15	0.2,0.5 0.13,0.32 0.06,0.15	6.64,26.42 3.22,8.84 0.8,2.26	All methods
IV. Selection of Valid Instruments	500	1.0	50	10	0.33	2 ; 0.1	0.5	11.13; 2.44	*CIV Variants* *sisVIVE Variants*

Table 2: Some of the parameter settings used in the four series of simulations. n: number of individuals. p: number of genotypes in total. p_z: number of pleiotropic genotypes with no effect on $\mathbf{X}$. MAF: minor allele frequency of all SNPs in the simulation. Methods: the methods compared in each simulation in one-sample setups. α_x: association parameter between $\mathbf{G}$ and $\mathbf{X}$ as in Figure 3. α_z: association parameter between $\mathbf{G}$ and $\mathbf{Z}$ as in Figure 3. F_x: the expected F-statistic values for testing the strength of instruments in $\mathbf{G}$ (for $\mathbf{X}$) given values of α_x, α_z and other parameters. All methods used in one sample analysis: 2SLS_naive, 2SLS_adj, *2SLS_mul*, *JIVE*, Allele Score, *Egger* Regression, *LIML*, *sisVIVE*_adj, *sisVIVE*_exo, *CUE*, *CIV*, *CIV_smooth* and *CIV_smooth.sel*.

4.2 Simulation Series I, II and III : Implementation

4.2.1 Simulation Series I : Standard Pleiotropy

In Series I, we simulated a standard pleiotropy problem with generated values of genotypes $\mathbf{G}$, a phenotype of interest $\mathbf{X}$, a pleiotropic phenotype $\mathbf{Z}$ and an outcome of interest $\mathbf{Y}$. The values of parameters were carefully chosen so that all the 9 genotypes in $\mathbf{G}$ were strong instruments for $\mathbf{X}$ and 2 of them had pleiotropic effects on $\mathbf{Y}$ via $\mathbf{Z}$. 200 replications of the simulation for both one-sample and two-sample setups were generated. The results were compared across all approaches described in the section 2. We generated Datasets as follows:

$$\begin{aligned} x_i &= \alpha_x \sum_{j=1}^{9} G_{ij} + u_{x_i} + \epsilon_{x,i} \\ z_i &= \alpha_z \sum_{j \in G_z} G_{ij} + u_{z_i} + \epsilon_{z,i} \\ y_i &= z_i + u_i + \epsilon_{y,i} \end{aligned} \tag{11}$$

where x_i, z_i, y_i are the ith observation of $\mathbf{G}, \mathbf{Z}, \mathbf{Y}$. Let G_{ij} denote the value of ith observation of jth genotype. u_{x_i}, u_{z_i}, u_i represent the effect of confounding factor $\mathbf{U}$ on $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{Y}$ respectively. $\epsilon_{x,i}, \epsilon_{z,i}, \epsilon_{y,i}$ are the independent error of $\mathbf{G}, \mathbf{Z}, \mathbf{Y}$ for ith observation respectively.

The simulation of Equation (11) was based on 9 ($p = 9$) independent variants $\mathbf{G}$ simulated with a minor allele frequency 0.3 and a sample size of $n = 500$. The pleiotropic subset,

Method	Label/Variants	Parameter Choice	Treatment of $\mathbf{Z}$
2SLS	2SLS_naive	NA	NA
	2SLS_adj	NA	Adjust $\mathbf{G} \sim \mathbf{Z}$
	2SLS_mul	NA	Combine $(\mathbf{X}, \mathbf{Z})$ to be $\mathbf{X}^*$
JIVE	*JIVE*	NA	Combine $(\mathbf{X}, \mathbf{Z})$ to be $\mathbf{X}^*$
Allele	Allele	10 fold cross-validation	Combine $(\mathbf{X}, \mathbf{Z})$ to be $\mathbf{X}^*$
	Allele_sim	"precise weight" with standard deviation 0.01 (Burgess and Thompson, 2013)	Combine $(\mathbf{X}, \mathbf{Z})$ to be $\mathbf{X}^*$
Egger	*Egger*	NA	NA
LIML	*LIML*	NA	NA
	*LIML*_exo	NA	Adjust $(\mathbf{G}, \mathbf{X}, \mathbf{Y}) \sim \mathbf{Z}$
sisVIVE	*sisVIVE*_adj	10 fold cross-validation; A numeric vector of penalization parameter must be given for cross-validation.	Adjust $\mathbf{G} \sim \mathbf{Z}$
	*sisVIVE*_exo	10 fold cross-validation; list of penalization parameter	Adjust $(\mathbf{G}, \mathbf{X}, \mathbf{Y}) \sim \mathbf{Z}$
CUE	*CUE*	NA	Combine $(\mathbf{X}, \mathbf{Z})$ to be $\mathbf{X}^*$
CIV	*CIV*	NA	Embedded
	CIV_boot	100 bootstrap samples	Embedded
	CIV_smooth	10 fold cross-validation; A numeric vector of penalization parameter must be given for cross-validation; 100 random initial points	Embedded
	CIV_smooth.sel	10 fold cross-validation; A numeric vector of penalization parameter must be given for cross-validation; 100 random initial points; select variants whose coefficients >0.2 * max(coefficients)	Embedded

Table 3: The methods and parameter settings used in the simulations. Label: the actual label used in figures. Parameter Choice: the parameter specifications for each of the methods in the simulations. Treatment of $\mathbf{Z}$: the way to incorporate pleiotropic phenotype $\mathbf{Z}$ in different methods.

G_z, containing 2 SNPs ($p_z = 2$) sampled without replacement, that were also associated with $\mathbf{Z}$. Given that the SNPs were assumed to be independent with the same minor allele frequency, each SNP (coded as 0,1,2 for the number of minor alleles) had variance $\sigma_G^2 = 0.42$. The concentration parameter, μ_x^2, the amount of the variation in the exposure that was jointly explained by the instruments, could therefore be written as $\mu_x^2 = \frac{np\alpha_x^2\sigma_G^2}{\sigma_{X_{u\epsilon}}^2} = 0.21np\alpha_x^2$, where $\sigma_{X_{u\epsilon}}^2 = 2$ is the variance of the regression residual $(\epsilon_x + u_x)$ in the first stage of Equation (11). The concentration parameter divided by p, given the OLS estimates of α_x and $\sigma_{X_{u\epsilon}}^2$, was equal to the F-statistic for testing the null hypothesis $\alpha_x = 0$. We set $\alpha_x = \alpha_z = 1$ to ensure $\mathbf{G}$ were strong instruments (F statistics > 10) for both $\mathbf{X}$ and $\mathbf{Z}$.

We implemented simulations in both one-sample and two-sample setups for Series I, in which specific MR methods (in Table 3) were compared. The one-sample setup corresponded to a sample of $n = 500$ observations of $\mathbf{G}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}$. In addition, we assessed performance of our *CIV* instruments and compared with appropriate comparators, including *Allele score* methods and *sisVIVE* methods in the two-sample setup. Specifically, we generated two different sets of $\mathbf{G}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}$, and each set contains 500 observations. We analyzed the first dataset with the variants of *Allele score* methods, *sisVIVE* methods and *CIV* methods (see Table 3). The weights obtained from the first dataset for instrumental variable construction were then applied to the second data set. The corresponding measurements of instrument strength (F-statistic), correlations between new instruments ($\mathbf{G}^*$) and pleiotropic phenotype $\mathbf{Z}$, and causal effect estimation bias from second data were recorded and compared in Figure 5. We chose the F-statistic as a measure of instrument strength instead of the concentration parameter (in one-sample analysis) because we have different numbers of instrumental variables from different methods, and only the F-statistic takes that into account.

4.2.2 Simulation Series II : Direct Causal Effect between X and Z

In Series II, we simulated direct causal links between $\mathbf{X}$ and $\mathbf{Z}$ to study the impact of pleiotropy on causal effect estimation of all applicable methods in both one-sample and two-sample set-ups. Specifically, $\mathbf{G}$ were generated as instruments for $\mathbf{X}$ and a proportion (p_z) of them were generated as genetic causes for phenotype $\mathbf{Z}$. Direct causal relationships between $\mathbf{X}$ and $\mathbf{Z}$, either $\mathbf{X} \to \mathbf{Z}$ or $\mathbf{Z} \to \mathbf{X}$, were then simulated. The outcome $\mathbf{Y}$ was depending on both of $\mathbf{X}$ and $\mathbf{Z}$. In this way, the correlation between $\mathbf{X}$ and $\mathbf{Z}$ was induced partially through direct causal relationships and partially through overlapping genetic causes.

As far as the direct causal relations between $\mathbf{X}$ and $\mathbf{Z}$ is concerned, we find that the violation of MR assumption (A2) is inevitable when $\mathbf{X} \to \mathbf{Z}$. Specifically, when there are direct causal relations $\mathbf{Z} \to \mathbf{X}$ and not all genetic variants in $\mathbf{G}$ are associated with $\mathbf{Z}$, the situation can be considered as pleiotropy and only the genetic variants not directly related with $\mathbf{Z}$ are valid instruments. However, if the causal relation goes from $\mathbf{X} \to \mathbf{Z}$, then all genetic variants in $\mathbf{G}$ are invalid and this should not be considered as pleiotropy. In addition, when there is a strong link between $\mathbf{X}$ and $\mathbf{Z}$, the two sets of variables are likely to be highly correlated, and this in itself can lead to instability of the estimation of causal effect, β.

In Simulation Series II, datasets were generated containing direct causality between x

and $\mathbf{Z}$ as follows:

$$
\begin{aligned}
&y_i = \beta x_i + z_i + u_i + \epsilon_{y,i} \\
&z_i = \alpha_z \sum_{j \in G_z} G_{ij} + u_{z_i} + \epsilon_{z,i}, \quad x_i = \alpha_x \sum_{j=1}^{p} G_{ij} + z_i + u_{x_i} + \epsilon_{x,i} \\
&\text{or} \\
&y_i = \beta x_i + z_i + u_i + \epsilon_{y,i} \\
&x_i = \alpha_x \sum_{j=1}^{p} G_{ij} + u_{x_i} + \epsilon_{x,i}, \quad z_i = \alpha_z \sum_{j \in G_z} G_{ij} + x_i + u_{z_i} + \epsilon_{z,i}
\end{aligned}
\tag{12}
$$

In each simulated dataset, 50 SNPs ($\mathbf{G}$ in Equation (12)) with a minor allele frequency 0.3 and $n = 500$ observations were generated. Both directions of the causality between $\mathbf{X}$ and $\mathbf{Z}$ were considered, i.e. in one case $G \to X \to Z$ was generated, and in the other case $G \to Z \to X$ was generated. In each case, $\alpha_x \in (0.1, 0.5)$ encompassed one scenario with weak instruments and another with strong instruments. 200 datasets were generated for each scenario, and results compared the estimates and variance of the causal effect, β.

We ran both one-sample and two-sample simulations in Series II using the methods introduced in Table 3. The instrumental variable strength, correlation between $\mathbf{G}^*$ and $\mathbf{Z}$, and causal effect estimation bias in two-sample simulations of selected methods were summarized in Figure 6 and Figure 7. The performance in one-sample simulations under the causal null, when there were no true associations ($\beta = 0$), was reported in Figures 10 and 11 for the scenarios $\mathbf{X} \to \mathbf{Z}$ and $\mathbf{Z} \to \mathbf{X}$.

4.2.3 Simulation Series III : Weak Instruments

In Series III, a set of genotypes $\mathbf{G}$ ($p = 9, 25$ or 100) were generated as weak instruments for $\mathbf{X}$. A proportion of them (2/9, 5/25 or 22/100) had pleiotropic effects on $\mathbf{Y}$ via $\mathbf{Z}$, in which case both phenotypes ($\mathbf{X}$ and $\mathbf{Z}$) were endogenous variables. We generated samples of ($\mathbf{X}$, $\mathbf{Z}$, $\mathbf{Y}$) according to Equation 13. The choices of parameter values, as in Table 2, were based on Davies et al. (2015), and the results of all approaches (in both one-sample and two-sample setups) were displayed and discussed.

$$
\begin{aligned}
x_i &= \alpha_x \sum_{j=1}^{p} G_{ij} + u_{x_i} + \epsilon_{x,i} \\
z_i &= \alpha_z \sum_{j \in G_z} G_{ij} + u_{z_i} + \epsilon_{z,i} \\
y_i &= \beta x_i + z_i + u_i + \epsilon_{y,i}
\end{aligned}
\tag{13}
$$

Sets of independent variants $\mathbf{G}$ in Equation (13) were generated with $p = 9, 25$, or 100, each with a minor allele frequency of 0.3 and a sample size of $n = 500$. The pleiotropic

subset, G_z ($p_z = 2, 5$ or 22 respectively), containing genotypes sampled without replacement, was also associated with $\mathbf{Z}$. The values for parameter α_x were carefully chosen so that, for different numbers of SNPs p, all variants in $\mathbf{G}$ were weak instruments (F statistics < 10) with the same magnitude (μ_x). We also chose the values for α_z and $p_z = |G_z|$ (number of pleiotropic genotypes) to make sure the magnitudes of the concentration parameter for the pleiotropic effect (μ_z) were the same for different values of p.

The parameter combinations evaluated in specific simulation scenarios were recorded in Table 4, capturing simultaneously the effects of weak instruments and pleiotropic variants, with a range of different numbers of instruments ($p = 9, 25, 100$). The performances of *CIV* instruments in two-sample setup were illustrated in Figure 8. Performances under the causal null, when there were no true associations ($\beta = 0$), in one-sample setup were summarized in Figures 12, 13 and 14, for different values of p. The performance of all methods under true value $\beta = 1$ can be found in the Appendix (Figures 19, 20 and 21).

	p	p_z	α_x	α_z	μ_x^2	μ_z^2
Scenario1	9	2	0.2	0.2	37.8	8.4
Scenario2	9	2	0.2	0.5	37.8	52.5
Scenario3	9	2	0.5	0.2	236.25	8.4
Scenario4	9	2	0.5	0.5	236.25	52.5
Scenario5	25	5	0.12	0.13	37.8	8.4
Scenario6	25	5	0.12	0.32	37.8	52.5
Scenario7	25	5	0.3	0.13	236.25	8.4
Scenario8	25	5	0.3	0.32	236.25	52.5
Scenario9	100	22	0.06	0.06	37.8	8.4
Scenario10	100	22	0.06	0.15	37.8	52.5
Scenario11	100	22	0.15	0.06	236.25	8.4
Scenario12	100	22	0.15	0.15	236.25	52.5

Table 4: Simulation scenarios for Series III simulations with weak instruments. Data were generated for each simulation scenario with two values of $\beta \in (0, 1)$, and replicated 200 times (i.e. 200 datasets were generated).

4.3 Simulation Series I, II and III : Results

4.3.1 Instrument Strength

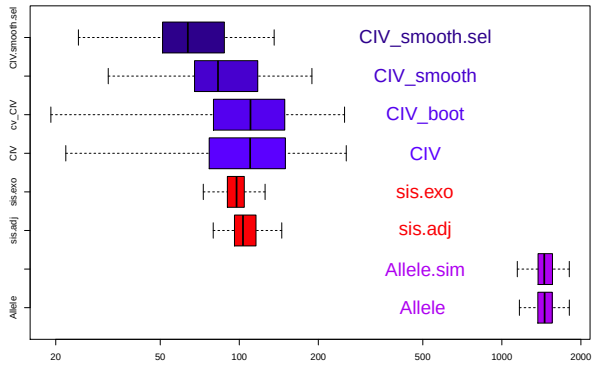
Figure 5 shows one example of the instrument strength results from the two-sample simulations in Series I. It can be seen from panel (a) in Figure 5 that all methods (Allele score methods, sisVIVE methods and *CIV* methods) form strong instrumental variables (F-statistics > 10), and *CIV* methods create instrumental variables that have weak correlation with $\mathbf{Z}$ (panel (b)). Allele score methods form the strongest instrumental variables among all competitors; however, their causal effect estimations are biased (panel (c)) because of the relatively large correlations of allele scores and $\mathbf{Z}$ (panel (b)). sisVIVE methods also lead to strong instruments (panel (a)) with unbiased causal effect estimation (panel (c)), however, the correlation with $\mathbf{Z}$ is a bit larger than *CIV* (panel (b)). In summary, *CIV* methods (including *CIV*, *CIV_boot* and *CIV_smooth.sel*) form instrumental variables that have reduced correlation with pleiotropic phenotype $\mathbf{Z}$, and lead to unbiased causal effect estimates in simulation Series I.

In Figure 6 and Figure 7 where results are shown from Series II with $Z \to X$ and $X \to Z$ respectively, all the variants of *Allele* methods and *sisVIVE* methods do not perform well. There are pleiotropic correlations of allele scores with $\mathbf{Z}$ and biased causal effect estimates from *Allele* methods. The *sisVIVE.exo* method selects instrumental variables that are relatively weak, and *sisVIVE.adj* method select pleiotropic instruments. Both*sisVIVE* methods give slightly biased causal effect estimates. In contrast, all *CIV* instruments show lower pleiotropic correlations with $\mathbf{Z}$ than *sisVIVE* or *Allele* methods, and the resulting causal effect estimates are less biased than the *Allele* and *sisVIVE* methods.

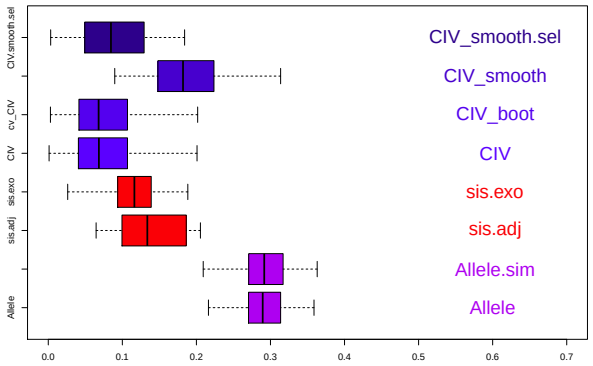
However, in Figure 8 where results are shown for Scenario 9 in Series III where $p = 100$, i.e. many weak instruments, the *CIV* methods do not perform well compared to Allele score methods. It can be seen from panel (a) in Figure 8 that all *CIV* methods form relatively weak instrumental variables compared to *Allele* score methods. *CIV* and *CIV_boot* result in instrumental variables with smaller correlations with $\mathbf{Z}$. Of the *CIV* methods, *CIV_smooth*, with or without selection, results in a stronger instrument with comparable strength to *sisVIVE*, and with the advantage of lower correlations with $\mathbf{Z}$ than sisVIVE, although the pleiotropic correlations with $\mathbf{Z}$ are no longer zero after smoothing of *CIV*. Here, the *Allele* methods perform extremely well, forming strong instruments that are uncorrelated with $\mathbf{Z}$. It is worth noting that any weighted score based on a non-sparse weight will have a weak correlation with $\mathbf{Z}$, since there are many valid but weak instruments (75/100) and the variance of $\mathbf{Z}$ is relative large in this case. Similar results were obtained for other scenarios in Series III simulations. The weak instruments from *CIV_boot* and *CIV* then result in large variability in the causal effect estimates (panel (c)). In conclusion, one limitation of *CIV* methods lies in the fact that the strength of *CIV* instruments may be excessively reduced when there are many weak instruments present in $\mathbf{G}$ ($F < 10$).

Figure 5: Boxplots of instrument strength measurements (a), correlation between new instruments ($\mathbf{G}^*$) and $\mathbf{Z}$ (b), and causal effect estimation bias (c) from a two-sample set-up, with $p = 9, \alpha_x = \alpha_z = 1$ instruments across 200 simulations in series I, when true $\beta = 0$.

(a) F-statistics of $\mathbf{X} \sim \mathbf{G}^*$ regression (in the second sample).



(b) $\mathbf{G}^* - \mathbf{Z}$ correlation (in the second sample).



(c) causal effect estimation bias (in the second sample).

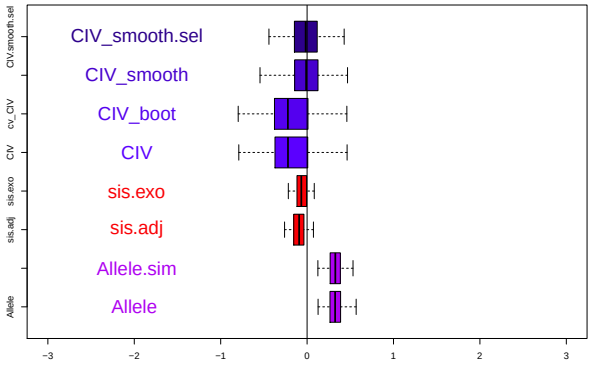
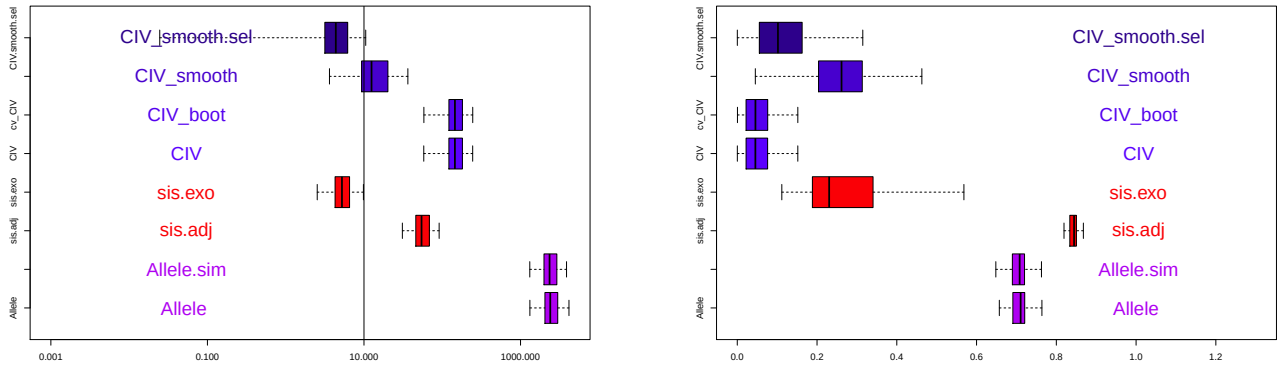


Figure 6: Boxplots of instrument strength measurements (a), correlation between new instruments ($\mathbf{G}^*$) and $\mathbf{Z}$ (b), and causal effect estimation bias (c) from a two-sample set-up, with $\mathbf{Z} \to \mathbf{X}$ and $\alpha_x = \alpha_z = 0.5$ across 200 simulations in series II, when true $\beta = 0$. A vertical line of $F = 10$ is drawn on the F-statistics plot.

(a) F-statistics of $\mathbf{X} \sim \mathbf{G}^*$ regression (in the second sample).

(b) $\mathbf{G}^* - \mathbf{Z}$ correlation (in the second sample).



(c) Causal effect estimation bias (in the second sample).

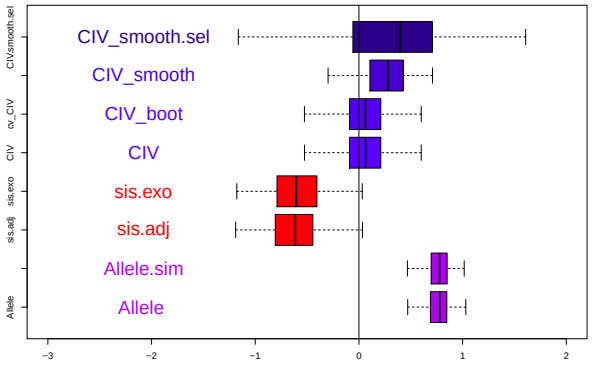
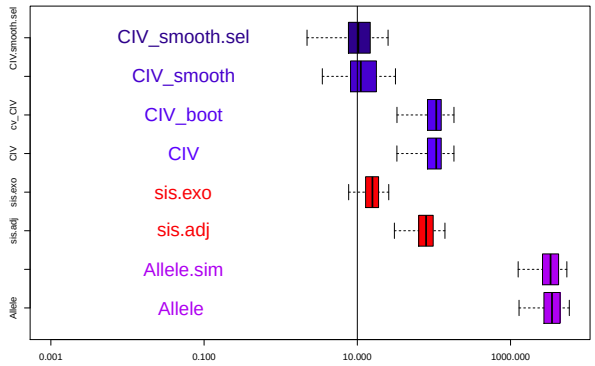
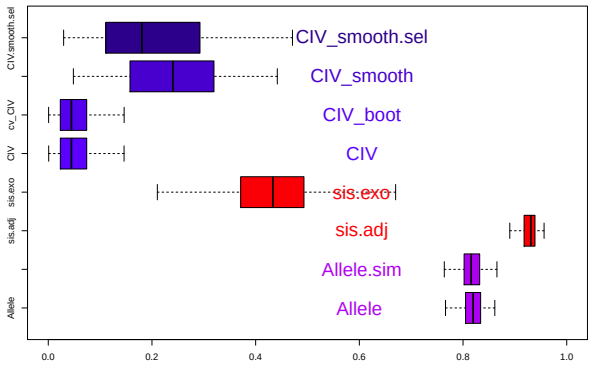


Figure 7: Boxplots of instrument strength measurements (a), correlation between new instruments ($\mathbf{G}^*$) and $\mathbf{Z}$ (b), and causal effect estimation bias (c) from a two-sample set-up, with $\mathbf{X} \to \mathbf{Z}$ and $\alpha_x = \alpha_z = 0.5$ across 200 simulations in series II, when true $\beta = 0$. A vertical line of $F = 10$ is drawn on the F-statistics plot.

(a) F-statistics of $\mathbf{X} \sim \mathbf{G}^*$ regression (in the second sample).



(b) $\mathbf{G}^* - \mathbf{Z}$ correlation (in the second sample).



(c) Causal effect estimation bias (in the second sample).

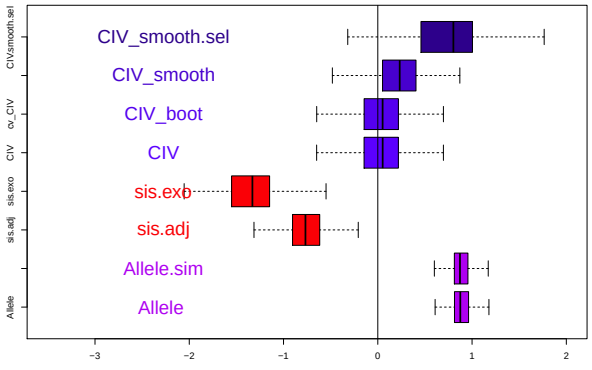
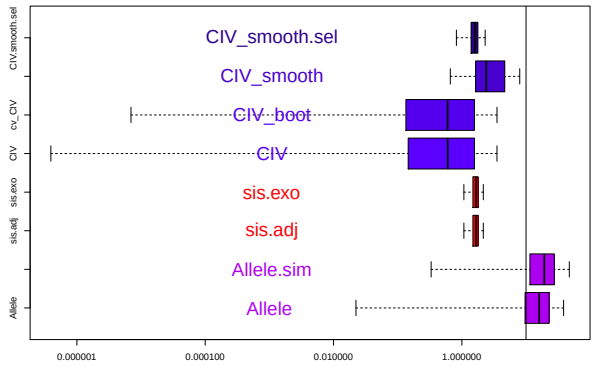
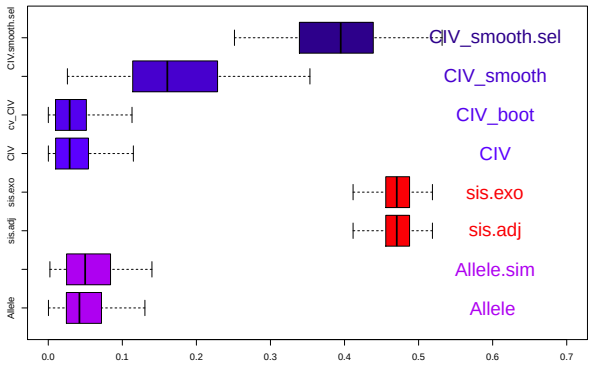


Figure 8: Boxplots of instrument strength measurements (a), correlation between new instruments ($\mathbf{G}^*$) and $\mathbf{Z}$ (b), and causal effect estimation bias (c) from a two-sample set-up, with $p = 100, \alpha_x = 0.06, \alpha_z = 0.06$ instruments across 200 simulations in series III, when true $\beta = 0$. A vertical line of $F = 10$ is drawn on the F-statistics plot.

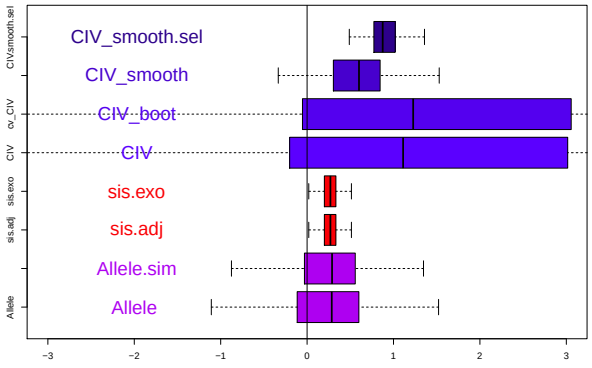
(a) F-statistics of $\mathbf{X} \sim \mathbf{G}^*$ regression (in the second sample).



(b) $\mathbf{G}^* - \mathbf{Z}$ correlation (in the second sample).



(c) causal effect estimation bias (in the second sample).

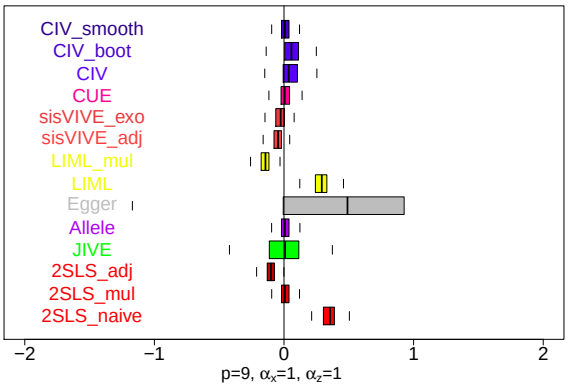


4.3.2 Causal Effect Estimation

Figure 9 shows causal effect estimation results from Series I in a one sample set-up. It can be seen that the estimates from *2SLS_adj* have consistent negative bias and *2SLS_naive* have positive bias, while *LIML* methods seem to show either a positive or negative bias, depending on whether or not adjustments for **Z** are included. The causal effect estimates from *sisVIVE* and *Egger* methods are slightly biased, while *JIVE* and *Egger* methods have large variability. *CUE*, *Allele* and *2SLS_mul* have consistently unbiased causal effect estimates and little variability across replications. All 3 flavors of *CIV*, especially *CIV_smooth*, also perform well with little bias and variability that is comparable to that of *2SLS_mul*, *Allele* or *CUE*. In summary, the causal effect estimation bias is small and comparable for *2SLS_mul*, *Allele*, *sisVIVE*, *CUE* and all three *CIV* variants.

In fact, *CIV_smooth* seems to perform the best among three *CIV* variants, with less bias and variability than the others. *CIV_smooth* method also has slightly less bias than *sisVIVE*, the reason is that *sisVIVE* does not consistently select valid instruments across replications according to our observations. Here, the regression adjustments for **Z** used by *2SLS* is biased since the adjusted instrumental variables (regression residuals) violate the assumption (A1). Note that *Allele* score method performs well here in one-sample analysis in contrast to its biased result in two-sample analysis (8). In conclusion, using the smoothed solution of *CIV*, the impact of pleiotropic genotypes can be reduced and the causal effect estimates is consistently less biased than its closest competitors (*Allele* score and *sisVIVE* methods).

Figure 9: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series I, with true causal effect $\beta = 0$ across 200 simulations.



The results of one-sample simulations in Series II, when there are causal relations between **Z** and **X**, are found in Figures 10 and Figure 11. In Figure 10, we have a causal pathway from **X** to **Z**. Conditioning on **Z** directly could lead to collider bias, but since some SNPs are not directly associated with **Z**, *CIV* methods may be able to alliviate the bias by removing them. In Figure 11, the causal effect estimation results of all methods when there is a causal pathway from **Z** to **X** are shown. In this case, *CIV* methods are expected to eliminate the pleiotropic bias since they are designed to remove the pleiotropic genotypes and restore

the assumption (A2). In both these figures, particularly when the instruments are strong, there are large biases associated with *sisVIVE*, *LIML*, *2SLS_naive* and *2SLS_adj* methods. *sisVIVE* method has a particularly large bias in Figure 10. *CUE* continues to show good performance, as does *2SLS_mul* method. There is little bias for *Allele*, *Egger*, or *JIVE* but with weak instruments, these methods are quite variable.

All 3 flavours of *CIV* perform well here in Figures 10 and Figure 11, with little bias and variability that is comparable to that of *2SLS* or *CUE*. *CIV_smooth* continue to perform the best of the three *CIV* variants, with slightly less bias than the other versions.

Figure 10: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series II, when there are causal effects $\mathbf{X} \rightarrow \mathbf{Z}$ and $\beta = 0$ across 200 simulations.

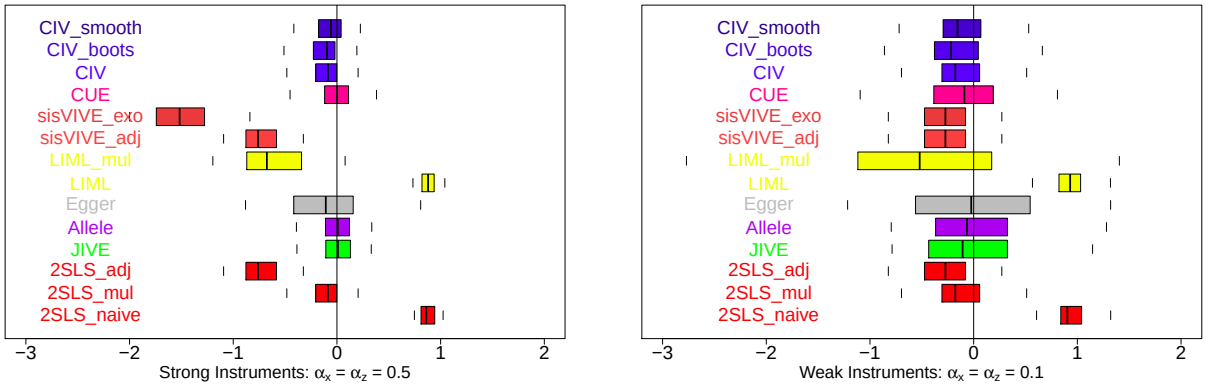
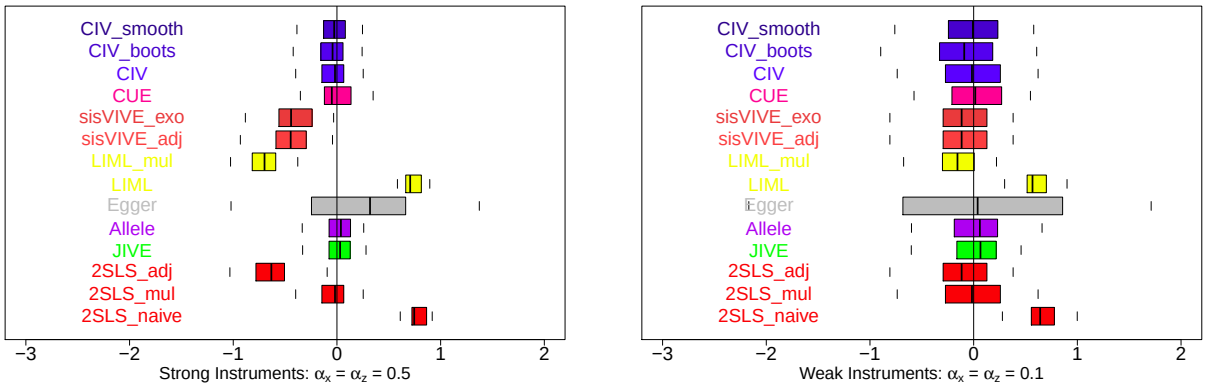


Figure 11: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series II, when there are causal effects $\mathbf{Z} \rightarrow \mathbf{X}$ and $\beta = 0$ across 200 simulations.



Figures 12, 13 and 14 show causal effect estimation results from Series III, for different numbers of SNPs, p, in one sample set-ups. In all 3 Figures, it can be seen that the *JIVE* estimates have huge variability across simulations. *2SLS_naive* and *Egger* methods show consistent and large positive bias, and the *LIML* methods seem to show either a positive or negative bias, depending on whether or not adjustments for $\mathbf{Z}$ are included. For *CUE* and

Allele score methods, the variability across simulations tends to increase with p. However, the variability of *CIV*, *sisVIVE* and *2SLS* appears to decrease as p increases.

The only method that gives an unbiased causal effect estimate in all of 3 Figures 12 - 14 is *CUE*, which was explicitly designed for situations with many weak instruments. Bias is also small for *LIML* (although estimates are variable), which agrees with the general claim that generalized methods of moments are less biased than *2SLS* when using many weak instruments (Hansen et al., 2008; Davies et al., 2015).

The causal effect estimation bias is small and comparable for *2SLS_adj*, *2SLS_mul*, *sisVIVE* and *CIV_smooth*. The similarity between *CIV_smooth* and *sisVIVE* is expected. These two methods both perform embedded feature selection, and hence are likely to underperform when only selecting a few weak instruments from a large set of candidates. The comparable performances of *2SLS_adj* and *2SLS_mul* here are probably due to the design of the simulations for Series III; there is no causal relationship between $\mathbf{X}$ and $\mathbf{Z}$ and all genotypes are weak instruments. Hence, the adjustments for $\mathbf{Z}$ used by *2SLS* work well. The biases for *Allele* score method and *CIV* without smoothing are larger; indicating that in general summarizing many weak instruments into one instrument does not provide substantial benefit here.

Figure 12: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series III, with $p = 9$ instruments across 200 simulations, when true $\beta = 0$. The panels display results for different values of α_x and α_z.

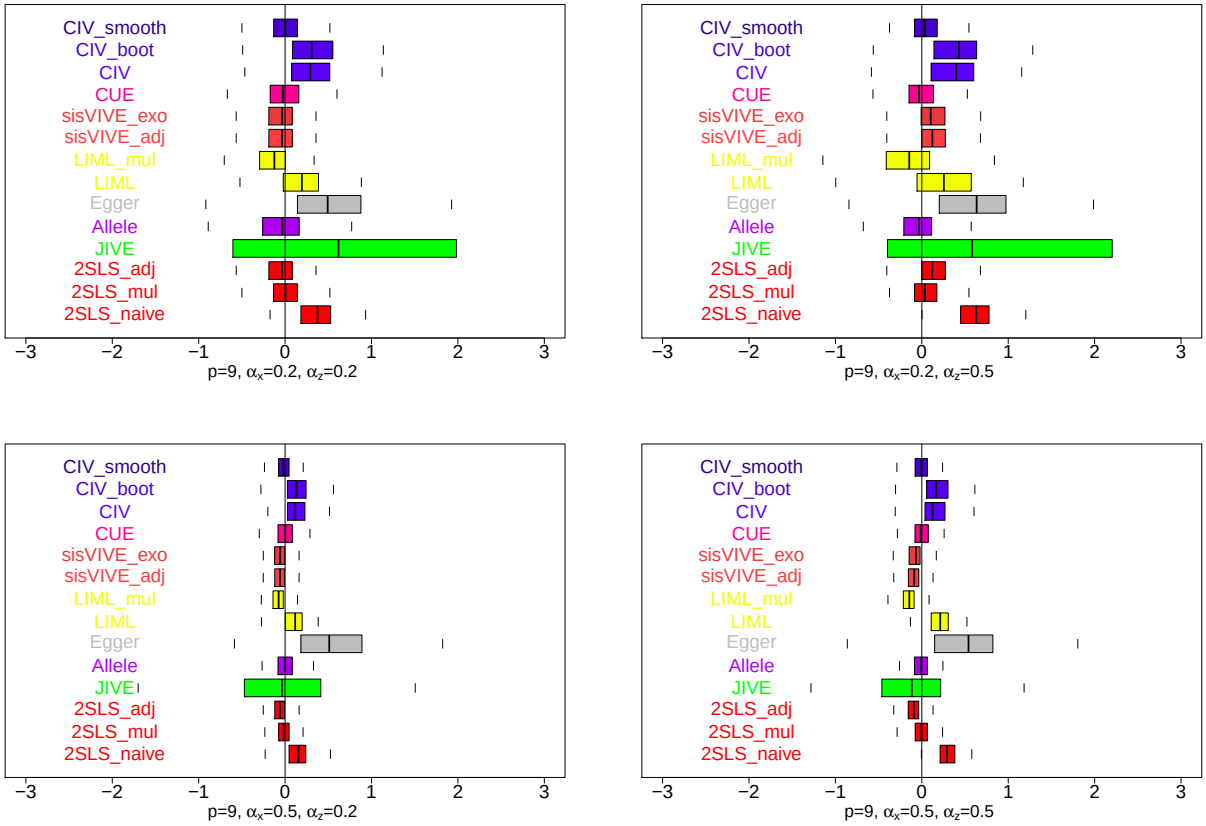


Figure 13: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series III, with $p = 25$ instruments across 200 simulations, when true $\beta = 0$. The panels display results for different values of α_x and α_z

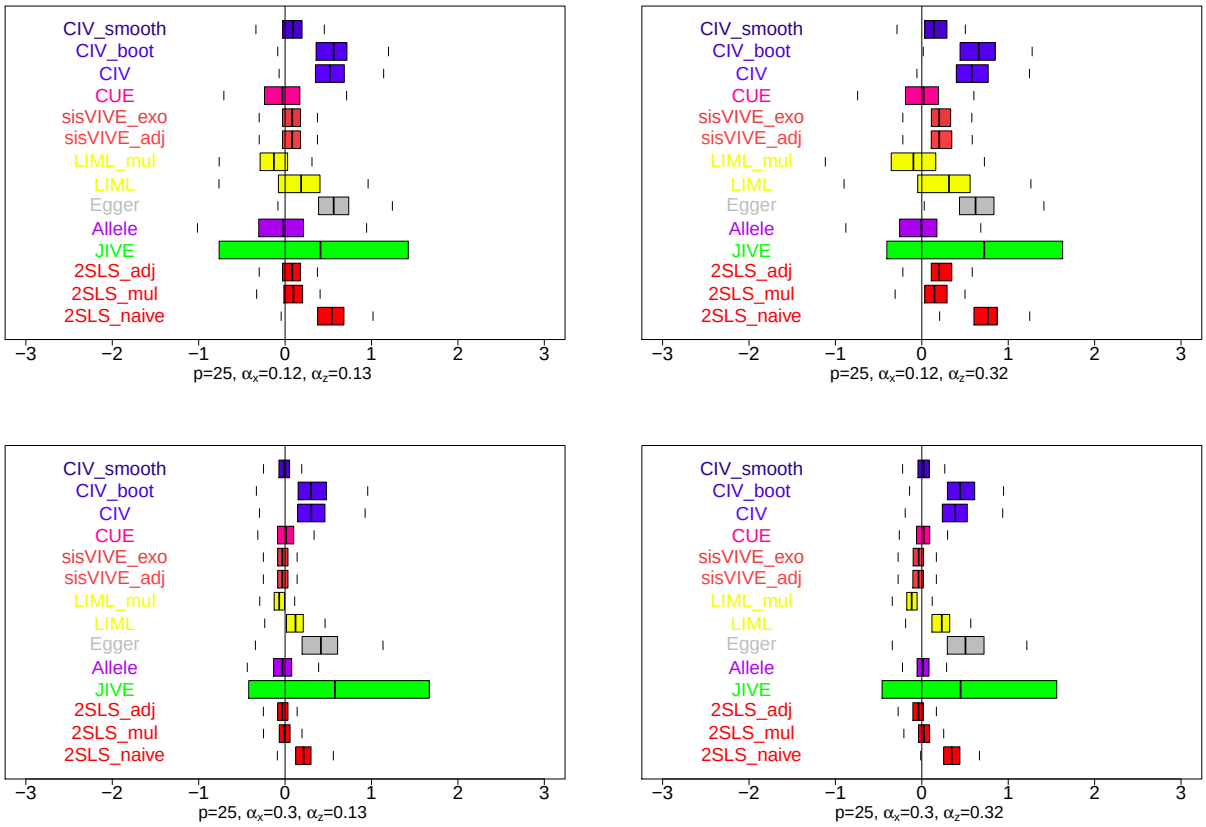
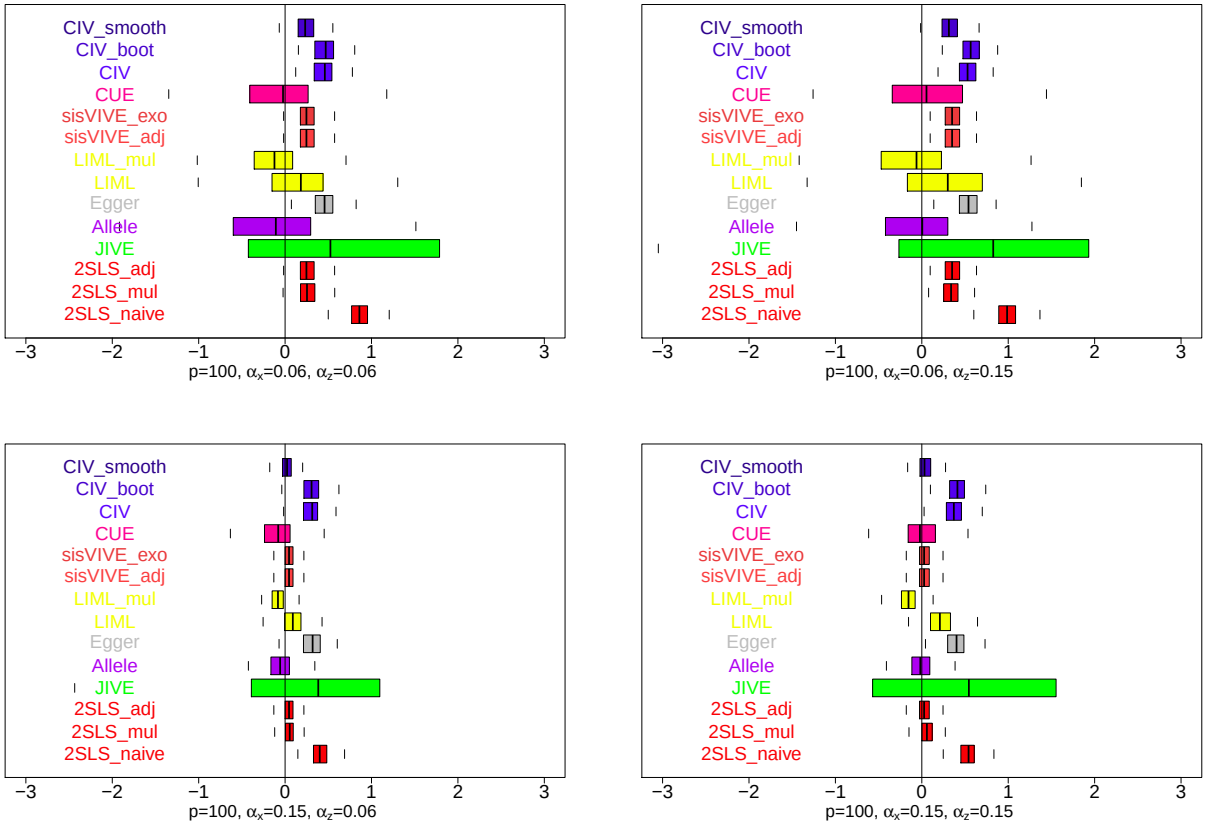


Figure 14: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series III, with $p = 100$ instruments across 200 simulations, when true $\beta = 0$. The panels display results for different values of α_x and α_z



4.4 Simulation Series IV : Selection of Valid Instruments

A fourth series of simulations was designed to explore the performance of valid genotype selection in one-sample setup. Here, $\mathbf{G}$ contained not only some strong instruments (some of which may have pleiotropic effects), but also many weak instruments. We referred to these weak instruments as "unnecessary genotypes"; perhaps they were put into the pool of instruments since they shared the same genetic pathway with a strong SNP, or their contributions were so weak that they were not adding information. Use of such SNPs as potential instruments jeopardized the validity of the whole set of instruments, $\mathbf{G}$, and causal effect estimates may be biased with some instrumental variable methods. Of particular interest here is the comparison of performance between CIV_smooth and sisVIVE, which both do genotype selection.

Data were simulated for $n = 500$ observations containing data $(\mathbf{G}, \mathbf{X}, \mathbf{Y}, \mathbf{Z})$. For each observation, $p = 50$ independent SNPs were generated with a minor allele frequency of 0.3, divided into two subsets G_{UN} (30 unnecessary SNPs) and $G_R = G/G_{\text{UN}}$ (20 relevant SNPs). We also sampled 10 SNPs (as G_{R_z}) without replacement in G_R to be pleiotropic, i.e. also associated with $\mathbf{Z}$. For each individual $i \in (1, ..., n)$, the phenotypes (X_i and Z_i) and response (y_i) were simulated as described below:

$$\begin{aligned} x_i &= 2 \sum_{j \in G_R} G_{ij} + 0.1 \sum_{j \in G_{\text{UN}}} G_{ij} + u_i + \epsilon_{x,i}, \\ z_i &= 0.5 \sum_{j \in G_{R_z}} G_{ij} + u_i + \epsilon_{z,i}, \\ y_i &= x_i + z_i + u_i + \epsilon_{y,i}, \end{aligned} \tag{14}$$

where G_{R_z} was the subset of relevant SNPs that were associated with $\mathbf{Z}$, and u_i was a confounding factor related to $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{Y}$. We generated u_i and $\epsilon_\cdot$ from the standard Gaussian distribution. These SNPs with $\alpha_x = 0.1$ were all weak instruments and not actually unrelated with x_i. We used the term "unnecessary" here only because the selection of these SNPs should be largely unnecessary for the causal effect estimation purpose. We repeated the simulation 200 times and aggregated the results.

The selection of genotypes using *CIV_smooth* was conducted in two different ways. In the first approach, we selected the "top" feature, i.e. the feature corresponding to the maximum absolute magnitude of $\mathbf{c}$, in each converged *CIV_smooth* solution. Then we combined all "top" features across different solutions and recorded the selection frequency of each feature. This method was referred as *CIV_smooth.top*. The number of features selected in this way would be limited; since many solutions only had slightly different weighting values of SNPs, but the 'top" feature in each solution could be the same. The second selection strategy was based on *CIV_smooth.sel*. We first labeled all SNPs j as "relevant" in a weighting vector $\mathbf{c}$ if $c_j \geq 0.2 \max |\mathbf{c}|$. Then we counted the number of times each SNP j was selected, and used the averaged value across solutions as the relative selection index for all genotypes.

The results of genotype selection were shown in Table 5. In general *CIV_smooth* selected

fewer of the pleiotropic genotypes in $\mathbf{G}$, as well as fewer unnecessary genotypes. For example, *CIV_smooth* (using only the top feature) selected 3.03 valid genotypes which accounted for 52% of all selected features. This means on average 52% of the top features in all converged modes $\mathbf{c}$ corresponded to valid genotypes, while sisVIVE with either the adjusted or the exogenous $\mathbf{Z}$ methods only achieved a sensitivity below 20%. The proportion of irrelevant features and pleiotropic features were also substantially lower when using the *CIV_smooth* methods. In conclusion, the valid instrument selection performances from *CIV_smooth* were substantially better than sisVIVE methods.

	Valid Genotypes	Pleiotropic Genotypes	Unnecessary Genotypes
CIV_smooth.top	3.03 (52%)	0.65(11%)	2.14(36%)
CIV_smooth.sel	7.96 (25%)	5.47 (17%)	18.31 (57%)
*sisVIVE*_adj	10.00 (20%)	10.00 (20%)	30.00 (60%)
*sisVIVE*_exo	6.83 (15%)	6.25(14%)	31.42 (70%)

Table 5: Simulation Series IV with unnecessary genotypes: feature selection results for CIV_smooth and sisVIVE among SNPs with different kinds of associations and across different methods. The table shows the average number (proportion) of selected features from different methods, averaging over 200 simulated data sets. *CIV_smooth.top* means only the SNP with highest $|c_j|$ were selected and used as instrument for MR analyses.

5 Data Analysis: Alzheimer's Disease

Alzheimer's disease (AD) is a chronic neurodegenerative disorder that causes a slow decline in memory and reasoning skills. It is well known that biomarkers including cerebrospinal fluid tau protein (CSF-tau) and CSF amyloid beta-protein ending at amino acid position 42 (CSF-Aβ 1-42) are reliable measures of AD progression (Iturria-Medina et al., 2016). Recently, other biomarkers such as glucose metabolism and neural functional activity have been added while exploring the mechanisms underlying late-onset Alzheimers disease (LOAD) using multi-factorial data analysis (Iturria-Medina et al., 2016). However, at this point, there is still some uncertainty whether the changes in these biomarkers "cause" AD progression or are simply associated with AD progression.

We have used instrumental variable methods on data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) to try to disentangle causal relationships for AD. ADNI began in 2004 and included 200 subjects with early AD, 400 subjects diagnosed with mild cognitive impairment (MCI), and 200 elderly control subjects. Biospecimens, including blood, urine, and cerebrospinal fluid (CSF) were collected from participants. FDG PET imaging was performed on participants within two weeks before or after the in-clinic assessments of memory composite scores. The global and regional standardized uptake value ratios (SUVR) of each subject were recorded after each scanning. Genotyping and sequencing data were also available for all subjects obtained from Illumina Human610-Quad BeadChip. Further details of of protocols and procedures of this data was available on the Image Data Archive (IDA).

The outcome (AD status) here is a binary variable, and MR methods assume linear additivity and are not designed for the situation of a binary outcome (Didelez and Sheehan, 2007). Alternatively, under more restrictive parameter assumptions (Vansteelandt et al., 2011), logistic regression and log-linear regression can be used in the second stage of MR analysis to estimate a causal risk ratio (CRR) when response is binary (Clarke and Windmeijer, 2012; Burgess et al., 2014). If only one exposure ($\mathbf{X}$) is analyzed with one binary outcome under linear assumption in the second stage, then any bias towards the null in the causal effect estimate would be largely due to the impact of confounding factors (Palmer et al., 2008).

A very important limitation of MR analysis in ADNI data is the retrospective nature of ADNI study design. Ascertainment in ADNI was retrospective by disease status, and therefore, instruments that would be valid for a prospective study design may not remain valid after retrospective sampling in the ADNI data (Didelez and Sheehan, 2007). Specifically, the estimated first stage ($\mathbf{G}$-$\mathbf{X}$) association from case-control samples may be biased relative to the true association in a general population sample (Tapsoba et al., 2014; Tchetgen Tchetgen, 2013). If the disease being studied is rare, it is possible to use only the control samples in a first stage regression, in MR methods where two-stage methods are appropriate (Lin and Zeng, 2009). Therefore, since we realize that MR analysis of the ADNI samples is not ideal, we use this dataset simply to illustrate performance of our methods, and not to make strong causal statements. Furthermore, we only present results where the first stage associations are estimated from the controls.

5.1 Outcome, Exposures and Instruments

Outcome $\mathbf{Y}$***:*** A subject is either from control group or diagnosed with MCI or AD. We combine AD and MCI subjects into the group with the same response variable. We collected $n = 491$ subjects including 151 controls with outcome $Y = 0$ and 340 patients with outcome $Y = 1$.

Exposures $\mathbf{X}$***:*** We are interested in estimating the causal effect of several exposures/biomarkers including CSF amyloid beta-protein (Aβ) (X_1), Phosphorylated Tau Protein (Ptau) (X_2), Total Tau protein (Ttau) (X_3) and FDG_SUVR (X_4) on AD progression. It is well known that the isoforms of Apolipoprotein E (ApoE), a class of apolipoprotein that mediates cholesterol metabolism, have effects on both Amyloid beta aggregation and Tau protein phosphorylation, and thus there may exist pleiotropic effects in this case. Natural logarithm scales of Ttau and Ptau are used to obtain X_2 and X_3. All exposures (X_k, k=1,...,4) (and also outcome $\mathbf{Y}$) are adjusted before analysis with covariates including age, sex and education. Profiles of the subjects are summarized in Table 6.

Instruments $\mathbf{G}$***:*** For each of the exposure $X_k, k = 1, ..., 4$, the strongly associated SNPs reported by the NHGRI-EBI Catalog of published genome-wide association studies (Burdett et al., 2016) were collected from the ADNI Imputed Genotype data. The missing samples were imputed to the 1000 Genome Project utilizing the same protocol for the ROS/MAP and AddNeuroMed study. When there were very highly correlated ($\rho \geq 0.8$) sets of SNPs,

	Number	Age (years) (mean ± SD)	Gender(M/F)	education (years) (mean ± SD)
Control	151	75.93 ± 5.86	86/65	16.3+2.7
MCI/AD	340	74.08 ± 7.63	212/128	15.89 ± 2.92
MCI	277	73.64 ± 7.53	173/104	16.03 ± 2.81
AD	63	76.03 ± 7.78	39/24	15.27 ± 3.31

Table 6: Characteristics of subjects studied in ADNI.

we kept only one representative SNP of each correlated sets. Specifically, we first built a genotype subset containing these SNPs with all pairwise correlations (ρ <0.8). Then we added SNPs into the feature subset one at a time and would reject the SNP if its pairwise correlations with the selected subset were high ($\rho \geq$0.8). The final SNP set was then further reduced with univariate feature selection using significant F-statistics ($p \leq 0.05$). Hence, reduced sets of SNPs containing 20 SNPs for Aβ (X_1), 8 SNPs for Ptau (X_2), 5 SNPs for Ttau (X_3) and 25 SNPs for glucose metabolism (X_4) were retained for use with Mendelian randomization methods.

5.2 Mendelian Randomization Analysis

The assumption (A1) of Mendelian randomization stated that the SNPs must be associated with biomarkers of interest. Strong instruments with F-statistics bigger than 10 are usually preferred in MR applications. The F-statistics for instrument strength of each biomarker here (Aβ, Ptau, Ttau, SUVR) are 11.86, 12.40, 3.89 and 5.20 respectively, indicating strong instruments only for Aβ and Ptau. We also perform Sargan test for over-identification (Baum et al., 2003) to test the MR assumption (A2) and (A3). The p-values of the Sargan test are 0.22, 0.01, 0.81 and 0.92 for $X_k, k = 1, ..., 4$, implying the existence of invalid instruments in $\mathbf{G}$ for Mendelian randomization for Ptau (X_2) on AD progression ($\mathbf{Y}$). The reason is because the selected SNPs that are strongly associated with Ptau have even stronger associations with Aβ. Hence this creates a pleiotropy problem for causal inference of Aβ and Ptau on AD progression using similar groups of genotypes. More information about the associations of instruments $\mathbf{G}$ with Aβ (X_1) and Ptau (β) are shown in the figure 15.

Mendelian randomization was performed to evaluate the potential causal effects of variability in all biomarkers (X) on the AD progression (Y) in a two-sample set-up. First we used only the control samples to obtain weights with the three applicable methods (*Allele*, *sisVIVE* and *CIV* and their variants). In the second step we constructed instrumental variables using these obtained weights with genotype information on the whole sample, and inferred causal effects of each biomarker X_k on AD progression (Table 3) while treating other biomarkers as secondary phenotypes, if applicable. In this way, the retrospective nature of ADNI is respected, if we assume that the control sample is similar to the whole population from which the individuals were drawn. We can only include *CIV*, *Allele scores* and *sisVIVE* in the two-sample analysis because not all methods can be adapted to a two-stage approach

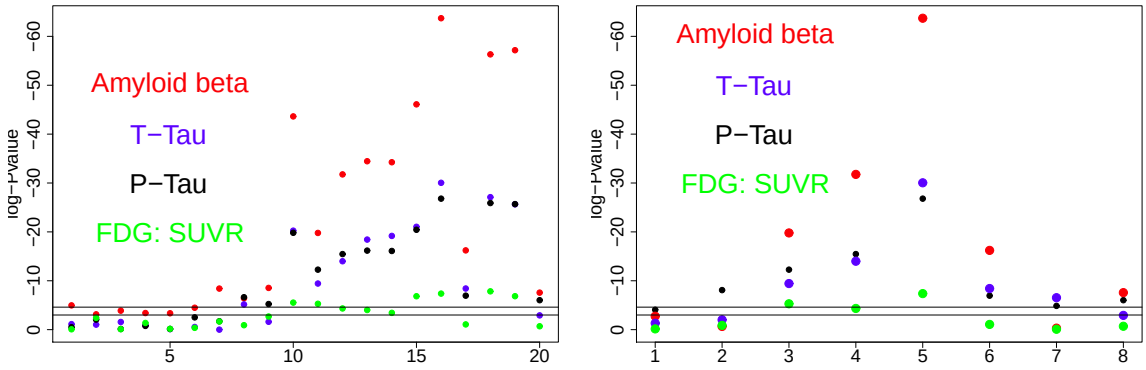


Figure 15: The associations between instruments selected for phenotypes and phenotypes of interest. Left: log-Pvalues for association test between the selected SNPs (for Aβ) and Aβ (X_1). Right: log-Pvalues for association test between the selected SNPs (for Ptau) and Ptau (X_2).

here. Hence, the results of this MR analysis are illustrative, and require further validation, ideally on a separate prospective study.

5.3 Results

The 95% confidence intervals of the causal effect estimates for all four biomarkers obtained from two-sample analyses are reported in Figure 16 and Figure 17. All variants of *Allele score* methods and *CIV* methods (except *CIV_boot*) identified significantly negative causal effect of Amyloid beta 1-42 protein levels on AD progression. Only the *Allele score* methods and *CIV_smooth.sel* showed a significantly positive causal effect of Ptau protein levels on AD progression. Furthermore, only Allele.sim and *CIV_smooth.sel* methods identified significant causal effects of Ttau protein levels on AD progression. For glucose metabolism levels none of the instrumental variables methods showed significant causal effect estimates.

The observation of significant causal impact for Aβ and Tau protein on Alzheimer's disease study is consistent with some previous publications. In fact, multiple observational studies have reported decreasing Aβ and increasing Tau levels in cerebrospinal fluid of patients with Alzheimer Disease compared to normal control subjects (Sunderland et al., 2003; Maruyama et al., 2001). There is no conclusive evidence to support a significant causal relationship between glucose metabolism levels and AD progression. However, limited research has been conducted to evaluate the causal relationships of CSF Aβ proteins, Ptau, Ttau and glucose metabolism on AD using Mendelian randomization.

The range of causal conclusions from different instrumental variable methods with Amyloid beta and Ptau protein levels may be due to the existence of pleiotropy. In fact, most of the SNPs strongly associated with Ptau in this dataset were also strongly associated with Amyloid beta, and most of the associated SNPs were located in the APOE gene. As a result, some instrumental variable methods using these invalid instruments could be biased.

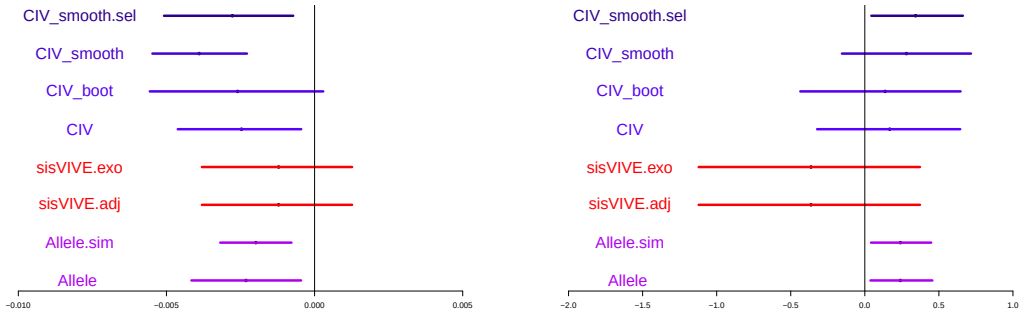


Figure 16: 95% bootstrapped confidence interval of causal estimation of Amyloid beta 1-42 protein levels (supposedly negative effects) and Ptau protein levels (supposedly positive effects) on AD progression using different instrumental variable methods in a two-sample set-up. Left: Amyloid beta protein levels on AD. Right: Ptau protein levels on AD.

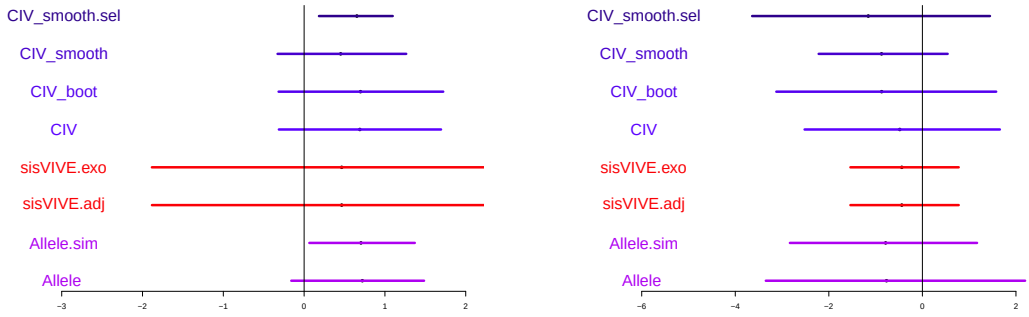


Figure 17: 95% bootstrapped confidence interval of causal effect estimation of Ttau protein levels (supposedly positive effects) and glucose metabolism on AD progression using different instrumental variable methods in a two-sample set-up. Left: Ttau protein levels on AD. Right: glucose metabolism levels on AD.

Weighted score methods, including allele score methods and *CIV_smooth.sel*, were able to reduce bias due to pleiotropy compared to *sisVIVE*. In conclusion, when pleiotropy exists, *CIV_smooth.sel* could select valid instruments and perform adjusted causal effect estimation to account for pleiotropy.

6 Discussion

In this paper we proposed constrained instrumental variable methods for causal inference when pleiotropy is suspected. We have presented the performance of our methods and compared with other popular methods in different simulation scenarios (i.e. standard pleiotropy, direct causal effects between phenotypes, weak instruments, valid and invalid instruments selection). To illustrate the performance of *CIV* methods in real data, we conducted MR analysis on data from Alzheimer's Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005) and demonstrated that the previously known associations for Amyloid beta 1-42 and Tau protein levels showed evidence for a causal relationship, even after making adjustments for other potentially-pleiotropic biomarkers.

One important feature of our *CIV* methods is that they find a balance between strength and validity in instrument construction. *CIV* solutions are designed to preserve maximum instrument strength while adjusting for the pleiotropic phenotype $\mathbf{Z}$. In general strong instruments will provide consistent causal effect estimates, while approximately valid instruments will reduce the pleiotropy-induced bias. When the instruments are strong in a standard pleiotropy problem (Section 4.2.1), our *CIV_smooth* method yields better estimates in both one-sample and two-sample setups, and outperforms *2SLS_adj*, *sisVIVE* methods, *LIML* methods, *Egger* regression, and *Allele* score with less bias. In simulations with a direct causal link between $\mathbf{X}$ and $\mathbf{Z}$ (Section 4.2.2), *CIV* methods yield strong instruments while retaining less pleiotropic genotypes, compared to *sisVIVE* and Allele methods. When the instruments are weak (Section 4.2.3), our *CIV_smooth* method performs slightly worse than *CUE* and *Allele* score in terms of consistence and bias; indicating that relatively valid but weak instrumental variables may still lead to more biased results. In simulation IV (Section 4.4) where we examined valid instrument selection, we showed that *CIV_smooth* has a higher rate of valid instrument selection compared to its closest competitor, *sisVIVE*.

Another advantage of the *CIV* methods is the separation of instrument construction and causal effect estimation. In fact, the construction of *CIV* only relies on the first stage pathway $\mathbf{G} \rightarrow \mathbf{X}$. Then, any estimation method for linear structural equation modeling can be applied to CIV_smooth instruments $\mathbf{G}^* \rightarrow \mathbf{X} \rightarrow \mathbf{Y}$ for causal inference. Due to this separation of first-stage and second-stage analysis, *CIV* (IV_smooth) and *Allele scores* can be trained and assessed on different datasets. In conclusion, *CIV* methods have substantial flexibility in terms of the model assessment and implementation of estimation methods.

Multiple solutions can be obtained with *CIV_smooth*–i.e. the solution to the problem (7) may not be unique. An example of this potential multi-modal problem is shown in Figure 18, where one simulated dataset from Series II was analyzed. The hierarchical cluster dendrogram shows the solution space for this simulation, demonstrating that there exist multiple

different *CIV_smooth* solutions. However, a principal component analysis of the converged solutions, across simulations, shows that in many simulations only 1 unique solution stands out. So although multiple distinct solutions do occur, often they are extremely similar to each other. To obtain our *CIV_smooth* estimator, we attempt to sample the possible solution space by starting our converging iterations from multiple initial points, and combining all converged distinct solutions into a matrix $\mathbf{c}$. This approach provides a spectrum of the possible instruments that achieve the maximized association and low correlation with $\mathbf{Z}$.

In our simulations, MR analyses were mostly restricted to the case of a single risk factor $\mathbf{X}$, although most of the methods mentioned above can be extended in some way to allow for a multivariate $\mathbf{X}$. *CIV* and *CIV_smooth* methods can be adapted to allow for multivariate $\mathbf{X}$ as seen in Equation 5, and the corresponding multiple solutions $\mathbf{c}$ can be used with multivariate 2SLS to infer the causal effect of $\mathbf{X}$ on $\mathbf{Y}$. *sisVIVE* is also flexible to the dimension of $\mathbf{X}$ as seen in Equation (4). Generalized moment methods (e.g. *LIML*, *CUE*) and 2SLS can also be adapted to handle multiple risk factors together. However, the nature of *Egger* regression restricts this approach primarily to univariate risk factor analysis; users can only analyze multiple risk factors one-by-one.

In simulations, we conducted two-sample analyses, by separating the weight construction process and causal inference process onto two different samples, to evaluate the properties of applicable instrumental variable methods. In such two-sample approaches, winner's curse, which may result in over-estimates of the predictive power of constructed instruments in one-sample analyses, is less likely to occur.

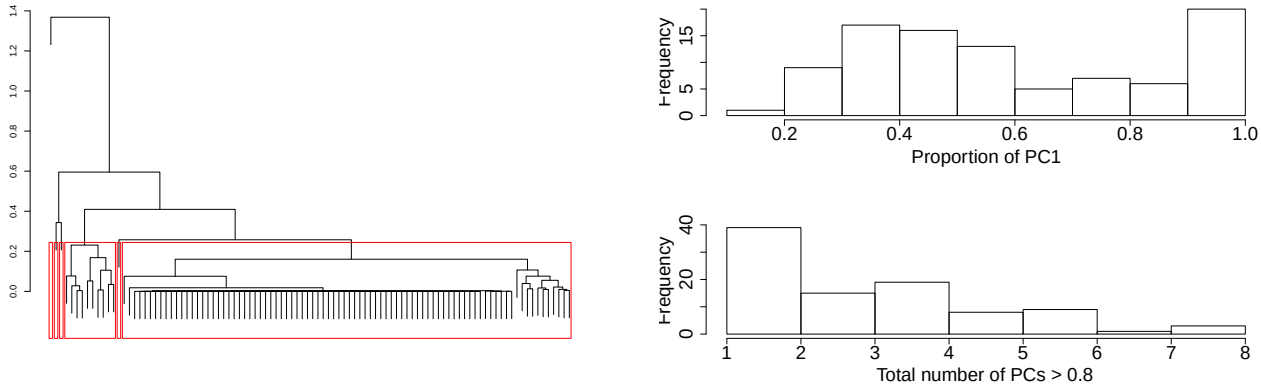
In the ADNI data analysis, we applied a two-stage approach in a similar way, by constructing instruments using only the control samples. These instruments were then used in the whole dataset to estimate causal effect of each individual biomarker while treating others as secondary phenotypes. However, we do realize this two-stage ADNI analysis is still not an ideal solution due to the retrospective nature of the sampling in this case-control study sampling. There are additional techniques, including inverse weighting with sampling probability (Monsees et al., 2009) and maximum likelihood (Lin and Zeng, 2009), that can be considered in the future to correct for the case-control sampling. It is also worth noting that we transformed the causal effect estimation problem with multiple phenotypes into multiple causal effect estimations, each with an individual phenotype – versus the rest – for comparison purposes, since only *CIV* can process multiple causal effect estimations simultaneously while both *Allele* and *sisVIVE* are designed for individual phenotypes. This "one versus the rest" strategy leaves room for improvement in future work. The result of ADNI analysis in this paper simply serve as a demonstration of our *CIV* methods, rather than a strong causal statement of Alzheimer's disease.

In conclusion, this paper proposed new approaches (*CIV* and *CIV_smooth* methods) for conducting Mendelian randomization analyses when the pleiotropy is observed. Under the assumption of linear structural equation models, these approaches can be used to implement valid instruments selection and adjust causal effect estimation when potential pleiotropic phenotypes are measured. Our methods (*CIV* and *CIV_smooth*) perform consistently in simulations and real data analysis. Hopefully, these methods will help to make the MR analyses a much more common practice even when pleiotropy is observed.

Figure 18: Cluster dendrogram (a) and principal component analysis (b) from a one-sample set-up, with $\mathbf{Z} \to \mathbf{X}$ and $\alpha_x = \alpha_z = 0.1$ across 200 simulations in Section 4.2.2.

(a) Cluster dendrogram of all (100) converged *CIV_smooth* solutions in one simulation. Red block denotes the identified hierarchical clusters using the number of cluster determined by the silhouette value.

(b) Top: The proportion of the top eigenvalue (among all values) of all *CIV_smooth* solutions, across 200 simulations. Bottom: The total number of principal components (with eigenvalue $\geq$ 0.8), across 200 simulations.



A Solution to the Constrained Instrumental Variable Problem

Let $M = X(X^\top X)^{-1}X^\top$ then $c^\top G^\top Xv = c^\top G^\top X(X^\top X)^{-1}X^\top Xv = c^\top G^\top MXv$,

$$\max_{c\in R^p, v\in R^r} c^\top G^\top Xv = \max_{c\in R^p, v\in R^r} c^\top G^\top MXv \leq ||c^\top G^\top M||\ ||Xv||,$$

where $||Xv|| = \sum_{i=1}^{n} |(Xv)_i|^2$ denotes the norm of vector Xv in the inner product space R^p. The equality holds if and only if Xv and MGc are collinear. Let $w = (G^\top G)^{\frac{1}{2}}c$ then the problem is equivalent to

$$\max_{w\in R^p} w^\top Aw$$

subject to conditions:

$$w^\top w = 1$$

$$B^\top w = 0$$

where $A = (G^\top G)^{-\frac{1}{2}}G^\top MG(G^\top G)^{-\frac{1}{2}}$ and $B = (G^\top G)^{-\frac{1}{2}}G^\top Z$.

If we have $\text{rank}(A) = p \geq \text{rank}(Z) = k$ (columns are uncorrelated), then there exists solution for w since this is a quadratic optimization problem with quadratic/linear constraints (Golub, 1973).

Consider the QR decomposition of B:

$$B = Q^\top \begin{bmatrix} R & S \\ 0 & 0 \end{bmatrix} \Delta \tag{15}$$

where R is a k by k upper triangular matrix with positive diagonal elements (thus invertible). Q is an orthogonal matrix. S is a k by $p-k$ matrix and Δ represents the column permutation matrix (Gu and Eisenstat, 1996) to ensure that the diagonal elements of R are positive and non-increasing, i.e.R is invertible. R is then unique under these conditions (Golub and Van Loan, 1996).

Now we let $w = Q^\top \begin{pmatrix} y \\ d \end{pmatrix}$ where $y \in R^k, d \in R^{p-k}$ and $QAQ^\top = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$.

The problem then becomes:

$$\max_{d \in R^{(p-k)}} d^\top A_{2,2}^\top d \tag{16}$$

subject to conditions:

$$d^\top d = 1$$

We now know that the solution for d is the eigenvector corresponding to the largest eigenvalue of $A_{2,2}$. There are at most $p-k$ eigenvectors.

In conclusion, when $n > p$ the (unique) solution of the constrained instrumental variable (CIV) is $Gc = G(G^\top G)^{-\frac{1}{2}} Q^\top \begin{pmatrix} 0 \\ d \end{pmatrix}$, where Q is an orthogonal matrix defined by (15) and d is an eigenvector defined by (16).

B Simulation III : Non-zero Null Hypothesis

Figure 19: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series III, with $p = 9$ instruments across 200 simulations, when true $\beta = 1$. The panels display results for different values of α_x and α_z.

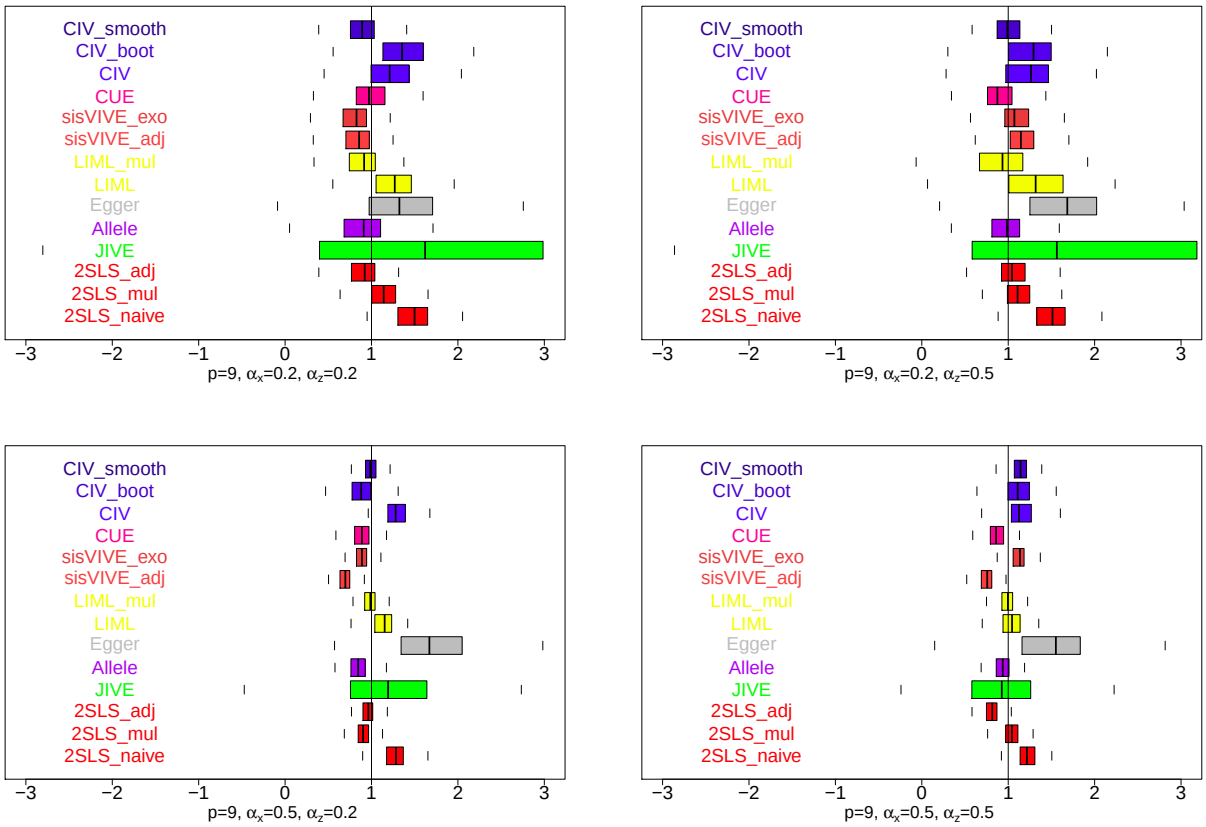


Figure 20: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series III, with $p = 25$ instruments across 200 simulations, when true $\beta = 1$. The panels display results for different values of α_x and α_z

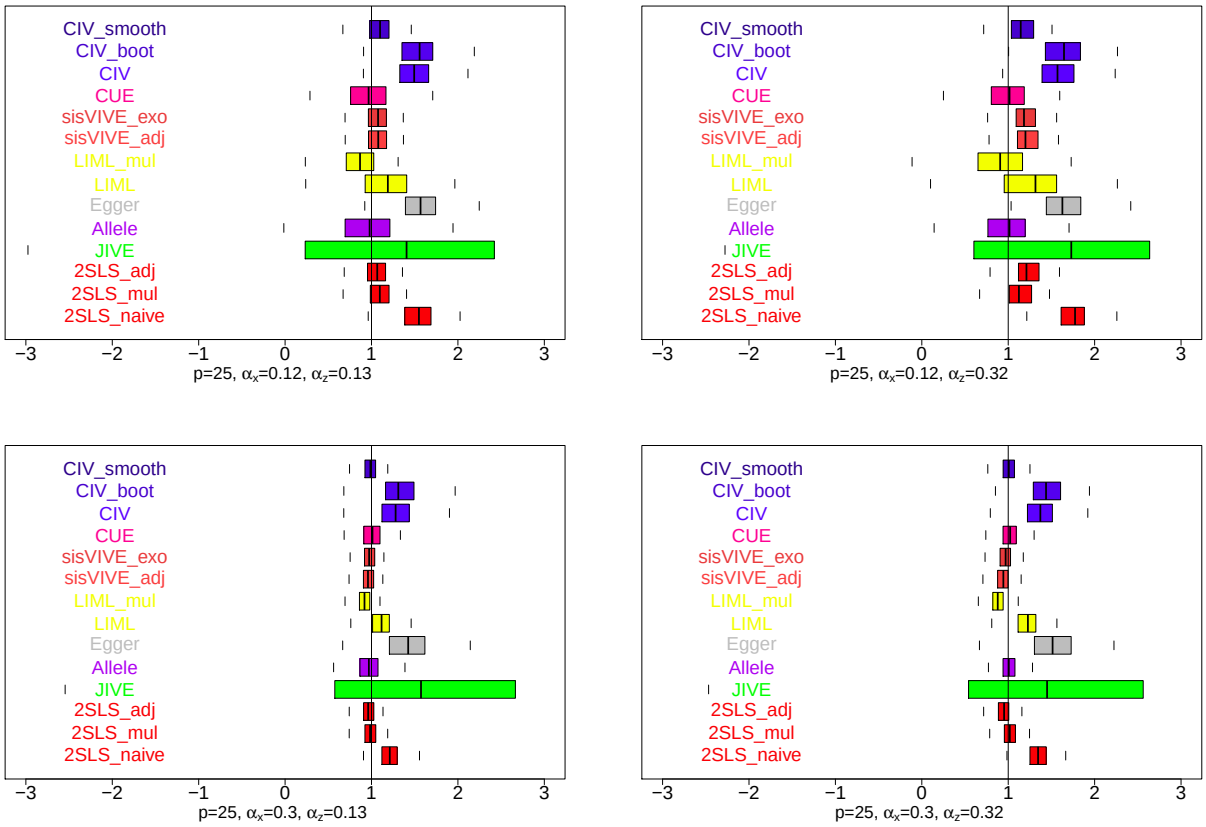
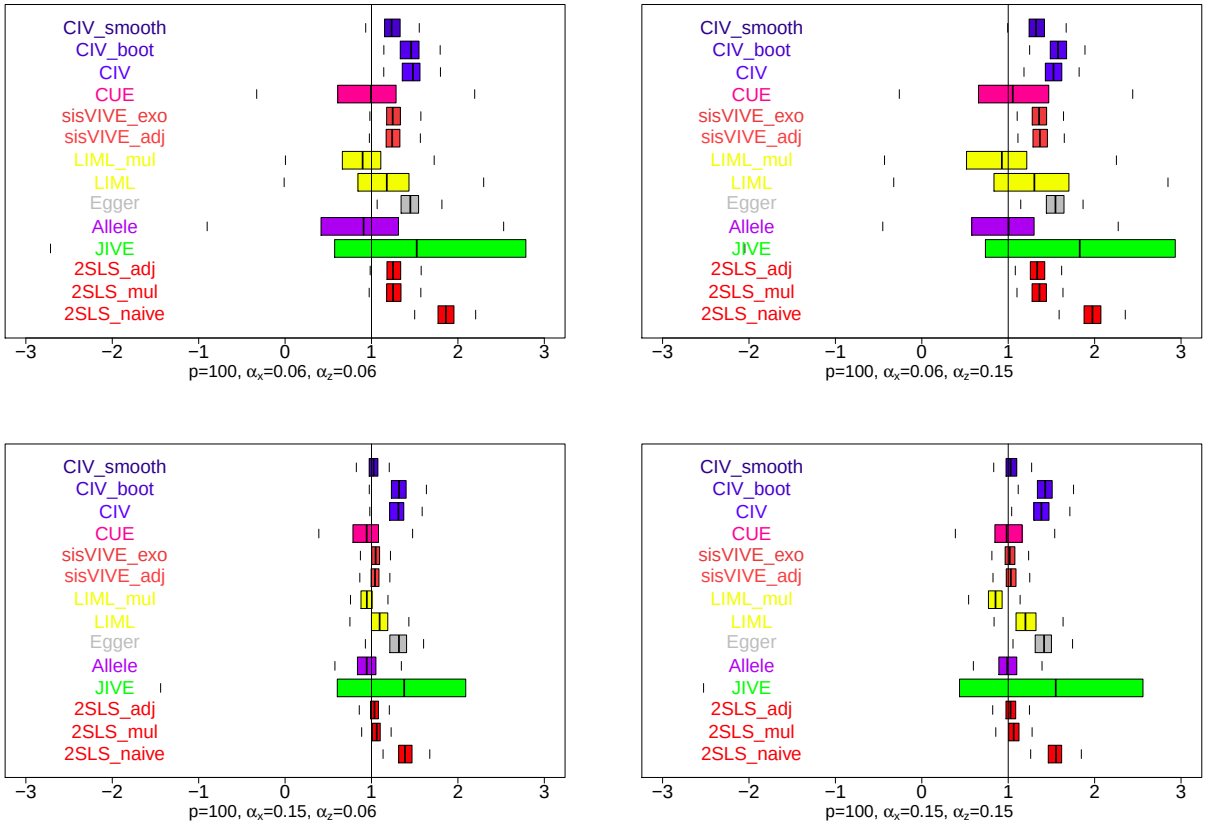


Figure 21: Boxplots of estimates of the causal effect estimates, β, from a one-sample set-up in simulation series III, with $p = 100$ instruments across 200 simulations, when true $\beta = 1$. The panels display results for different values of α_x and α_z



References

Joshua Angrist, Guido Imbens, and Alan B Krueger. Jackknife instrumental variables estimation, 1995.

Joshua D Angrist and Alan B Krueger. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, 87(418):328–336, 1992.

Jeffrey C Barrett, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, M Michael Barmada, et al. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nature genetics*, 40(8):955–962, 2008.

Christopher F Baum, Mark E Schaffer, Steven Stillman, et al. Instrumental variables and gmm: Estimation and testing. *Stata journal*, 3(1):1–31, 2003.

Colin B Begg and Madhuchhanda Mazumdar. Operating characteristics of a rank correlation test for publication bias. *Biometrics*, pages 1088–1101, 1994.

Sören Blomquist, Matz Dahlberg, et al. Small sample properties of liml and jackknife iv estimators: experiments with weak instruments. *Journal of Applied Econometrics*, 14(1): 69–88, 1999.

George J Borjas. Food insecurity and public assistance. *Journal of Public Economics*, 88 (7):1421–1443, 2004.

John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450, 1995.

Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015.

T Burdett, PN Hall, E Hasting, LA Hindorff, HA Junkins, AK Klemm, J MacArthur, TA Manolio, J Morales, H Parkinson, et al. The nhgri-ebi catalog of published genome-wide association studies. *Available at: www. ebiacuk/gwas*, 2016.

Stephen Burgess and Simon G Thompson. Use of allele scores as instrumental variables for mendelian randomization. *International journal of epidemiology*, 42(4):1134–1144, 2013.

Stephen Burgess, Simon G Thompson, and CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in mendelian randomization studies. *International journal of epidemiology*, 40(3):755–764, 2011.

Stephen Burgess, Raquel Granell, Tom M Palmer, Jonathan AC Sterne, and Vanessa Didelez. Lack of identification in semiparametric instrumental variable models with binary outcomes. *American journal of epidemiology*, 180(1):111–119, 2014.

Paul S Clarke and Frank Windmeijer. Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652, 2012.

Zhifu Cui, Hang Zhang, and Wei Lu. An improved smoothed l0-norm algorithm based on multiparameter approximation function. In *Communication Technology (ICCT), 2010 12th IEEE International Conference on*, pages 942–945. IEEE, 2010.

George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics*, 23(R1):R89–R98, 2014.

Neil M Davies, Stephanie Hinke Kessler Scholder, Helmut Farbmacher, Stephen Burgess, Frank Windmeijer, and George Davey Smith. The many weak instruments problem and mendelian randomization. *Statistics in medicine*, 34(3):454–468, 2015.

Thomas S Dee and William N Evans. Teen drinking and educational attainment: evidence from two-sample instrumental variables estimates. *Journal of Labor Economics*, 21(1): 178–209, 2003.

Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16(4):309–330, 2007.

Eric C Donny, Kasey M Griffin, Saul Shiffman, and Michael A Sayette. The relationship between cigarette use, nicotine dependence, and craving in laboratory volunteers. *Nicotine & Tobacco Research*, 10(3):447–455, 2008.

Matthias Egger, George Davey Smith, Martin Schneider, and Christoph Minder. Bias in meta-analysis detected by a simple, graphical test. *Bmj*, 315(7109):629–634, 1997.

Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

Gene H Golub. Some modified matrix eigenvalue problems. *Siam Review*, 15(2):318–334, 1973.

Gene H Golub and Charles F Van Loan. Matrix computations. 1996. *Johns Hopkins University, Press, Baltimore, MD, USA*, pages 374–426, 1996.

Michael H Graham. Confronting multicollinearity in ecological multiple regression. *Ecology*, 84(11):2809–2815, 2003.

Rajdeep Grewal, Joseph A Cote, and Hans Baumgartner. Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science*, 23(4):519–529, 2004.

Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

Jinyong Hahn and Atsushi Inoue. A monte carlo comparison of various asymptotic approximations to the distribution of instrumental variables estimators. *Econometric Reviews*, 21(3):309–336, 2002.

Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422, 2008.

Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 2012.

Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

Fumio Hayashi. Econometrics. 2000. *Princeton University Press. Section*, 1:60–69, 2000.

Rayjean J Hung, James D McKay, Valerie Gaborieau, Paolo Boffetta, Mia Hashibe, David Zaridze, Anush Mukeria, Neonilia Szeszenia-Dabrowska, Jolanta Lissowska, Peter Rudnai, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, 452(7187):633–637, 2008.

Y Iturria-Medina, RC Sotero, PJ Toussaint, JM Mateos-Pérez, AC Evans, Alzheimers Disease Neuroimaging Initiative, et al. Early role of vascular dysregulation on late-onset alzheimer/'s disease based on multifactorial data-driven analysis. *Nature Communications*, 7, 2016.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

Roz Azinur Che Lamin, Nursyuhadah Othman, and Che Noriah Othman. Effect of smoking behavior on nicotine dependence level among adolescents. *Procedia-Social and Behavioral Sciences*, 153:189–198, 2014.

Debbie A Lawlor. Commentary: Two-sample mendelian randomization: opportunities and challenges. *International journal of epidemiology*, 45(3):908, 2016.

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008.

DY Lin and D Zeng. Proper analysis of secondary phenotype data in case-control association studies. *Genetic epidemiology*, 33(3):256–265, 2009.

América Martínez-Calleja, Irma Quiróz-Vargas, Isela Parra-Rojas, José Francisco Muñoz-Valle, Marco A Leyva-Vázquez, Gloria Fernández-Tilapa, Amalia Vences-Velázquez, Miguel Cruz, Eduardo Salazar-Martínez, and Eugenia Flores-Alfaro. Haplotypes in the crp gene associated with increased bmi and levels of crp in subjects with type 2 diabetes or obesity from southwestern mexico. *Experimental diabetes research*, 2012, 2012.

Masahiro Maruyama, Hiroyuki Arai, Mitsunori Sugita, Haruko Tanji, Makoto Higuchi, Nobuyuki Okamura, Toshifumi Matsui, Susumu Higuchi, Sachio Matsushita, Hiroshi Yoshida, et al. Cerebrospinal fluid amyloid β 1–42 levels in the mild cognitive impairment stage of alzheimer's disease. *Experimental neurology*, 172(2):433–436, 2001.

Laurence Menard, David Saadoun, Isabelle Isnardi, Yen-Shing Ng, Greta Meyers, Christopher Massad, Christina Price, Clara Abraham, Roja Motaghedi, Jane H Buckner, et al. The ptpn22 allele encoding an r620w variant interferes with the removal of developing autoreactive b cells in humans. *The Journal of clinical investigation*, 121(9):3635–3644, 2011.

Genevieve M Monsees, Rulla M Tamimi, and Peter Kraft. Genome-wide association scans for secondary traits using case-control samples. *Genetic epidemiology*, 33(8):717–728, 2009.

Marcelo J Moreira, Jack R Porter, and Gustavo A Suarez. Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics*, 149(1):52–64, 2009.

Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimers disease: the alzheimers disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Whitney K Newey and Frank Windmeijer. Gmm with many weak moment conditions. 2005.

Tom M Palmer, John R Thompson, Martin D Tobin, Nuala A Sheehan, and Paul R Burton. Adjusting for bias and unmeasured confounding in mendelian randomization studies with binary responses. *International journal of epidemiology*, 37(5):1161–1168, 2008.

Tom M Palmer, Debbie A Lawlor, Roger M Harbord, Nuala A Sheehan, Jon H Tobias, Nicholas J Timpson, George Davey Smith, and Jonathan AC Sterne. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical methods in medical research*, 21(3):223–242, 2012.

Brandon L Pierce, Habibul Ahsan, and Tyler J VanderWeele. Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. *International journal of epidemiology*, page dyq151, 2010.

Mary Rieck, Adrian Arechiga, Suna Onengut-Gumuscu, Carla Greenbaum, Patrick Concannon, and Jane H Buckner. Genetic variation in ptpn22 corresponds to altered function of t and b lymphocytes. *The Journal of Immunology*, 179(7):4704–4710, 2007.

Lior Shamai, Einar Lurix, Michael Shen, Gian M Novaro, Samuel Szomstein, Raul Rosenthal, Adrian V Hernandez, and Craig R Asher. Association of body mass index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese. *Obesity surgery*, 21(1):42–47, 2011.

George Davey Smith and Shah Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International journal of epidemiology*, 33(1):30–42, 2004.

Nadia Solovieff, Chris Cotsapas, Phil H Lee, Shaun M Purcell, and Jordan W Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14(7): 483–495, 2013.

Frank W Stearns. One hundred years of pleiotropy: a retrospective. *Genetics*, 186(3): 767–773, 2010.

James H Stock, Jonathan H Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 2012.

Trey Sunderland, Gary Linker, Nadeem Mirza, Karen T Putnam, David L Friedman, Lida H Kimmel, Judy Bergeson, Guy J Manetti, Matthew Zimmermann, Brian Tang, et al. Decreased β-amyloid1-42 and increased tau levels in cerebrospinal fluid of patients with alzheimer disease. *Jama*, 289(16):2094–2103, 2003.

Jean de Dieu Tapsoba, Charles Kooperberg, Alexander Reiner, Ching-Yun Wang, and James Y Dai. Robust estimation for secondary trait association in case-control genetic studies. *American journal of epidemiology*, 179(10):1264–1272, 2014.

Eric J Tchetgen Tchetgen. A general regression framework for a secondary outcome in case–control studies. *Biostatistics*, 15(1):117–128, 2013.

Thorgeir E Thorgeirsson, Frank Geller, Patrick Sulem, Thorunn Rafnar, Anna Wiste, Kristinn P Magnusson, Andrei Manolescu, Gudmar Thorleifsson, Hreinn Stefansson, Andres Ingason, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, 452(7187):638–642, 2008.

Nicholas J Timpson, Debbie A Lawlor, Roger M Harbord, Tom R Gaunt, Ian NM Day, Lyle J Palmer, Andrew T Hattersley, Shah Ebrahim, Gordon DO Lowe, Ann Rumley, et al. C-reactive protein and its role in metabolic syndrome: mendelian randomisation study. *The Lancet*, 366(9501):1954–1959, 2005.

John A Todd, Neil M Walker, Jason D Cooper, Deborah J Smyth, Kate Downes, Vincent Plagnol, Rebecca Bailey, Sergey Nejentsev, Sarah F Field, Felicity Payne, et al. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics*, 39(7):857–864, 2007.

Stijn Vansteelandt, Jack Bowden, Manoochehr Babanezhad, Els Goetghebeur, et al. On instrumental variables estimation of causal odds ratios. *Statistical Science*, 26(3):403–422, 2011.

Günter P Wagner and Jianzhi Zhang. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, 12(3):204–213, 2011.

George L Wehby, Robert L Ohsfeldt, and Jeffrey C Murray. mendelian randomizationequals instrumental variable analysis with genetic instruments. *Statistics in medicine*, 27(15): 2745–2749, 2008.