METHODS FOR CORRECTING THE ACCURACY IN MENDELIAN RANDOMIZATION

METHODS FOR CORRECTING THE ACCURACY IN MENDELIAN RANDOMIZATION

BY

MENGJIE BIAN, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS AND STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Doctor of Philosophy (2023) McMaster University
(Mathematics and Statistics) Hamilton, Ontario, Canada

TITLE:	Methods for Correcting the Accuracy in Mendelian Randomization
AUTHOR:	Mengjie Bian M.Sc. (Statistics), McMaster University, Hamilton, Canada M.Sc. (Finance), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Angelo J. Canty
NUMBER OF PAGES:	xxiv, 213

Abstract

Mendelian randomization (MR) uses genetic variants as instrumental variables (IVs) to investigate the causal relationship between exposure and outcome. It has become widely popular due to its versatile applications in epidemiological research. Its rising popularity is largely driven by the ease of accessing summary-level data from large consortia, making it a cost-effective choice for researchers.

In this thesis, we focus on three issues in MR that result in potential bias in causal inference. We first address the "winner's curse" in MR, which arises from selecting genetic markers based on their significance or ranking. To mitigate this bias, we adapt the bootstrap-based BR-squared method to function with summary-level data. Our findings reveal that the correction methods can effectively reduce bias, albeit with an increase in variability. We then develop a method that accounts for the correlation caused by sample overlap while addressing potential bias from weak instruments. This proposed method yields stable causal estimates, although the standard errors of causal estimates may not be precisely estimated. Lastly, we introduce a novel approach for identifying invalid instrumental variables showing signs of horizontal pleiotropy. We recommend using the bootstrap method to account for the data-driven process of IV selection. Our results indicate that the bootstrap intervals approach the nominal level of coverage rate when the proportion of invalid IVs is less than 50%.

Acknowledgements

I would like to take this moment to express my profound gratitude to all those who have contributed to the completion of my Ph.D. thesis. These past three years have presented unique challenges, particularly due to the unforeseen impact of the COVID-19 pandemic, which coincided with my doctoral studies. My supervisor and I had to adapt to online communication during the first two years of the pandemic, which was a significant transition and inevitably affected the progress of my research. However, it is with unwavering support that we have overcome these challenges.

First and foremost, my deepest thanks go to my supervisor, Dr. Angelo Canty. Not only has he provided academic guidance, but he has also been a pillar of support for my mental well-being. His encouragement and confidence in my abilities, coupled with his invaluable financial support, have been pivotal to my journey. He consistently went the extra mile to ensure I had the resources and guidance I needed, even in the face of unexpected challenges. His mentorship has been a guiding light throughout this arduous journey. He not only believed in my academic potential but also in my personal resilience, instilling in me the confidence to overcome the hurdles that came my way.

I am also immensely grateful to my committee members, Dr. Guillaume Paré and Dr. Roman Viveros-aguilera. Dr.Paré's extensive knowledge in my research field enriched my work, providing valuable insights and clear, direct guidance. Dr. Viveros, a compassionate and understanding individual, played a crucial role in encouraging and supporting me during these unique times.

I extend my warmest appreciation to Dr. Katherine Davies, my new committee member, for joining my academic journey during this critical phase. Your guidance and assistance in preparing for the defense are highly valued.

My heartfelt thanks go to my parents and Xuezheng, Huang, who have been the unwavering pillars of strength throughout my Ph.D. pursuit. Their unending support and belief in me have been my motivation to persevere.

The completion of this Ph.D. thesis would not have been possible without the collective contributions of these remarkable individuals. I am profoundly grateful for the support and encouragement I have received from everyone mentioned.

Contents

List of Figures

List of Tables

Notation and Abbreviations

Notation

X	exposure or risk factor
Y	outcome
G	genetic variants
U	confounding factor
ε	random error in the regression model
β_c	causal effect
$\hat{\beta}_{gx}$	estimate of genetic-variant-exposure association (G-X)
$\hat{\beta}_{gy}$	estimate of genetic-variant-outcome association (G-Y)
α_u	effect of confounder on exposure
β_u	effect of confounder on outcome
α	effect of genetic variant on exposure

ϕ	direct effect of genetic variant on outcome
F	F-statistic from regression of X on G
n	the total number of observation
M	the total number of instruments
R	the number of bootstrap samples
z	z-statistic: the ratio of estimated G-X association and its standard error
μ	the mean of z-statistics
n_c	the number of overlapping samples
n_x	the sample size for the G-X association
n_y	the sample size for the G-Y association
r_{xy}	the sample correlation between X and Y

Abbreviations

AD	Alzheimer's disease
CHD	coronary heart disease
CI	confidence interval
Compromise	Compromise estimator

CRP	C-reactive protein
DAG	directed acyclic graph
DNA	Deoxyribonucleic Acid
FDR	false discovery rate
FIQT	FDR Inverse Quantile Transformation
FPR	false positive rate
GIANT	Genetic Investigation of ANthropometric Traits
GLGC	Global Lipid Genetics Consortium
GWAS	genome-wide association study
HDL-C	high-density lipoprotein cholesterol
HWE	Hardy–Weinberg equilibrium
IGAP	International Genomics of Alzheimer's Project
IL-10	Interleukin-10
InSIDE	Instrument Strength Independent of Direct Effect
IV	instrumental variable
IVW	inverse-variance weighted
LD	linkage disequilibrium
LDL-C	low-density lipoprotein cholesterol

LIML	limited information maximum likelihood
MAF	minor allele frequency
MBE	mode-based estimator
MCMC	Markov Chain Monte Carlo
MLE	maximum likelihood estimator
MR	Mendelian randomization
MR-Egger	Mendelian randomization-Egger
MR-PRESSO	MR Pleiotropy Residual Sum and Outlier
MR-RAPS	Robust Adjusted Profile Score
MSE	mean squared error
NOME	no measurement error
Normalized	normalized conditional likelihood
OLS	ordinary least squares
Projack	Predicion by Re-Ordered Jackknife and Cross-Validation K-fold
RCT	Randomized Controlled Trials
SD	standard deviation
SE	standard error
SID	Šidák correction

SNP	single nucleotide polymorphisms
TGL	triglycerides
T2D	type 2 diabetes
TPR	true positive rate
TSLS	two-stage least squares
VIS	Variation in Instrument Strength
WHR	waist-to-hip ratio
ZEMPA	Zero Modal Pleiotropy Assumption

Chapter 1

Introduction

1.1 What is a Causal Effect

A causal effect represents the specific impact of a risk factor or treatment on an outcome, indicating how changes in the risk factor or treatment directly lead to changes in the outcome variable. Following Pearl (1995), we can describe the causality of exposure (X) on the distribution of outcome (Y) as: $P(Y = y|do(X = x))$, where the "do" notation $do(X = x)$ indicates an intervention on the variable X. This expression represents the probability (or frequency) that event ($Y = y$) would occur if treatment condition $X = x$ were enforced uniformly over the population.

In observational studies, the researchers do not manipulate the causal variable. Instead, they observe the natural variation in the causal variable and its effect on the outcome variable. As a result, the observed conditional probability of the effect variable given the causal variable, denoted as $P(Y = y|X = x)$, represents the probability of Y taking the value y given the observed value of X as x in the data.

The relationship between $P(Y = y|do(X = x))$ and $P(Y = y|X = x)$ depends on

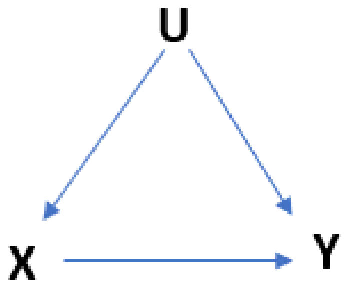


Figure 1.1: Directed acyclic graph (DAG) for confounded association between X and Y

the presence of confounding variables (U). Confounders are additional variables that influence both the causal variable X and the outcome variable Y. They can distort the observed relationship between X and Y and lead to a biased estimator of the causal effect.

When there are no confounders, the causal effect $P(Y = y|do(X = x))$ and the observed conditional probability $P(Y = y|X = x)$ are equal. In such cases, the observed association between X and Y can be interpreted causally (Pearl and Mackenzie, 2018).

However, in the presence of confounders, $P(Y = y|do(X = x))$ and $P(Y = y|X = x)$ may differ. In this scenario, the observed conditional probability $P(Y = y|X = x)$ does not represent the true causal effect of X on Y due to the confounding influence.

In reality, the presence of unknown or unmeasured confounders remains uncertain. Causal inference serves as a means to untangle true causality from observed correlations, especially when confronted with unfamiliar confounders, surpassing the observational studies.

1.2 Observational Studies and Randomized Controlled Trials

Observational studies involve the examination and analysis of data from existing populations or groups without manipulating exposures or interventions. They rely on existing data, which can include data from medical records, surveys, registers or other sources. While observational studies are valuable for generating hypotheses and exploring associations, they have limitations such as confounding, selection bias, and potential challenges in establishing causal inferences. Confounding arises when unmeasured variables are associated with both the exposure and the outcome, leading to biased estimates of their relationship. For example, a study investigating the link between smoking and human overall health may be confounded by factors like diet, alcohol consumption and lifestyles, which are related to both smoking and health (Tjønneland *et al.*, 1999).

Reverse causality is another issue in observational studies, occurring when the outcome influences the exposure instead of the other way around. For instance, in a study examining the association between child adiposity (the amount of body fat in a child) and physical activity, individuals with lower physical activity also showed increased adiposity, resulting in an incorrect interpretation of causality (Richmond *et al.*, 2014).

In contrast, randomized trials, also known as randomized controlled trials (RCTs), can help address potential confounding and reverse causation by randomly assigning participants to different groups or interventions. Participants are randomly allocated to receive either the experimental treatment or a control group (placebo or standard treatment). The comparison of outcomes between these groups establishes the causal relationship. Randomization minimizes bias and confounding, making RCTs the gold standard for assessing causal effects. However, RCTs have limitations, such as their high cost, time-consuming nature, and logistical challenges. Some research questions or interventions may not be feasible or ethical to investigate through RCTs. For example, studying lifestyle factors such as smoking, alcohol consumption, or diet through RCTs could be ethically questionable or practically unfeasible due to potential harm to participants. As for the relationship between greater adiposity or body mass index (BMI) and cardiovascular risk, observational evidence has long suggested a strong link (Hubert *et al.*, 1983). However, randomized trials exploring weight reduction as an intervention have been relatively scarce and challenging to conduct, leaving the question of causality unanswered.

1.3 The Rise of Mendelian Randomization

Mendelian randomization (MR) (Katan, 1986) is an innovative approach in epidemiology and genetics that employs genetic variants as instrumental variables to investigate causal relationships between exposures and outcomes. This method has gained significant popularity in the field of causal inference due to its unique strengths and advantages over traditional observational studies and randomized controlled trials. Its historical roots can be traced back to the concept of instrumental variables (IV) in the field of econometrics. Instrumental variables were first used by Philip G. Wright (Wright, 1928) when he sought to analyze the supply and demand for butter in the United States. Facing challenges in constructing accurate demand and supply curves due to the influence of price on both variables, Wright introduced an instrumental variable – regional rainfall – which correlated with supply but not demand. By using the instrumental variable in his regression analysis, he estimated the causal effect of price on supply. The term "instrumental variable" was later formalized by Olav Reiersøl in 1945 when he applied the same approach for the errors in variables problem in his dissertation, giving the method its name (Reiersøl, 1945).

In MR, genetic variants act as IVs, enabling the investigation of exposure-outcome causality while minimizing bias from confounding factors and reverse causation. The concept of Mendelian randomization was first proposed by Katan (1986) in a study examining the link between low serum cholesterol levels and cancer. Traditional observational studies were susceptible to biased estimates due to unaccounted confounders. Katan suggested comparing individuals with different Apolipoprotein E (APOE) genotypes, which are associated with serum cholesterol levels but unlikely to be influenced by confounders. This approach was later termed "Mendelian Randomization" by Gray and Wheatley (1991). Unlike RCTs, where individuals are randomly assigned to treatment groups, MR studies compare individuals with different genotypes (Figure 1.2). Genotypes are fixed at conception and are independent of confounding factors. This makes genetic variants ideal IVs for causal inference. While intrumental variables has been extensively studied in econometrics (Anderson and Rubin, 1949; Basmann, 1957; Hansen, 1982; Bound *et al.*, 1995; Angrist *et al.*, 1996; Stock *et al.*, 2002), they have also gained popularity in epidemiology as a powerful tool for assessing causal effects. This instrumental variable framework provides a natural experiment-like setting to estimate the causal effect of the exposure on the outcome. These genetic variants, often single nucleotide polymorphisms (SNPs), are randomly allocated during meiosis and are thus independent of environmental and

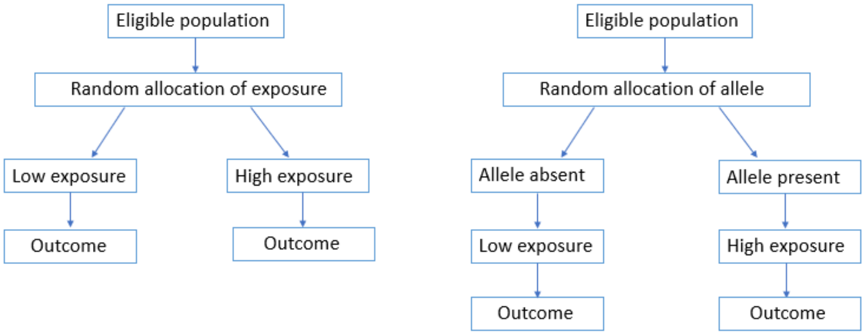


Figure 1.2: Randomized Controlled Trial vs Mendelian Randomization trial

lifestyle factors. By leveraging genetic variants as instrumental variables, MR aims to mimic a randomized controlled trial in which individuals are randomly assigned to different levels of exposure. MR is particularly valuable for studying long-term effects of exposures that are challenging or ethically impractical to manipulate in RCTs. The growing availability of large-scale genomic data and advancements in genotyping technologies contribute to the popularity of MR. Genome Wide Association Studies (GWAS) have facilitated the identification of genetic variants associated with various exposures and outcomes, making MR analyses more feasible and informative.

1.4 A Brief Review of Genetics

Genes are the fundamental units of heredity that carry the instructions for the production of proteins. These genes are encoded within DNA (Deoxyribonucleic Acid), the genetic material that stores and transmits genetic information. DNA is organized into structures called chromosomes, which exist in pairs in humans and contain numerous genes along their length. Each gene is composed of a specific sequence of nucleotides, which are the building blocks of DNA. The nucleotides consist of four bases: adenine (A), thymine (T), cytosine (C), and guanine (G). The sequence of these bases within a gene determines the sequence of amino acids in a protein.

Variations can arise within the DNA sequence at specific positions known as SNPs. SNPs represent the most prevalent form of genetic variation, where a single nucleotide base (A, T, C, or G) may differ among individuals at a particular position. Genotype refers to the combination of alleles at a particular SNP within an individual's genome. Alleles are alternative forms or variants of a gene at the same position on paired chromosomes. For instance, at a given SNP, individuals can have different genotypes. Homozygous genotypes (e.g., AA or TT) refer to individuals carrying two identical alleles at a specific SNP. In these cases, both copies of the allele at the given SNP are the same nucleotide base. Heterozygous genotypes where they carry one copy of each allele (e.g., AT). Linkage disequilibrium (LD) refers to the phenomenon whereby the genetic makers that are close to each other tend to be inherited together.

The frequency of different alleles within a population is measured by the Minor Allele Frequency (MAF). MAF represents the proportion of individuals in a population who carry the less common allele at a specific SNP. For instance, if a SNP has an A allele and a T allele, and the T allele is less common, the MAF would reflect the frequency of the T allele among individuals in that population. Hardy–Weinberg equilibrium (HWE) states that, in the absence of external evolutionary factors, the frequencies of alleles and genotypes within a population will remain stable across successive generations. Under the assumption of HWE, the distribution of genotypes in a population can be modeled using the binomial distribution. Specifically, if we consider a bi-allelic locus (two alleles, often denoted as A (minor allele) and a), the genotypes AA, Aa, and aa can be treated as the outcomes of a binomial experiment.

For example, if p and q denote the frequencies of allele A and allele a, respectively, then the probabilities of the three genotypes AA, Aa, and aa under HWE are p^2, $2pq$ and q^2, respectively. These probabilities follow the binomial distribution $Bin(n, p)$ with $n = 2$ and $p=$ MAF.

A Genome Wide Association Study (GWAS) (Ozaki *et al.*, 2002) is a powerful approach used to identify genetic variants associated with complex traits or diseases. It involves scanning the entire genome for SNPs and assessing their associations with specific traits or phenotypes. For example, a GWAS might investigate the association between SNPs and the risk of developing a particular disease, such as diabetes or cancer. The primary goal of GWAS is to identify SNPs that are significantly associated with a particular trait or disease of interest. By comparing the SNP patterns among individuals with and without the trait or disease, researchers can determine which genetic variations are more prevalent in one group compared to the other.

GWAS is typically conducted in large cohorts that include thousands to hundreds of thousands of participants. The participants undergo genotyping, where their DNA is analyzed to determine the presence of specific genetic markers, usually SNPs. Detailed phenotypic information, such as medical history, clinical measurements, or lifestyle factors, is also collected to correlate with the genotypic data. The GWAS threshold (a significance level), typically at 5×10^{-8}, is rooted in the Bonferroni correction—a technique that adjusts for multiple testing issues in GWAS. In GWAS, this correction divides the significance level (like 0.05) by the total number of independent tests conducted across the genome.

1.5 Mendelian Randomization Application

A systematic review of applied Mendelian randomization studies was published in several works (Sheehan *et al.*, 2008; Pierce *et al.*, 2018; Markozannes *et al.*, 2022). Table 1.1 gives a range of examples of MR applications in the field of epidemiology. As observational studies cannot establish the causality due to the potential confounding or biases, MR provides an alternative approach to assess the causality. An illustration of this can be seen in the establishment of a causal relationship between elevated low-density lipoprotein cholesterol (LDL-C) and the occurrence of coronary heart disease (CHD) events. This association has been derived from observational studies and confirmed through randomized trials investigating the effectiveness of LDL-C-lowering medications (Cholesterol Treatment Trialists' (CTT) Collaborators, 2005, 2012). However, there is a lack of consensus between observational studies and randomized trials regarding the causal association of high-density lipoprotein cholesterol (HDL-C) and triglycerides with CHD. While observational studies clearly indicate a positive association between triglycerides and CHD and an inverse association for HDL-C (The Emerging Risk Factors Collaboration, 2009), the anticipated benefits have not been demonstrated thus far in randomized trials involving HDL-C or triglyceride modifying drugs (Schwartz *et al.*, 2012; The FIELD Study Investigators, 2005). Subsequent MR studies provide additional support for a causal relationship between triglycerides and the risk of coronary heart disease (CHD). However, the causal role of HDL-C, although plausible, remains less certain (Holmes *et al.*, 2015; Do *et al.*, 2013).

Table 1.1: A list of Mendelian Randomization application. BMI: body mass index; WHR: waist hip to ratio; TGL:triglycerides; LDL-C: low-density lipoprotein cholesterol; T2D: type 2 diabetes; CRP: C-reactive protein; IL-10: Interleukin-10

Exposure	Outcome	Reference
Dietary intake and micronutrient concentration		
Alcohol	Esophageal	Lewis and Davey Smith (2005)
Alcohol	Head and neck	Boccia *et al.* (2009)
Alcohol	Colorectal cancer	Wang *et al.* (2011)
Vitamin D & B12	Prostate cancer	Bonilla *et al.* (2013)
Vitamin D & B12	Osvarian cancer	Ong *et al.* (2016)
Coffee	Prostate cancer	Taylor *et al.* (2017)
Magnesium	Breast cancer	Papadimitriou *et al.* (2021)
Ferrition	Liver cancer	Yuan *et al.* (2020a)
Anthropometric traits		
BMI	Colerectal cancer	Thrift *et al.* (2015); Gao *et al.* (2016)
BMI	Ovarian cancer	Painter *et al.* (2016)
WHR	Colerectal cancer	Jarvis *et al.* (2016)
Lipid traits		
TGL	Breast cancer	Orho-Melander *et al.* (2018)
LDL-C	Endometrial cancer	Kho *et al.* (2021)
Disease		
T2D	Esophageal	Yuan *et al.* (2020b)
Schizophrenia	Breast cancer	Shi *et al.* (2018)
Inflammation		
CRP	Colerectal cancer	Nimptsch *et al.* (2017)
IL-10	Digestive cancer	Niu *et al.* (2016)

1.6 Assumptions and Selection of Instrument Variables

To make a genetic variant a valid instrumental variable, three assumptions must be met (Martens *et al.*, 2006):

- IV1: it is strongly associated with the exposure of interest.
- IV2: it is independent of any confounding factors of the exposure-outcome association.
- IV3: it is conditionally independent of the outcome given the exposure and the confounding factors.

Figure 1.3 shows the standard Directed Acyclic Graph (DAG) for the assumptions of MR, in which a genetic instrument is associated with an exposure X, and its influence on the outcome variable Y arises exclusively through its impact on X.

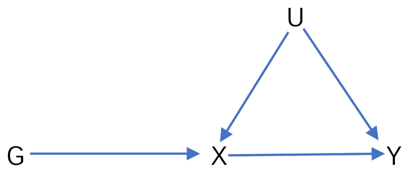


Figure 1.3: Graph of instrument variable assumptions

1.6.1 IV 1: Relevance

The first assumption in MR states that the genetic variant used as an instrumental variable should be strongly associated with the exposure variable of interest. This assumption is crucial to avoid weak instrument bias. Weak instrument bias happens if the association between the genetic variants and exposure is weak. If the association between the genetic variant and the exposure variable is weak, it means that the genetic variant explains only a small proportion of the variation in the exposure. In such cases, the variation in confounding factor may account for a greater portion of the variation in the exposure than the instruments do. Consequently, using such a weak IV in MR analysis can lead to a biased estimator of the causal effect between the exposure and the outcome. The bias is in the direction of the observational confounded association (Burgess *et al.*, 2011). The first-stage F-statistic is often used to judge whether the instrument is weak or not. The first-stage F-statistic is obtained from the regression of exposure X on the genetic variants, which can be calculated as

$$F = \frac{\sum_{i=1}^{n}(\hat{X}_i - \overline{X})^2/(M-1)}{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2/(n-M)} \tag{1.6.1}$$

where n is the number of observations and M is the number of IVs. Typically, the first-stage F must be larger than 10 for the causal estimate to be reliable (Staiger and Stock, 1997; Stock *et al.*, 2002). It was shown that the relative mean bias, which is defined as the ratio of the bias of the IV estimator to the bias of the ordinary least square (OLS) estimator, is approximately inversely proportional to $1/F$ (Staiger and Stock, 1997; Burgess and Thompson, 2011). If the genetic variant-exposure (G-X) and genetic variant-outcome (G-Y) association estimates are obtained from two non-overlapping samples, then these estimates are uncorrelated. In this case, Pierce and Burgess (2013) showed that the bias due to the weak IVs is in the direction of null when the sample size for the exposure is small compared to that for the outcome. However, when the sample size for the exposure is large compared to that for the outcome, the bias is in the direction of confounded (observational) association. The direction of confounded (observational) association depends on the product of the signs of the confounder effect on both the exposure and the outcome.

1.6.2 IV 2: Independence

Assumption 2 relies on the concept that the genetic variants that are used as instruments are independent of other variables that could confound the relationship between the exposure and outcome. This independence assumption relies on Mendel's law of independent assortment, which underlies the idea that different traits are inherited independently. Mendel's law of independent assortment ensures that the genetic instrument used is not correlated with the confounders. Nevertheless, the independence assumption can be violated by environmental and social factors such as assortative mating, dynastic effects, and population structure, which induce an association between genetic variants and outcome (Brumpton *et al.*, 2020).

Population structure occurs when there are geographic or regional differences in allele frequency relating to a trait of interest. Techniques such as principal component analysis and linear mixed models have been demonstrated to effectively mitigate the confounding effects introduced by population stratification (Price *et al.*, 2010). Dynastic effect occurs when the parent's genotype affects the outcome of the offspring that are mediated through the parent's phenotype. Assortative mating refers to the tendency of individuals to choose mates who have similar characteristics to themselves. It is a non-random pattern of mate selection based on certain traits or characteristics (Hartwig *et al.*, 2018). These challenges can be addressed by using within-family designs which focus on family units, such as parent-offspring trios or siblings. Brumpton *et al.* (2020) showed that the within-family designs can be applied in the context of Mendelian randomized with genomic data to obtain a more robust estimate of causal effect.

1.6.3 IV 3: Exclusion restriction

The assumption of conditional independence between genetic variants and the outcome is violated when genetic variants impact the outcome through pathways other than the exposure, known as horizontal pleiotropy (Figure 1.4) (Bowden et al., 2015). Vertical pleiotropy, on the other hand, occurs when genetic variants are associated with other variables on the same pathway as the exposure of interest (Figure 1.5). It is worth noting that vertical pleiotropy does not introduce bias in causal estimation, whereas horizontal pleiotropy can induce bias, especially when the average pleiotropic effect across genetic variants is non-zero, referred to as directional pleiotropy. However, if the average pleiotropy effect is zero, also known as balanced pleiotropy, the causal estimate will not be biased (Bowden *et al.*, 2015). Numerous approaches have been developed to deal with horizontal pleiotropy (Koller and Stahel, 2011; Bowden *et al.*, 2015, 2016b; Hartwig *et al.*, 2017; Rees *et al.*, 2019; Qi and Chatterjee, 2019; Zhao *et al.*, 2020; Burgess *et al.*, 2020). We will introduce some of these methods in details in the following sections.

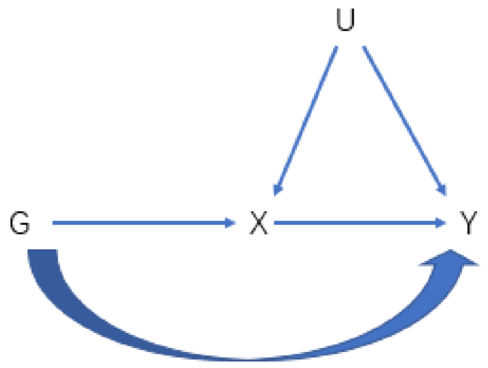


Figure 1.4: Horizontal pleiotropy

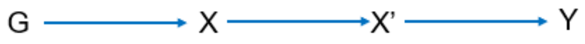


Figure 1.5: Vertical pleiotropy

1.6.4 Selection of Instrument Variables

The simplest application of Mendelian randomization involves using a single genetic variant as an instrument for an exposure. However, a single variant usually explains only a small portion of phenotype variation. Studies using such variants can be statistically underpowered and biased. To address this, we can use multiple independent SNPs as instruments to explain more variation in the exposure. Alternatively, the SNPs can be combined into an allele score, acting as a single IV for predicting the exposure in MR analyses (Burgess and Thompson, 2013).

Typically, the unrelated genetic variants across multiple regions of genome which meet the GWAS significance threshold, are chosen as IVs for investigating the causal role of an exposure on a outcome. Recently, MR has also gained popularity in drug target validation (Gill *et al.*, 2021), where variants within a single gene region are selected as IVs, a practice also known as *cis*-MR. In this application, the protein expression is taken as the exposure. This is attributed to the fact that each coding gene section encodes a distinct protein and proteins often serve as targets for medications. By using the variants selected from that region for protein encoding, *cis*-MR studies provide insights into the potential of the encoded protein as a drug target for the outcome (Gill *et al.*, 2021). For example, the application of MR using IVs selected from the SNPs in the interleukin-6 receptor (IL6R) gene has indicated that targeting of IL6R could provide a novel therapeutic approach to prevention of coronary heart disease (The Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium, 2012).

1.7 One-sample Mendelian Randomization

In one-sample Mendelian randomization, the data on genetic variants, exposure and outcome are from the same sample of individuals while in two-sample MR, the G-X and G-Y association estimates are measured in two different samples, and typically, these two samples need to be independent to avoid bias. The advantage of one-sample MR over two-sample MR is that one-sample MR allows us to conduct analysis in specific subgroups or to choose which variables to adjust for when generating the summarized data (Burgess *et al.*, 2023). In one-sample MR, a common approach to estimate the causal effect is using the two-stage least squares method (TSLS). To illustrate this method, let's assume the following linear model. The exposure X_i for individual i is a linear combination of M independent genetic variants, a confounder factor U_i and random error ε_{x_i}; the outcome Y_i is a linear combination of exposure X_i, confounding factor U_i and random error ε_{y_i}.

$$\begin{aligned}
&X_i = \sum_{k=1}^{M} \alpha_k G_{ik} + \alpha_u U_i + \varepsilon_{x_i} \\
&Y_i = \beta_c X_i + \beta_u U_i + \varepsilon_{y_i} \\
&U_i \sim N\left(0, \sigma_u^2\right); \varepsilon_{x_i} \sim N\left(0, \sigma_x^2\right); \varepsilon_{y_i} \sim N\left(0, \sigma_y^2\right) \qquad (1.7.1)
\end{aligned}$$

We let $v_{y_i} = \beta_u U_i + \varepsilon_{y_i}$ and $v_{x_i} = \alpha_u U_i + \varepsilon_{x_i}$. Then v_{y_i} can be written in form of v_{x_i}: $v_{y_i} = \frac{\beta_u}{\alpha_u}(v_{x_i} - \varepsilon_{x_i}) + \varepsilon_{y_i}$. From this expression, we can see that X_i is correlated with v_{y_i} since X_i is a linear combination of v_{x_i} and genetic variants and v_{x_i} is correlated with v_{y_i}. With individual-level data, the ordinary least square estimator of β_c is a biased estimator and is not consistent as the exposure X_i is correlated with v_{y_i}. Two assumptions are needed for the consistency of OLS (Wooldridge, 2010):

Assumption 1 $E(\boldsymbol{X'v_y}) = \mathbf{0}$

Because $\boldsymbol{X}$ contains a constant, Assumption 1 is equivalent to saying that $\boldsymbol{v_y}$ has mean zero and is uncorrelated with the regressor.

Assumption 2 $\text{rank}(E(\boldsymbol{X'X})) = K$, with $\boldsymbol{X} = [1, \boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_{K-1}}]$

Note that K is equal to 2 with a single exposure and greater than 2 with multiple exposures. For simplicity, we take $K = 2$ for the illustration of TSLS.

Using genetic variants $\boldsymbol{G} = [\boldsymbol{G_1}, \boldsymbol{G_2}, \cdots, \boldsymbol{G_M}]$ as IVs, TSLS regression produces a consistent and unbiased estimator. The IVs need to satisfy the three assumptions outlined in Section 1.6. Since the IVs are assumed to have no correlation with the confounder, which means that $\text{cov}(\boldsymbol{G}, \boldsymbol{v_y}) = 0$ and $\text{cov}(\boldsymbol{G}, \boldsymbol{v_x}) = 0$, it follows that $\boldsymbol{X^*}$ is uncorrelated with $\boldsymbol{v_x}$, where

$$\boldsymbol{X^*} = \alpha_1 \boldsymbol{G_1} + \alpha_2 \boldsymbol{G_2} + \cdots + \alpha_M \boldsymbol{G_M}. \tag{1.7.2}$$

We can get an estimate of X_{1i} for individual i from the fitted values of an OLS regression of $\boldsymbol{X_1}$ on $\boldsymbol{G}$:

$$\hat{X}_{1i} = \hat{\alpha}_1 G_{i1} + \hat{\alpha}_2 G_{i2} + \cdots + \hat{\alpha}_M G_{iM} \tag{1.7.3}$$

Now, let's define the vector $\boldsymbol{\hat{X}_i} = [1, \hat{X}_{1i}]$ for each individual i to construct the matrix $\boldsymbol{\hat{X}}$ with a dimension of $n \times 2$. Then, the TSLS IV estimator is obtained by running a second OLS regression of $\boldsymbol{Y}$ on fitted values $\boldsymbol{\hat{X}}$:

$$\boldsymbol{\hat{\beta}_c = (\hat{X}'\hat{X})^{-1}\hat{X}'Y} \tag{1.7.4}$$

To summarize, $\boldsymbol{\hat{\beta}_c}$ can be derived through the following two procedures:

1. Perform a first-stage regression, where $\boldsymbol{X_1}$ is regressed on 1, $\boldsymbol{G_1}$, $\boldsymbol{G_2}$, ..., $\boldsymbol{G_M}$.
2. Conduct an OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{\hat{X}}$ in the second-stage regression, resulting in $\boldsymbol{\hat{\beta}_c}$.

Define the projection matrix $\boldsymbol{P = G(G'G)^{-1}G'}$. Then the two-stage least squares estimator can be expressed as:

$$\begin{aligned}\boldsymbol{\hat{\beta}_c} &= \boldsymbol{(\hat{X}'\hat{X})^{-1}\hat{X}'Y} \\ &= \boldsymbol{(X'PX)^{-1}\hat{X}'PY}\end{aligned} \tag{1.7.5}$$

The TSLS residuals are defined as

$$\boldsymbol{\hat{\varepsilon}_i} = Y_i - \boldsymbol{X_i}\boldsymbol{\hat{\beta}_c}, \quad i = 1, 2, ...n \tag{1.7.6}$$

Given the TSLS residuals, a consistent estimator of $\boldsymbol{\sigma^2}$ is

$$\boldsymbol{\hat{\sigma}^2} = (n-2)^{-1}\sum_{i=1}^{n}\boldsymbol{\hat{\varepsilon}_i^2} \tag{1.7.7}$$

So the estimator of asymptotic variance of $\boldsymbol{\hat{\beta}_c}$ can be expressed as

$$\mathrm{var}(\boldsymbol{\hat{\beta}_c}) = \boldsymbol{\hat{\sigma}^2(\hat{X}'\hat{X})^{-1}} \tag{1.7.8}$$

When the instruments have low correlation with the exposure variable X, they are considered "weak", leading to increased bias in the TSLS estimator. If the instruments are both numerous and weak, the TSLS estimator becomes biased towards the probability limit of the corresponding OLS estimator. This bias can be quantified using F-statistic (Equation (1.6.1)), which assesses the joint significance of all instruments in the first-stage regression, i.e.,

$$E[\hat{\beta}_c - \beta_c] \approx \frac{\sigma_{v_x v_y}}{\sigma_{v_x^2}}\frac{1}{F+1} \tag{1.7.9}$$

From this, we see that as the first-stage F statistic gets small, the bias of TSLS approaches $\frac{\sigma_{v_x v_y}}{\sigma^2_{v_x}}$. Meanwhile, the bias of OLS estimate is $\frac{\sigma_{v_x v_y}}{\sigma^2_x}$, which is equivalent to $\frac{\sigma_{v_x v_y}}{\sigma^2_{v_x}}$ when the G-X association is 0. In general, we can assert that TSLS estimates tend to be"biased towards" OLS estimates when there is limited information in the first stage. Conversely, the bias of the TSLS estimator diminishes as the first-stage

F-statistic increases (Stock *et al.*, 2002). It is worthwhile to note that because $\boldsymbol{v_x}$ and $\boldsymbol{v_y}$ are both a linear combination of a confounder and random error, the direction of bias depends on the product of confounder effect on exposure and outcome. A commonly cited criterion to determine the weakness of an instrument is when the F-statistic in the G-X association is below 10, as mentioned by Lawlor *et al.* (2008).

The limited information maximum likelihood (LIML) method, as introduced by Anderson and Rubin (1949), offers advantages over TSLS, particularly when dealing with weak instruments. Notably, the median of the LIML distribution remains close to unbiased even in the presence of weak instruments, as indicated by Pierce and VanderWeele (2012). However, LIML estimates do not have defined moments for any number of instruments (Hahn *et al.*, 2004).

1.8 Two-sample Mendelian Randomization

Two-sample Mendelian randomization has distinct advantages over one-sample MR. Firstly, it allows for a larger sample size by combining data from two independent studies. This increased sample size enhances the statistical power and improves the ability to detect causal relationships. Moreover, two-sample MR only relies on summary statistics rather than individual-level data, simplifying data handling and addressing privacy concerns associated with sharing sensitive information.

The two samples are assumed to come from the same population, and they are typically independent. If the two samples have some overlap, then the causal estimate might be biased. The size of bias is linearly dependent on the degree of overlap. For a case-control setting, if risk factors are only included for the control participants, estimates are still unbiased in a one-sample setting (Burgess *et al.*, 2016). A more thorough discussion of the issue of overlap will be provided in Chapter 3.

Within this section, we will present several widely used approaches for summary-level MR, emphasizing the benefits and underlying assumptions associated with each method.

1.8.1 Inverse Variance Weighted

The ratio method, also known as the Wald ratio method, is a simple approach used in single-instrument MR to estimate the causal effect of an exposure variable on an outcome variable.

The key idea behind the ratio method is to calculate the ratio of the genetic variant association effect on the outcome variable to its effect on the exposure variable. Specifically, we obtain the estimates of the genetic associations of Y ($\hat{\beta}_{gy}$) and X ($\hat{\beta}_{gx}$) and standard errors, obtained from large-scale GWAS.

The formula for the causal estimate $\hat{\beta}_c$ using the ratio method is:

$$\hat{\beta}_c = \frac{\hat{\beta}_{gy}}{\hat{\beta}_{gx}} \tag{1.8.1}$$

The inverse variance weighted (IVW) method is a widely used approach in multiple-instrument MR to estimate the causal effect of an exposure variable on an outcome variable. In this method, we use multiple genetic variants (instruments) that are associated with the exposure variable to derive a weighted average of their individual causal estimates on the outcome.

The key idea behind the IVW method is to combine the effect estimates and their corresponding standard errors for each genetic variant in a meta-analysis-like framework. The weight assigned to each ratio estimate is inversely proportional to its variance (the square of its standard error).

Mathematically, the IVW causal estimate ($\hat{\beta}_{IVW}$) is calculated as follows:

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{M} w_j \hat{\beta}_{c_j}}{\sum_{j=1}^{M} w_j} \tag{1.8.2}$$

where w_j is the inverse variance for the jth ratio estimate:

$$w_j = \widehat{\text{var}(\hat{\beta}_{c_j})}^{-1} \tag{1.8.3}$$

with $\widehat{\text{var}(\hat{\beta}_{c_j})} = \frac{1}{\hat{\beta}_{g_j x}^2} \text{se}^2(\hat{\beta}_{g_j y})$.

Here, $\hat{\beta}_{c_j}$ represents the ratio estimate for IV j, and $\widehat{\text{var}(\hat{\beta}_{c_j})}$ is the corresponding estimated variance. Then the variance of $\hat{\beta}_{IVW}$ is expressed as

$$\text{var}(\hat{\beta}_{IVW}) = \frac{1}{\sum_{j=1}^{M} w_j} \tag{1.8.4}$$

Note that the calculation of $\widehat{\text{var}(\hat{\beta}_{c_j})}$ for each individual genetic instrument (j) assumes the NOME (no measurement error) assumption, where the uncertainty in G-X association is ignored. The IVW estimator can be biased if the three core assumptions of MR are not met. However, there is a unique scenario where the IVW estimator remains unbiased even when the IVs are not valid. This occurs when the pleiotropy effect is balanced across all IVs. In this case, the pleiotropy effect has a zero mean and adheres to the Instrument Strength Independent of Direct Effect (InSIDE) assumption, as described by Bowden *et al.* (2017). The InSIDE assumption implies that the genetic association with the exposure should not be related to the pathway between the instrument and the outcome variable that does not involve the exposure of interest. In such circumstance, the IVW method can yield unbiased causal estimators.

As previously mentioned, we can assess violations of Assumption IV1 using F-statistic. For Assumptions IV2 and IV3, we detect violations by measuring the heterogeneity in the ratio estimates. The Cochran's Q statistic is used to assess the heterogeneity of causal estimates across different genetic instruments.

$$Q = \sum_{i=1}^{M} w_j(\hat{\beta}_c - \hat{\beta}_{IVW})^2 \tag{1.8.5}$$

The Q statistic follows a chi-squared distribution with the degree of freedom (ν) equal to the number of IVs minus 1 ($M - 1$) under the null hypothesis, which states that there is no heterogeneity between instruments, and the effect estimates are consistent. If the p-value from the chi-squared test is significant, it suggests the presence of heterogeneity. To address heterogeneity, researchers can use statistical methods that account for effect heterogeneity, such as random-effects meta-analysis or MR-Egger regression. In two-sample MR, the multiplicative random-effects model is more prevalent than the additive version, as it maintains the weight for each ratio estimate within the fixed-effects model (Bowden *et al.*, 2017).

The estimate from the multiplicative random-effects model will be the same as that from the fixed-effects model because the weights remain the same. However, the variance will be inflated by a scale parameter ϕ to account for the heterogeneity. From Equation (1.8.6), we see that the IVW estimate can also be obtained from a weighted linear regression of the G-Y association with the G-X association using the weight $\text{se}(\hat{\beta}_{g_j y})^{-2}$. The multiplicative random-effects IVW method provides valid causal estimates when the average pleiotropic effect across all genetic variants is zero (referred to as balanced pleiotropy).

$$\begin{aligned}
\text{G-X}: \quad & \hat{\beta}_{g_jx} = \beta_{g_jx} \\
\text{G-Y}: \quad & \hat{\beta}_{g_jy} = \beta_c\beta_{g_jx} + \phi^{1/2}\,\mathrm{se}(\hat{\beta}_{g_jy})\varepsilon_j, \quad \mathrm{var}(\varepsilon_j) = 1
\end{aligned} \tag{1.8.6}$$

In this case, the variance of the multiplicative random-effects IVW estimator will be:

$$\mathrm{var}(\hat{\beta}_{IVW}) = \frac{\hat{\phi}}{\sum w_j} \tag{1.8.7}$$

The scale parameter ϕ is estimated by $\hat{\phi} = \frac{Q}{M-1}$.

1.8.2 MR-Egger

The MR-Egger method serves as an alternative statistical approach within Mendelian randomization to estimate causal effects when there is directional pleiotropy (average pleiotropic effect is non zero) that challenges the exclusion restriction assumption and introduces bias in the IVW estimator (Bowden *et al.*, 2015). This method involves a straightforward modification to the weighted linear regression as described earlier. Unlike the IVW method, where the intercept term is fixed at zero, MR-Egger allows for the estimation of the intercept term as part of the analysis.

$$\hat{\beta}_{g_jy} = \beta_0 + \beta_c\beta_{g_jx} + se(\hat{\beta}_{g_jy})\varepsilon_j, \quad \mathrm{var}(\varepsilon_j) = 1 \tag{1.8.8}$$

The intercept term (β_0) in Equation (1.8.8) captures the average pleiotropic effect across the genetic instruments on the outcome variable, independent of the exposure. It assesses the directional pleiotropy by conducting a test on the intercept. A non-zero intercept in MR-Egger indicates the presence of directional pleiotropy and suggests that the IVW estimator is biased.

When the assumptions of InSIDE and NOME are fully satisfied, MR-Egger provides an unbiased estimator for the causal effect. However, if InSIDE is met while NOME is violated, the estimator becomes biased, leading to what is known as regression dilution bias. This bias causes the MR-Egger slope to be attenuated towards zero. In addition to InSIDE and NOME, MR-Egger relies on the variation in gene-exposure associations, denoted as VIS (Variation in Instrument Strength). If there is minimal variation in the effect of the genetic instruments on the exposure variable, the estimate may suffer from dilution bias. The dilution bias can be assessed using I^2 statistic (Bowden *et al.*, 2016a), which quantifies the percentage of total variation in G-X association across studies attributed to heterogeneity rather than chance (Higgins *et al.*, 2003). When the G-X associations exhibit substantial variation and the measurement error in estimating the variance of G-X association estimate is relatively minor compared to the actual variability, the I^2 value tends to approach 1, indicating minimal impact from NOME violation. On the other hand, if the G-X associations show similar magnitudes or their estimates lack precision, the I^2 value can markedly fall below 1, leading to a pronounced dilution effect (Bowden *et al.*, 2016a).

1.8.3 Median-based Estimate

Recall that the IVW estimator remains unbiased only when all IVs fulfill the three Mendelian randomization assumptions or when the average pleiotropy effect is zero. In contrast, the MR median-based method provide a consistent estimator of the causal effect even when up to 50% of the IVs are not valid (Bowden *et al.*, 2016b).

This approach demonstrates greater robustness to the presence of invalid instruments, making it a valuable alternative in scenarios where the traditional IVW method might be affected by pleiotropy or other biases.

The simple median estimator often lacks efficiency, especially when the precision of individual estimates varies significantly. To address this issue, Bowden *et al.* (2016b) also introduced a weighted median estimator, which takes into account the varying precision of the ratio estimates. The weighted median is defined as follows: Let w_j represent the standardized weight assigned to the j-th ordered ratio estimate, and let s_j be the sum of standardized weights up to. The weighted median estimator calculates the median of a distribution with $\hat{\beta}_j$ as its $p_j = 100(s_j - w_j/2)$-th percentile. For other percentile values, linear extrapolation is performed between neighboring ratio estimates (Bowden *et al.*, 2016b). Similar to the IVW method, the weighted median estimator uses the inverses of the estimated variance of the ratio estimators as weights, which are then standardized. The variance of the median-based estimator and confidence intervals are computed using the parametric bootstrap method. In comparison, MR-Egger regression can yield a consistent estimator even when all genetic variants are invalid instrumental variables, whereas the weighted median method requires at least 50% of the variants are valid IVs. However, the advantage of the weighted median approach lies in its ability to accommodate a broader range of IV assumption violations, including potential breaches of the INSIDE assumption upon which MR-Egger heavily relies (Bowden *et al.*, 2016b).

1.8.4 Mode-based Estimate

The mode-based estimate (MBE) offers an alternative approach to obtain a consistent causal effect estimator, especially when IVs are susceptible to pleiotropy bias. This method is based on the Zero Modal Pleiotropy Assumption (ZEMPA), which states that the largest weights among subsets of IVs are contributed by valid instruments, irrespective of the presence of horizontal pleiotropy (even when the INSIDE assumption is not met) (Hartwig *et al.*, 2017).

The MBE is computed as using the mode of the smoothed empirical density function of all ratio estimates $\hat{\beta}_{cj}$ as the causal effect estimate. This approach allows for different weights to individual IVs. For example, they refer to the mode of the unweighted empirical density function as "simple MBE" and the mode of inverse-variance weighted empirical density function as the "weighted MBE" (Hartwig *et al.*, 2017). In simple MBE, each IV is assigned an equal weight $w_1 = w_2 = \cdots = 1/M$, wheras in weighted MBE, w_j is the standardized inverse variance weight. The MBE causal estimate takes the value that maximizes $f(x)$:

$$f(x) = \frac{1}{h\sqrt{2\pi}} \sum_{j=1}^{M} w_j \text{exp} \left[-\frac{1}{2} \left(\frac{x - \hat{\beta}_{cj}}{h} \right)^2 \right] \tag{1.8.9}$$

Here, the parameter h controls the balance between bias and variance in the MBE, where larger values of h result in increased precision but also higher bias.

1.8.5 Other Methods

The MR Pleiotropy Residual Sum and Outlier (MR-PRESSO) method also requires that at least 50% of the variants are valid instruments and relies on the InSIDE assumption (Verbanck *et al.*, 2018). It addresses the issue of horizontal pleiotropy by identifying and removing outliers that exhibit pleiotropic effects. The MR-PRESSO method involves two key steps. Firstly, it identifies potential outlier genetic variants by calculating their residual sum of squares (RSS) after iteratively removing each variant from the analysis. If the RSS shows a significant decrease compared to a simulated expected distribution, it indicates the presence of pleiotropy, and the variant is considered an outlier and excluded from further analysis. In the second step, MR-PRESSO addresses the impact of these outliers by re-estimating the causal effect after removing them from the analysis.

Like MR-PRESSO, the MR-Lasso method (Rees *et al.*, 2019) also removes some IVs for further analysis. It considers the objective function for the MR-Egger model and adds a Lasso penalty term for the intercept of the regression. If the removed genetic variants which are detected as the outlier or heterogeneous estimates are valid IVs, then it would be not appropriate to remove them. Instead, we can downweight the contributions of them rather than remove them directly (Rees *et al.*, 2019). Two methods, MR-Robust and MR-RAPS (Robust Adjusted Profile Score), employ a downweighting approach for handling outliers rather than directly removing them (Slob and Burgess, 2020). MR-Robust provides robustness to outliers by using MM-estimation instead of ordinary least squares for the IVW method. MM-estimation consists of an initial S-estimate followed by an M-estimate of regression (Koller and Stahel, 2011), combined with Tukey's biweight loss function (Rees *et al.*, 2019). On the other hand, MR-RAPS, introduced by Zhao *et al.* (2020), involves adjusting the profile score and addresses pleiotropy and extreme outliers through robustification. To achieve robustness, Tukey's biweight loss function is applied to the adjusted profile score.

MRMix (Qi and Chatterjee, 2019) and contamination mixture (Burgess *et al.*, 2020) are two robust methods that employ a mixture model approach to handle outliers. MRMix categorizes a SNP into four different types of effects: (1) direct effect on X and an indirect effect on Y only through X, (2) direct effects on both X and Y, (3) direct effect on Y only, and (4) no relationship with either X or Y. It estimates causal effects using a spike-detection algorithm by fitting the mixture model $\hat{\beta}g_jy - \beta_c\hat{\beta}g_jx \sim \pi_0 N(0, \sigma_0^2) + \pi_1 N(0, \sigma_1^2)$ and identifying β_c that maximizes the probability concentration at the null component $N(0, \sigma_0^2)$ corresponding to valid IVs (Qi and Chatterjee, 2019). This approach requires ZEMPA. The contamination mixture method also employs a mixture model, characterizing two clusters of instruments: one cluster for valid IVs and another for invalid IVs. For a variant to be a valid instrument, its ratio estimate $\hat{\beta}_{cj}$ is assumed to be normally distributed around the true causal effect β_c with variance equal to $\mathrm{var}(\hat{\beta}_{cj})$. If a variant is not a valid instrument, its ratio estimator is assumed to be normally distributed around zero with a wider variance $\psi^2 + \mathrm{var}(\hat{\beta}_{cj})$, where ψ^2 represents the variance of the estimates from invalid IVs. The analyst specifies this parameter.

1.9 Three-sample Genome-wide Design

A three-sample genome-wide design (Zhao *et al.*, 2019) involves using three separate non-overlapping GWAS studies to analyze causal relationships (Figure 1.6). In this approach, the first GWAS study, known as the Discovery GWAS, is utilized to select IVs for the MR analysis. The second GWAS focuses on estimating the association between the genotype and exposure using the selected IVs. The third GWAS is then used to estimate the association between the genotype and outcome using the same set of selected IVs .

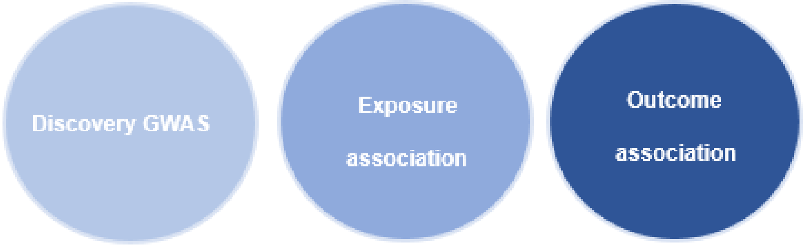
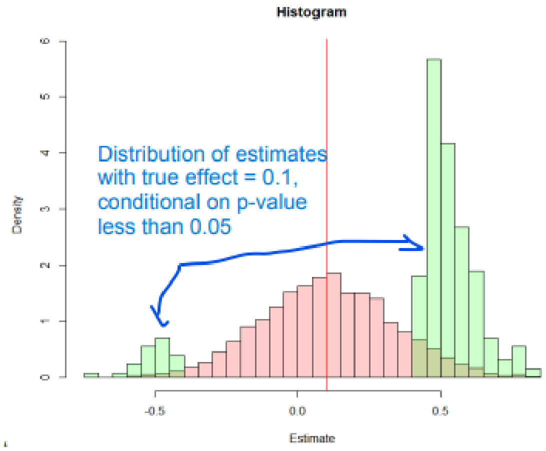
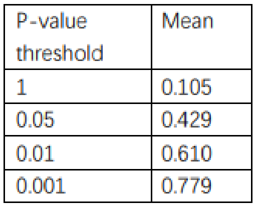


Figure 1.6: Three-sample design

Employing three distinct GWAS samples helps to avoid the winner's curse (detailed in Chapter 2). In MR, it is common to select the variants with the most significant associations as IVs. However, if the significant estimates from the Discovery GWAS are used for MR analysis rather than an independent dataset, the G-X association is likely to be biased away from 0. This phenomenon is known as the winner's curse. Importantly, this bias in the G-X association can subsequently lead to biased estimator of causal effect.

To illustrate this issue, let's consider an example where y is a linear combination of x and random error: $y = \beta x + \varepsilon$, with the true effect β set to 0.1. Running an OLS regression $y \sim x$ multiple times, we obtain a distribution of estimates $(\hat{\beta})$ as shown in Figure 1.7 (red area). This distribution is asymptotically normally distributed around the true effect (0.1).





P-value threshold	Mean
1	0.105
0.05	0.429
0.01	0.610
0.001	0.779

Figure 1.7: Example of winner's curse

However, when we condition the estimates on a threshold of p-values for the t-statistics less than 0.05 (highlighted in green), we observe more weights above the true effect (0.1) than below it. The right-hand table displays the mean estimates with different thresholds. As the threshold becomes more stringent, the mean of estimates is more away from the true effect. Mathematically, we describe the winner's curse as follows: $E(|\hat{\beta}| \; ||\hat{\beta}| > c\hat{\sigma}) > |\beta|$, where $c > 0$ and $\hat{\sigma}$ is the standard error of $\hat{\beta}$.

1.10 Software and Public Databases

Numerous publicly accessible databases and software packages are available, which can function as data sources and facilitate the Mendelian randomization analyses. In the following section, we will explore these databases and software packages in detail.

The GWAS Catalog (`https://www.ebi.ac.uk/gwas/`) is a comprehensive database that consolidates summary statistics from genome-wide association studies. It includes information on genetic variants associated with various traits and diseases, which can be employed as instrumental variables in MR analyses. As of Oct 11, 2023, the GWAS Catalog comprises 6586 publications, 555899 top associations, and 65846 full summary statistics.

The Neale lab (`http://www.nealelab.is/uk-biobank/`) provides GWAS results for the UK Biobank data, featuring 4203 phenotypes. The GWAS results are derived from a subset of 361,194 samples in the round 2 data.

PhenoScanner is a curated database housing over 65 billion associations between human genotypes and phenotypes, encompassing more than 150 million unique genetic variants, predominantly SNPs (Staley *et al.*, 2016). It is accessible at `www.phenoscanner.medschl.cam.ac.uk` and offers a convenient R command, "pheno_input" within the *MendelianRandomization* package. By utilizing this function, users can extract summarized data from PhenoScanner, providing the rsid (which stands for reference SNP ID number, a unique identifier assigned to a SNP) for a SNP, exposure and outcome variables, associated PubMed IDs, individual ancestry for exposures and outcomes (e.g., "European", "Mixed", "Asian" populations), and genetic variant correlations. The output includes rsid, genetic association with exposure and outcome, and corresponding standard errors.

MR-BASE is a platform integrating a database of thousands of GWAS summary datasets with a web interface and an R package called *TwoSampleMR* for automated causal inference through MR (Hemani *et al.*, 2018). The database contains 11 billion SNP associations from 1673 GWAS and is regularly updated. The MR-BASE web application facilitates MR analysis in three steps: selecting instruments for the exposure, choosing instruments for the outcome, and conducting the MR analysis. MR-BASE offers various MR analysis methods, including Wald ratio, maximum likelihood, MR

Egger, median-based, mode-based, IVW, and MR-RAPS. The entire process can be performed in R using the *TwoSampleMR* package.

The MRC IEU OpenGWAS database contains 126 billion genetic associations from 14582 complete GWAS datasets. It is freely accessible at `https://gwas.mrcieu.ac.uk` and can be accessed through an application programming interface (API) or via R (ieugwasr) or Python (ieugwaspy) packages (Elsworth *et al.*, 2020). The *TwoSampleMR* package utilizes this database to obtain data, with functions such as "extract_instruments" and "extract_outcome_data" for extracting instruments for exposure and outcome, respectively. The extracted datasets are then harmonized using the "harmonise_data" function before conducting MR analysis.

MendelianRandomization and *TwoSampleMR* are two software packages utilized not only for extracting summary statistics but also for conducting MR analysis. Although both packages offer functions for common MR methods such as IVW, median-based, and mode-based methods, each package includes some unique methods that are absent in the other. For instance, *MendelianRandomization* incorporates MR-Lasso, along with offering multivariable version analysis for MR-Egger, median-biased MR, and MR-Lasso. On the other hand, *TwoSampleMR* provides functions for MR-RAPS and MR-PRESSO. Furthermore, *TwoSampleMR* includes capabilities for selecting exposure instruments through LD clumping and harmonizing effect sizes of instruments on exposures and outcomes to align with the same reference allele.

In this thesis, we acquired the summary statistics for the UK Biobank information from Neale lab. Additionally, we utilized the MRC IEU OpenGWAS, facilitated by the R package (ieugwasr), to retrieve data from the Genetic Investigation of Anthropometric Traits (GIANT) consortium and the International Genomics of Alzheimer's

Project (IGAP). These datasets are examined in practical examples presented in subsequent chapters.

1.11 The Structure of This Thesis

This chapter establishes the fundamental concepts that underpin the novel techniques developed in this thesis. In the subsequent three chapters, we will primarily address bias in the causal estimation arising from winner's curse, overlapping samples and horizontal pleiotropy.

In Chapter 2, we focus on the winner's curse in MR, which arises from the selection of genetic markers based on their significance or ranking in GWAS. We compare the performances of various methods for dealing with the winner's curse in GWAS when applied to MR. Additionally, we adapt the existing bootstrap-based BR-squared method (Faye *et al.*, 2011) to work with summary-level data.

In Chapter 3, we tackle the challenge of sample overlapping with weak instrument bias. We develop a method that accounts for the correlation due to overlapping and weak instrument bias, building on the work of Bowden *et al.* (2019).

In Chapter 4, we introduce a novel method for identifying invalid IVs that exhibit pleiotropy. We propose employing the bootstrap method to account for the selection process in dealing with data-driven instrument variable selection.

In the concluding chapter, we will provide a comprehensive overview of the entire thesis and explore potential avenues for future research.

Chapter 2

Winner's Curse in Mendelian Randomization

The chapter primarily focuses on examining the impact of the winner's curse on causal estimation, which arises when IVs are chosen based on their significance or ranking. It delves into the investigation of the winner's curse within the context of Mendelian Randomization and compares various methods for handling this issue in Genome Wide Association Studies (GWAS) when applied to MR. In the course of this research, we recognized the need to extend the BR-squared method. Consequently, we put forth a modified bootstrap method derived from the BR squared approach, tailored to address this bias with summary-level data. The adapted method performs similarly to the original one. We demonstrate that the correction method effectively mitigates bias, albeit at the expense of increased variability and broader confidence intervals.

The structure of this chapter can be outlined as follows. First, we introduce the existing methods for correcting the winner's curse bias in GWAS. We adapt an existing bootstrap method to work in common situations when only summary-level data is available. Next, we perform simulation studies to assess the potential bias from winner's curse in Mendelian randomization and apply the existing methods to correct the genetic variant-exposure association which are then used in causal estimation. Finally, we apply the approaches to real data investigating the causal relationship between Body Mass Index (BMI) and schizophrenia risk, and between Low Density Lipoprotein Cholesterol (LDL-C) and Alzheimer's Disease (AD) risk. These two examples are also explored in Rees *et al.* (2019).

2.1 Introduction

Mendelian randomization is a method used to estimate causal effects using genetic markers. The genetic markers to be used in Mendelian randomization are typically chosen from GWAS based on their significance in a test of association with the causal exposure. However, regardless of the selection criteria employed, an upward bias is present when the association effect size is estimated from the discovery GWAS (Göring *et al.*, 2001; Garner, 2007). This bias phenomenon is commonly referred to as the "winner's curse". The upward bias in the estimated association between the genotype and exposure causes the causal estimator biased toward 0. This bias occurs because the estimated genetic association with exposure appears in the denominator of the ratio estimator (as shown in Equation (1.8.1)).

The winner's curse problem has been widely studied in the Genome Wide Association Studies (Sun and Bull, 2005; Garner, 2007; Jeffries, 2007; Zöllner and Pritchard, 2007; Ghosh *et al.*, 2008; Zhong and Prentice, 2008; Xiao and Boehnke, 2009; Faye *et al.*, 2011; Sun *et al.*, 2011; Xu *et al.*, 2011; Ferguson *et al.*, 2013; Zhou and Wright, 2016; Bigdeli *et al.*, 2016; Forde *et al.*, 2023). Many methods have been developed to address the winner's curse, with the earliest one originating from Sun and Bull (2005), which employs the bootstrap method to decrease selection bias. The bootstrap shrinkage estimator they proposed is the average of the difference between the within-sample and out-of-sample bootstrap estimates. Jeffries (2007) used a comparable bootstrap method to address the issue of bias in ranking. Faye *et al.* (2011) built on the work of Sun and Bull (2005) and proposed two crucial adjustments for the previous bootstrap estimator. They called their method BR-squared (Bias-Reduced estimates via Bootstrap Re-sampling) (Sun *et al.*, 2011). These two modifications enabled the method to consider differences in variance associated with the MAF of a SNP and account for the negative correlation between within-sample and out-of-sample estimates. BR-squared is computationally expensive since it requires individual-level data for bootstrapping. Building upon the BR-squared method, Forde *et al.* (2023) devised a new approach that employs the bootstrap method to address winner's curse bias, with the added advantage of only requiring summary-level data. Zhou and Wright (2016) proposed the projack method (Prediction by Re-Ordered Jackknife and Cross-Validation K-fold). This is a method that uses resampling techniques to provide corrected estimates of the ranked z-statistics. It works for both individual-level and summary-level data. Ghosh *et al.* (2008) proposed a conditional likelihood approach in which the association estimates for the variants which are declared as significant at specified threshold level are adjusted. Similarly, Zhong and Prentice (2008) derived three bias-reduced estimators: the first one maximizes the conditional likelihood at the naive estimate, the second one ensures the conditional expectation equals the naive estimate, and the third one has the median at the observed. They also proposed corresponding weighted estimators that combine these corrected estimators (chosen from the three) with uncorrected estimators to address selection bias. Likelihood-based techniques have also been developed specifically for case-control studies, such as those proposed by Zöllner and Pritchard (2007) and Xiao and Boehnke (2009). Zöllner and Pritchard (2007) focused on estimating the frequency of a variant and its penetrance parameter (the estimated probability of disease for the genotypes), whereas Xiao and Boehnke (2009) focused on estimating the difference in allele frequency.

An Emprical Bayes (EB) method (Ferguson *et al.*, 2013) was motivated by Tweedie's formula (Efron, 2009). It utilizes empirical estimates of the density of z statistic and is well-suited for a large number of variants. However, this method is less precise in the extreme tails of the distribution, where conditional likelihood methods are more accurate. To address this limitation, they proposed a method that combines the estimator with the conditional likelihood estimator (Ghosh *et al.*, 2008; Zhong and Prentice, 2008) based on the lengths of their respective 95% confidence and credible intervals. The new estimator combines the strengths of both approaches but requires more computational complexity. Bigdeli *et al.* (2016) introduced the FDR Inverse Quantile Transformation (FIQT). Xu *et al.* (2011) proposed a hierarchical Bayes method that uses a spike-and-slab prior to account for the possibility that the significant test result is due to chance. It can also deal with the case where only the summary statistic is available. However, it might also be time-consuming since it involves using Markov Chain Monte Carlo (MCMC) methods for sampling from a probability distribution.

The research on winner's curse in MR is limited. A recent study by Jiang *et al.* (2023) evaluated the impact of the winner's curse in MR by randomly dividing the

UK-biobank data into three equally sized subsets. The first subset is for selecting the IVs for exposure and the second subset is for estimating the effect size for IVs that are selected in the first sample. The last subset is for estimating the genetic variant-outcome association. Because the selection and estimation come from two non-overlapping datasets, the winner's curse can be avoided. Jiang *et al.* (2023) examined the causal relationship between BMI and coronary artery disease, showing that the winner's curse affected the causal estimates, though the impact was not substantial. Nevertheless, it remains valuable to investigate whether this holds true in various scenarios and to assess the overall impact of the winner's curse on causal inference. It is worth noting that our work on the winner's curse in MR predates the publication of Jiang *et al.* (2023).

To mitigate the impact of the winner's curse in Mendelian randomization, one can utilize a three-sample design similar to the approach employed by Jiang *et al.* (2023). Nonetheless, in practice, it's seldom achievable to find three distinct large participant samples that are similar enough. Even if such samples are available, the reduction in sample size, and consequently statistical power, due to excluding the discovery GWAS from genetic association estimation may be undesirable. Alternatively, one can apply corrective techniques to modify the estimates. More recently, Mounier and Kutalik (2023) introduced a method called MR-lap, which addresses weak instrument bias and winner's curse, while also considering the potential occurrence of sample overlap. They proposed a corrected IVW estimator when the genetic variants are selected based on the significance threshold. Unlike other estimators developed in GWAS which correct the bias in the estimation of each genetic effect size, MR-lap directly accounts for the impact of the winner's curse on the IVW estimator.

This chapter aims to assess the current approaches in GWAS for correcting the winner's curse when applied to MR. We modify the existing BR-squared method to work with summary-level data. Furthermore, we explore the influence of the winner's curse on causal estimation in MR.

2.2 Materials and Methods

We assume that the exposure $\mathbf{X}$ and outcome $\mathbf{Y}$ come from the following Model (2.2.1).

$$\begin{aligned} \mathbf{X} &= \mathbf{G}\boldsymbol{\alpha} + \mathbf{U}\alpha_u + \boldsymbol{\varepsilon_x} \\ \mathbf{Y} &= \mathbf{X}\beta_c + \mathbf{U}\beta_u + \boldsymbol{\varepsilon_y} \end{aligned} \tag{2.2.1}$$

The causal effect is denoted by β_c. The exposure $\mathbf{X}$ is a linear combination of genetic variants $\mathbf{G}$, a confounder $\mathbf{U}$, and random errors $\boldsymbol{\varepsilon_x}$. The outcome $\mathbf{Y}$ is a linear function of exposure, confounder and random errors $\boldsymbol{\varepsilon_y}$. Genetic variants that are chosen to be the instrument variables in MR analysis are assumed to be independent.

We also assume that

$$Z_j = \frac{\hat{\beta}_{g_jx}}{\widehat{\text{se}(\hat{\beta}_{g_jx})}} \sim N\left(\mu_j = \frac{\beta_{g_jx}}{\widehat{\text{se}(\hat{\beta}_{g_jx})}}, 1\right) \tag{2.2.2}$$

where β_{g_jx} is the true effect size of genetic variant-exposure association, $\hat{\beta}_{g_jx}$ is the estimate of exposure association for variant j, $\widehat{\text{se}(\hat{\beta}_{g_jx})}$ is the estimated standard error for the estimate $\hat{\beta}_{g_jx}$. The assumption shown in expression (2.2.2) relies on that $Z_j - \mu_j = \frac{\hat{\beta}_{g_jx} - \beta_{g_jx}}{\widehat{\text{se}(\hat{\beta}_{g_jx})}} \sim N(0, 1)$ for increasing sampple size (Wald, 1943). For a large sample, $\widehat{\text{se}(\hat{\beta}_{g_jx})}$ remains relatively stable across in multiple data realizations.

Let $z_{(1)} \geq z_{(2)}, ..., \geq z_{(M)}$ denote the order statistics, with their means that are also arranged in order and denoted by $\boldsymbol{\mu}_{(.)} = \{\mu_{(1)}, ..., \mu_{(M)}\}$, where $\mu_{(j)} \geq \mu_{(j+1)}$.

This chapter focuses on the common scenario where the winner's curse occurs in the estimate of exposure association. Specifically, it involves the selection of genetic variants that demonstrate the most significant associations with the exposure, followed by the correction of exposure association estimates for these variants. Then, the IVW method is used to estimate the causal effect. Recall that in section 1.7.1, we define IVW as

$$\hat{\beta}_{\text{IVW}} = \frac{\sum_{j=1}^{M} w_j \hat{\beta}_{c_j}}{\sum_{j=1}^{M} w_j} \tag{2.2.3}$$

where w_j is the inverse variance for the j-th ratio estimate. And the variance of IVW estimator is

$$\text{var}(\hat{\beta}_c) = \frac{1}{\sum_{j=1}^{M} w_j} \tag{2.2.4}$$

Let $\hat{\beta}_{\text{correct,IVW}}$ be the corrected estimate of IVW estimator, with the corrected exposure association estimate denoted as $\hat{\beta}_{\text{correct},g_j x}$, and let $w_{\text{correct},j}$ be the corrected weight for each variant j. Since $\hat{\beta}^2_{\text{correct},g_j x}$ is not greater than $\hat{\beta}^2_{g_j x}$ due to correction, the corrected weight for j-th genetic variant $\sum w_{\text{correct},j} = \sum \frac{\hat{\beta}^2_{\text{correct},g_j x}}{\text{se}^2(\hat{\beta}_{g_j y})}$ is also not greater than the original weights $\sum w_j = \sum \frac{\hat{\beta}^2_{g_j x}}{\text{se}^2(\hat{\beta}_{g_j y})}$. Theoretically, the corrected IVW estimator should have a higher variance compared to the naive estimator.

2.2.1 Methods for Correcting the Winner's Curse

We focus our attention on specific methods, including Ghosh's methods, FIQT, pro-jack, Forde's bootstrap method, and BR-squared. Not all methods mentioned earlier are under examination due to various factors, including time constraints, computational complexity, and the fact that certain methods were developed for case-control studies. However, we anticipate the possibility of exploring these omitted methods in the future.

Conditional likelihood method

Ghosh *et al.* (2008) corrected the estimated association for the variants which are declared at the pre-specified threshold based on a conditional likelihood approach. The conditional likelihood function for u_j given the variant j is significant is given by:

$$L_c(\mu_j) = p_{\mu_j}(z_j||Z_j| > c) = \frac{p_{\mu_j}(z_j)}{p_{\mu_j}(|Z_j| > c)} = \frac{\phi(z_j - u_j)}{\Phi(-c + \mu_j) + \Phi(-c - \mu_j)} \tag{2.2.5}$$

where c is the cut-off value associated with a specified significant threshold; ϕ and Φ are the probability density function (p.d.f) and cumulative distribution function (c.d.f) of the standard normal distribution, respectively.

Three estimators are devised based on the conditional likelihood function (Equation (2.2.5)). The first estimator is the maximum likelihood estimator (MLE):

$$\tilde{\mu}_{j1} = \arg\max L_c(\mu_j) \tag{2.2.6}$$

The second estimator is called the mean of normalized conditional likelihood (Normalized):

$$\tilde{\mu}_{j2} = \frac{\int_{-\infty}^{\infty} \mu_j L_c(\mu_j) d\mu_j}{\int_{-\infty}^{\infty} L_c(\mu_j) d\mu_j} \tag{2.2.7}$$

While $\tilde{\mu}_{j2}$ has a higher mean squared error (MSE) than $\tilde{\mu}_{j1}$ for values of μ_j that are close to zero, it performs better for values of μ_j that are farther away from zero.

The third estimator is the average of the MLE and Normalized estimators. It is called the compromise estimator, which incorporates the advantages of two estimators:

$$\tilde{\mu}_{j3} = (\tilde{\mu}_{j1} + \tilde{\mu}_{j2})/2 \tag{2.2.8}$$

To make the Ghosh method compatible with rank-based selection, we introduce a modification in which the cut-off value c for $|z|$ is not pre-determined but is rather selected based on the observed $|z|$ of the top variants. Specifically, if we aim to select 30 variants, for example, we would choose c as the $|z|$ value of the 30th ranked variant.

FDR Inverse Quantile Transformation (FIQT)

The FIQT technique, as described in Bigdeli *et al.* (2016) is a very simple method that comprises just two steps. Initially, the p-values for all variants are subjected to a multiple testing correction, such as the False Discovery Rate (FDR) method. Following this, the modified z-statistics are computed by applying an inverse normal distribution to the adjusted p-values. It can be formulated as:

$$z_j^* = \text{sign}(z_j)\Phi^{-1}(1 - p^*/2) \tag{2.2.9}$$

where p^* denotes the FDR adjusted p-values. Then the corrected estimate for association is obtained by multiplying z_j^* by the corresponding standard error $\hat{\sigma}_{g_jx}$.

Projack: Predicion by Re-Ordered Jackknife and Cross-Validation K-fold

The projack method (Zhou and Wright, 2016) uses cross-validation by dividing the columns into K equally sized subsets and holding out each subset sequentially. During this process, elements of column K are rearranged based on the row means of the remaining $K-1$ columns. To improve the stability of the results, this cross-validation process is repeated numerous times using multiple random K-fold data partitions. Finally, the estimates for the ordered parameters, denoted as $\boldsymbol{\mu}_{(.)} = \{\mu_{(1)}, ..., \mu_{(M)}\}$ are obtained by averaging over the results of those random partitions.

The basic idea for the above process is to construct a matrix, with element $C_{jk} \sim N(\mu_j, \sqrt{K})$, so that $E(z_j) = E(\sum_k C_{jk}/K) = \mu_j$ and $\text{var}(z_j)$ is equal to 1.

They also introduce an alternative approach that does not require the cross-validation on individual-level data and only relies on summary-level data to achieve the desired properties. The method can be described as follows:

- Simulate a vector $\mathbf{z}' = \mathbf{z} + \boldsymbol{\gamma}$, where each element in $\boldsymbol{\gamma}$ is independently drawn from $N(0, \sqrt{1/(K-1)})$.
- Compute a new vector $\mathbf{c} = K\mathbf{z} - (K-1)\mathbf{z}'$, and reorder to create $\mathbf{d}$, where re-ordering is based on $\mathbf{z}'$.
- Repeat the first two steps, averaging over many ordered simulated vectors $\mathbf{d}$ to obtain the estimator for the ordered parameter $\mu_{(j)}$.

Finally, we multiply the corrected test statistic by the original standard error to get the corrected estimate of G-X association.

BR-squared method

The BR-squared method orinigated from the shrinkage estimator (Sun and Bull, 2005):

$$\hat{\beta}^*_{\text{boot}(k)} = \hat{\beta}_{N(k)} - \frac{\sum_{r=1}^{R}(\hat{\beta}^*_{Dr(k)} - \hat{\beta}^*_{Er(k)})}{R} \tag{2.2.10}$$

Where $\hat{\beta}_{N(k)}$ is the naive estimate for the kth ranked variant selected in the original sample. $\hat{\beta}^*_{Dr(k)}$ and $\hat{\beta}^*_{Er(k)}$ denote the within-sample and out-of-sample bootstrap estimate for the kth ranked variant selected in the rth bootstrap sample. $\hat{\beta}^*_{Dr(k)}$ can be viewed as an estimate for the naive estimator while the out-of-sample estimate $\hat{\beta}^*_{Er(k)}$ imitates an estimate that would be obtained from an independent sample. Despite the fact that the out-of-sample estimate $\hat{\beta}^*_{Er(k)}$ and the within-sample estimate $\hat{\beta}^*_{Dr(k)}$ are based on two sets of observations that do not overlap, they are negatively correlated due to the finite size of the original sample. This is because the observations that are excluded from one sample must necessarily be included in the other. To fix this problem, Faye *et al.* (2011) introduce $\hat{\beta}^\dagger_{Er(k)}$ that accounts for the correlation between $\hat{\beta}^*_{Dr(k)}$ and $\hat{\beta}^*_{Er(k)}$. In addition, the MAF of the kth ranked variant selected from the bootstrap sample might differ from the MAF of the kth variant in the original sample. To account for the negative correlation between $\hat{\beta}^*_{Dr(k)}$ and $\hat{\beta}^*_{Er(k)}$ and the difference in MAF, Faye *et al.* (2011) propose a modified bootstrap shrinkage estimator:

$$\begin{aligned} \hat{\beta}^*_{\text{boot}(k)} &= \hat{\beta}_{N(k)} - \frac{\sum_{r=1}^{R}(\hat{\beta}^*_{Dr(k)} - \hat{\beta}^\dagger_{Er(k)})\sqrt{2p_{r(k)}(1-p_{r(k)})}}{R\sqrt{2p_{(k)}(1-p_{(k)})}} \\ \text{where } \hat{\beta}^\dagger_{Er(k)} &= \hat{\beta}^*_{Er(k)} - \frac{\hat{\sigma}^*_{DEr(k)}}{\hat{\sigma}^{2*}_{Dr(k)}}(\hat{\beta}^*_{Dr(k)} - \hat{\beta}_{Nr(k)}) \end{aligned} \tag{2.2.11}$$

where R is the number of bootstrap samples, $p_{(k)}$ is the MAF of the kth ranked variant in the original sample and $p_{r(k)}$ is the MAF for the kth ranked variant in the rth bootstrap sample. $\hat{\beta}_{Nr(k)}$ is the estimate from the original sample for the kth ranked variant selected in the rth bootstrap sample. (Note that $\hat{\beta}_{Nr(k)}$ is different from $\hat{\beta}_{N(k)}$ as the latter refers to the kth ranked variant selected in the original sample). $\hat{\sigma}^{2*}_{Dr(k)}$ is the variance for $\hat{\beta}^{*}_{Dr(k)}$, and $\hat{\sigma}^{*}_{DEr(k)}$ is the covariance between $\hat{\beta}^{*}_{Er(k)}$ and $\hat{\beta}^{*}_{Dr(k)}$. The variance and covariance terms for each variant are estimated by taking a seperate set of bootstrap samples.

Since this method depends on individual-level data, it is considerably more computationally intensive than the methods that only require summary statistics. The following two methods are based on the same idea as BR-squared but only require summary statistics.

Forde's Bootstrap Method

Forde *et al.* (2023) developed a bootstrap resampling method based on the summary statistics. The procedure for estimating the effect size of the variant with the kth ranked z-statistic can be described as follows:

- Generate the bootstrap estimate for variant j from the normal distribution based on the observed naive estimate.

$$\hat{\beta}^{*}_{j} \sim N(\hat{\beta}_j, \sigma(\hat{\beta}_j)) \tag{2.2.12}$$

- Then calculate the z-statistic for the bootstrap estimate for variant j:

$$z_j^* = \frac{\hat{\beta}_j^*}{\sigma(\hat{\beta}_j)} \tag{2.2.13}$$

- Order the z-statistics for the boostrap estimates in decreasing order. The kth largest z^* is denoted as $z^*_{(k)}$. Then the bias of variant k (with kth largest original z-statistic) is

  $$\text{bias}_k = \frac{\hat{\beta}^*_{A(k)} - \hat{\beta}_{A(k)}}{\sigma(\hat{\beta}_{A(k)})} \tag{2.2.14}$$

  where $\hat{\beta}^*_{A(k)}$ is the bootstrap value of the variant with the kth largest z^* value, $\hat{\beta}_{A(k)}$ and $\sigma(\hat{\beta}_{A(k)})$ are the orginal observed estimates of the same variant.

- Finally, fit a cubic smoothing spline to the data, with z-statistic (ordered) as inputs and bias_k as output. Let bias^*_k be the fitted value from this model as bias^*_k. Then, the corrected effect for the variant with the kth largest original z-statistic is defined as $\hat{\beta}^*_{(k)} = \hat{\beta}_{(k)} - \sigma(\hat{\beta}_{(k)}) * \text{bias}^*_k$.

BR-squared Based on Summary Statistics

We have also modified the BR-squared technique in a different way to work with summary statistics. Unlike Forde's approach that produces a single set of bootstrap estimates, we create two sets of estimates, one for within-sample bootstrap estimates and the other for out-of-sample bootstrap estimates. We also consider the negative correlation between these two estimates. Note that we normalize each variant before implementing our method so that there is no need to accout for the variability of MAF in our setting. The step-by-step procedures are explained below:

1. For each bootstrap sample r, generate pairs of estimates $(\hat{\beta}^*_{Drj}, \hat{\beta}^\dagger_{Erj})$ for each variant j from a bivariate normal distribution given the naive estimate $(\hat{\beta}_{Nj})$ from the original sample for variant j. The supplementary material of Faye *et al.* (2011) gives the result:

$$\left(\begin{pmatrix} \hat{\beta}^*_{Drj} \\ \hat{\beta}^\dagger_{Erj} \end{pmatrix} \middle| \hat{\beta}_{Nj} = B \right) \sim N \left(\begin{pmatrix} B \\ B \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{Dj} & 0 \\ 0 & \hat{\sigma}^2_{Ej}(1 - \hat{\rho}^2_{DEj}) \end{pmatrix} \right)$$

Since we normalized the genetic variants and exposure, we can write $\hat{\sigma}^2_{Dj} \approx \frac{1-B^2}{n}$, $\hat{\sigma}^2_{Ej} \approx \frac{1-B^2}{ne^{-1}}$, $\hat{\sigma}_{DEj} \approx \frac{-(1+2B^2)}{n}$, and $\hat{\rho}_{DEj} \approx \frac{\hat{\sigma}_{DEj}}{\hat{\sigma}_{Dj}\hat{\sigma}_{Ej}}$. See the details of derivation in Appendix A. It is worth noting that ne^{-1} is the approximate number of observations not included in the bootstrap sample.

2. Simulate the variance of within-sample bootstrap estimate $\hat{\sigma}^{2*}_{Drj} \sim \frac{\hat{\sigma}^2_{Dj}}{\nu}\chi^2_\nu$ for each bootstrap sample r, where ν is the degree of freedom.

3. Then we order the vectors $\hat{\beta}^*_{Dr}$ and $\hat{\beta}^\dagger_{Er(k)}$ based on z^* statistics for the within-sample bootstrap estimates $(\hat{\beta}^*_{Drj}/\hat{\sigma}^*_{Drj})$.

4. Repeat Steps 1-3 for R times. Then we can get the bootstrap estimate for each variant by

$$\hat{\beta}^*_{\text{boot}(k)} = \hat{\beta}_{N(k)} - \frac{\sum_{r=1}^{R}(\hat{\beta}^*_{Dr(k)} - \hat{\beta}^\dagger_{Er(k)})}{R} \tag{2.2.15}$$

where $\hat{\beta}^*_{Dr(k)}$ is the within-sample bootstrap estimate for the variant with kth largest z^* value in the rth bootstrap sample, $\hat{\beta}^\dagger_{Er(k)}$ is the corrected out-of-sample bootstrap estimate for the same variant in the rth bootstrap sample, and $\hat{\beta}_{N(k)}$ is the original estimate for the variant with kth largest z-statistic in the original sample.

Note that this method is compared to the original BR-squared approach as part of the simulation presented in Section 2.3.

2.3 Simulation Framwork

To evaluate the performance of the different correction methods, we conduct a simulation study using Model (2.2.1) to generate the data. Specifically, we set the G-X association α_i for variant i to be normally distributed with mean 0 and standard deviation 0.02. The causal effect β_c takes values of 0.2 or 0.05. The confounder effects on X and Y, denoted by α_u and β_u, are both set to 0.3. We assume that the genotype follows Hardy-Weinberg Equilibrium (HWE). Consequently, each genetic variant $\boldsymbol{G_j}$ $(j = 1, 2, ..., M)$ is independently generated from a binomial distribution with $n = 2$ and a probability p_j uniformly distributed between 0.1 and 0.5. We standardize $\boldsymbol{G_j}$ to have mean 0 and variance 1. The error terms ε_x and ε_y are independently generated from a normal distribution with mean 0 and variances chosen so that $\mathbf{U}$, $\mathbf{X}$, and $\mathbf{Y}$ all have unit variance.

We simulate 40,000 individuals with 100 genetic variants. The sample is split into half, with one half used for exposure association estimation and the other half used for outcome association estimation. Two selection criteria, namely threshold-based and rank-based, are evaluated by generating 250 data sets for each criterion. With threshold-based selection, three different thresholds are considered: $1 \cdot 10^{-4}$, $5 \cdot 10^{-4}$ and $1 \cdot 10^{-3}$. With rank-based selection, different number of top ranked variants are selected: 20, 25 or 30.

We utilize Ghosh's methods, FIQT, BR-squared, Projack, and Forde's method, for each situation to adjust the bias in the genetic association estimate with the exposure. Subsequently, we apply the modified estimate to estimate the causal effect by using IVW. Since we are aware that the magnitude of the corrected estimate of the genetic variant-exposure association should not be greater than the naïve estimate, we incorporate this restriction into all methods.

2.4 Results

The simulation result for a moderate causal effect is shown in Table 2.1 and Table 2.2, while the result for a small causal effect is shown in Table 2.3 and Table 2.4. For each scenario, we calculated the mean of IVW causal estimates, relative bias ((mean of IVW causal estimates - true causal effect) / absolute value of true causal effect), mean of the standard error (SE), standard deviation (SD), mean squared error (MSE), mean of interval lengths of 95% confidence intervals, coverage of 95% confidence intervals and power for the naive method, projack, BR-squared, FIQT, Forde's method, and Ghosh's three methods (conditional MLE, the mean of the normalized likelihood, and compromise).

The comparison between our adapted version of the BR-squared method and the original BR-squared method is illustrated in Figure 2.1. It can be observed that all points are clustered around the line, indicating that our method is similar to the original method in terms of performance.

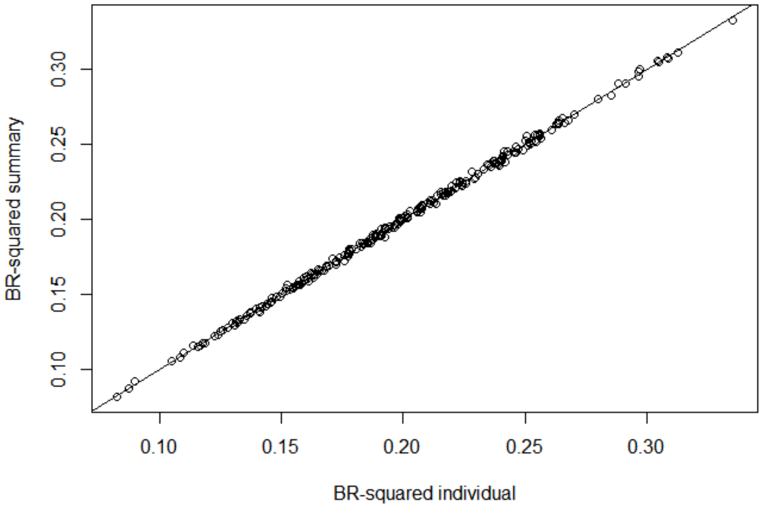


Figure 2.1: Comparison of IVW causal estimates between individual-level based BR-squared with summary data based BR-squared estimates in the case of rank-based selection (top 25 variants are selected). True causal effect = 0.2.

Table 2.1: Mean, relative bias, mean of standard error (SE), standard deviation (SD), mean squared error (MSE), mean of interval lengths, coverage, and power by naive method, projack method (K=5), BR-squared, FIQT, Forde's method, conditional MLE (MLE), the mean of the normalized likelihood (Normalized), and Compromise method (Compromise). We repeated 250 simulations for threshould-based selection. True causal effect=0.2.

Method	**Mean**	**Relative bias**	**SE**	**SD**	**MSE**	**CI length**	**Coverage**	**Power**
Threshold for p-value is $1 \cdot 10^{-4}$								
Naive	0.177	-0.116	0.038	0.039	0.0021	0.147	0.904	0.988
Projack $k = 5$	0.195	-0.023	0.043	0.044	0.0020	0.163	0.940	0.988
BR-squared	0.196	-0.022	0.043	0.044	0.0020	0.162	0.940	0.988
FIQT	0.192	-0.041	0.042	0.043	0.0019	0.160	0.940	0.988
Forde	0.194	-0.031	0.042	0.043	0.0019	0.161	0.936	0.988
MLE	0.189	-0.061	0.043	0.046	0.0022	0.167	0.940	0.976
Normalized	0.197	-0.017	0.043	0.048	0.0023	0.167	0.940	0.984
Compromise	0.193	-0.033	0.043	0.048	0.0023	0.166	0.948	0.980
Threshold for p-value is $5 \cdot 10^{-4}$								
Naive	0.177	-0.115	0.035	0.038	0.0020	0.140	0.896	0.988
Projack	0.196	-0.019	0.041	0.043	0.0018	0.155	0.936	0.992
BR-squared	0.196	-0.020	0.041	0.043	0.0018	0.155	0.940	0.988
FIQT	0.193	-0.037	0.039	0.040	0.0018	0.153	0.932	0.988
Forde	0.194	-0.029	0.040	0.042	0.0018	0.153	0.940	0.988
MLE	0.188	-0.058	0.040	0.042	0.0019	0.155	0.924	0.988
Normalized	0.196	-0.022	0.041	0.043	0.0019	0.157	0.940	0.984
Compromise	0.193	-0.037	0.041	0.043	0.0019	0.156	0.940	0.988
Threshold for p-value is $1 \cdot 10^{-3}$								
Naive	0.176	-0.119	0.035	0.037	0.0019	0.137	0.888	0.992
Projack $k = 5$	0.195	-0.024	0.040	0.041	0.0017	0.153	0.944	0.996
BR-squared	0.195	-0.025	0.040	0.041	0.0017	0.152	0.944	0.988
FIQT	0.192	-0.041	0.039	0.040	0.0017	0.149	0.936	0.992
Forde	0.193	-0.034	0.039	0.041	0.0017	0.151	0.932	0.992
MLE	0.188	-0.061	0.039	0.041	0.0018	0.151	0.924	0.988
Normalized	0.193	-0.036	0.039	0.044	0.0020	0.152	0.944	0.988
Compromise	0.190	-0.048	0.039	0.043	0.0019	0.152	0.924	0.984

Table 2.2: Mean, relative bias, mean of standard error (SE), standard deviation (SD), mean squared error (MSE), mean of interval length, coverage, and power by naive method, projack method (K=5), BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of the normalized likelihood (Normalized), and Compromise method (Compromise). We repeated 250 simulations for rank-based selection. True causal effect=0.2.

Method	**Mean**	**Relative bias**	**SE**	**SD**	**MSE**	**CI length**	**Coverage**	**Power**
Top 20 variants								
Naive	0.176	-0.120	0.039	0.042	0.0023	0.152	0.884	0.984
Projack $k = 5$	0.195	-0.025	0.044	0.047	0.0022	0.169	0.920	0.984
BR-squared	0.195	-0.025	0.044	0.047	0.0022	0.168	0.932	0.984
FIQT	0.191	-0.045	0.042	0.046	0.0021	0.164	0.936	0.984
Forde	0.193	-0.035	0.043	0.046	0.0022	0.166	0.928	0.984
MLE	0.191	-0.044	0.045	0.048	0.0024	0.173	0.928	0.960
Normalized	0.200	-0.000	0.045	0.050	0.0026	0.176	0.940	0.976
Compromise	0.196	-0.018	0.045	0.050	0.0025	0.175	0.936	0.972
Top 25 variants								
Naive	0.177	-0.116	0.037	0.039	0.0021	0.143	0.904	0.984
Projack $k = 5$	0.196	-0.019	0.041	0.044	0.0019	0.160	0.940	0.984
BR-squared	0.196	-0.020	0.041	0.044	0.0019	0.159	0.944	0.984
FIQT	0.192	-0.039	0.040	0.043	0.0019	0.156	0.940	0.984
Forde	0.194	-0.030	0.040	0.043	0.0019	0.157	0.932	0.984
MLE	0.189	-0.052	0.042	0.045	0.0021	0.162	0.928	0.988
Normalized	0.197	-0.015	0.042	0.047	0.0022	0.163	0.932	0.984
Compromise	0.194	-0.030	0.042	0.047	0.0022	0.163	0.936	0.988
Top 30 variants								
Naive	0.176	-0.120	0.035	0.037	0.0020	0.138	0.880	0.992
Projack $k = 5$	0.195	-0.023	0.040	0.042	0.0018	0.154	0.932	0.996
BR-squared	0.195	-0.024	0.040	0.042	0.0018	0.153	0.932	0.988
FIQT	0.192	-0.041	0.039	0.041	0.0017	0.150	0.92	0.992
Forde	0.193	-0.034	0.039	0.041	0.0017	0.151	0.936	0.988
MLE	0.188	-0.059	0.039	0.042	0.0019	0.153	0.924	0.984
Normalized	0.195	-0.024	0.040	0.043	0.0019	0.155	0.936	0.984
Compromise	0.193	-0.037	0.040	0.043	0.0019	0.154	0.936	0.984

Table 2.3: Mean, relative bias, mean of standard error (SE), standard deviation (SD), mean squared error (MSE), mean of interval lengths, coverage, and power by naive method, projack method (K=5), BR-squared, FIQT, Forde's method, conditional MLE (MLE), the mean of the normalized likelihood (Normalized), and Compromise method (Compromise). We repeated 250 simulations for threshould-based selection. True causal effect=0.05.

Method	**Mean**	**Relative bias**	**SE**	**SD**	**MSE**	**CI length**	**Coverage**	**Power**
Threshold for p-value is $1 \cdot 10^{-4}$								
Naive	0.042	-0.151	0.039	0.041	0.0017	0.152	0.928	0.220
Projack $k = 5$	0.047	-0.061	0.044	0.045	0.0020	0.169	0.928	0.208
BR-squared	0.047	-0.061	0.044	0.045	0.0020	0.167	0.928	0.216
FIQT	0.046	-0.078	0.043	0.043	0.0020	0.164	0.932	0.216
Forde	0.047	-0.069	0.043	0.043	0.0020	0.166	0.928	0.212
MLE	0.046	-0.081	0.044	0.047	0.0020	0.169	0.944	0.240
Normalized	0.047	-0.051	0.045	0.047	0.0020	0.172	0.924	0.224
Compromise	0.047	-0.067	0.045	0.047	0.0020	0.171	0.928	0.232
Threshold for p-value is $5 \cdot 10^{-4}$								
Naive	0.043	-0.146	0.037	0.039	0.0016	0.144	0.920	0.232
Projack	0.047	-0.052	0.042	0.044	0.0019	0.160	0.932	0.236
BR-squared	0.047	-0.056	0.042	0.044	0.0019	0.160	0.928	0.236
FIQT	0.046	-0.071	0.041	0.044	0.0018	0.157	0.928	0.228
Forde	0.047	- 0.063	0.041	0.043	0.0019	0.158	0.924	0.220
MLE	0.045	-0.092	0.041	0.043	0.0019	0.159	0.928	0.228
Normalized	0.047	-0.054	0.042	0.044	0.0020	0.161	0.932	0.220
Compromise	0.047	-0.070	0.042	0.044	0.0019	0.160	0.932	0.220
Threshold for p-value is $1 \cdot 10^{-3}$								
Naive	0.042	-0.161	0.037	0.038	0.0015	0.141	0.932	0.224
Projack $k = 5$	0.046	-0.069	0.041	0.042	0.0018	0.154	0.936	0.224
BR-squared	0.046	-0.073	0.041	0.042	0.0018	0.156	0.944	0.220
FIQT	0.046	-0.086	0.040	0.042	0.0018	0.154	0.940	0.220
Forde	0.046	-0.080	0.040	0.042	0.0018	0.155	0.944	0.216
MLE	0.045	-0.099	0.040	0.042	0.0018	0.155	0.928	0.220
Normalized	0.046	-0.076	0.040	0.043	0.0018	0.157	0.928	0.208
Compromise	0.046	-0.088	0.040	0.043	0.0018	0.156	0.928	0.220

Table 2.4: Mean, relative bias, mean of standard error (SE), standard deviation (SD), mean squared error (MSE), mean of interval length, coverage, and power by naive method, projack method (K=5), BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of the normalized likelihood (Normalized), and Compromise method (Compromise). We repeated 250 simulations for rank-based selection. True causal effect=0.05.

Method	**Mean**	**Relative bias**	**SE**	**SD**	**MSE**	**CI length**	**Coverage**	**Power**
Top 20 variants								
Naive	0.042	-0.167	0.040	0.043	0.0019	0.156	0.932	0.212
Projack $k = 5$	0.046	-0.076	0.045	0.048	0.0023	0.174	0.920	0.204
BR-squared	0.046	-0.077	0.045	0.048	0.0023	0.173	0.920	0.220
FIQT	0.045	-0.094	0.044	0.047	0.0022	0.169	0.936	0.212
Forde	0.046	-0.087	0.044	0.047	0.0022	0.171	0.932	0.212
MLE	0.046	-0.078	0.046	0.048	0.0024	0.178	0.936	0.188
Normalized	0.048	-0.040	0.047	0.050	0.0025	0.182	0.928	0.196
Compromise	0.047	-0.056	0.046	0.049	0.0024	0.181	0.924	0.192
Top 25 variants								
Naive	0.042	-0.151	0.038	0.040	0.0017	0.148	0.920	0.232
Projack $k = 5$	0.047	-0.055	0.043	0.045	0.0020	0.165	0.928	0.216
BR-squared	0.047	-0.060	0.042	0.045	0.0020	0.163	0.928	0.228
FIQT	0.046	-0.076	0.041	0.044	0.0019	0.160	0.928	0.232
Forde	0.047	-0.068	0.042	0.044	0.0020	0.162	0.924	0.224
MLE	0.045	-0.091	0.043	0.045	0.0020	0.167	0.928	0.208
Normalized	0.048	-0.050	0.043	0.046	0.0021	0.168	0.928	0.204
Compromise	0.047	-0.069	0.043	0.046	0.0021	0.168	0.932	0.208
Top 30 variants								
Naive	0.042	-0.170	0.036	0.038	0.0015	0.142	0.920	0.204
Projack $k = 5$	0.046	-0.074	0.041	0.043	0.0018	0.158	0.924	0.216
BR-squared	0.046	-0.080	0.041	0.043	0.0018	0.157	0.924	0.216
FIQT	0.045	-0.094	0.039	0.042	0.0018	0.154	0.924	0.204
Forde	0.046	-0.088	0.039	0.042	0.0018	0.155	0.928	0.208
MLE	0.045	-0.097	0.039	0.043	0.0019	0.158	0.932	0.240
Normalized	0.047	-0.067	0.041	0.043	0.0019	0.159	0.924	0.228
Compromise	0.046	-0.078	0.041	0.043	0.0019	0.159	0.924	0.232

In each scenario, the naive estimates exhibit a bias towards the null. For instance, when the causal effect is 0.2, the relative bias for the naive method is approximately 12%. However, this bias is notably reduced when employing correction methods. The BR-squared and Projack methods lead to a substantial reduction in relative bias, bringing it down to 2%, while the MLE method results in a 6% reduction, which is relatively less effective compared to the other correction methods. The BR-squared method consistently outperforms other approaches in terms of reducing relative bias. Nevertheless, the correction process does have its trade-offs, as it tends to increase variability. Ghosh's methods, in particular, exhibit higher variability compared to other correction techniques. We also note that the standard errors have a tendency to underestimate the standard deviation across all methods. This tendency is particularly pronounced in Ghosh's three methods.

In the comparison of various correction methods, it is observed that Projack, FIQT, BR-squared, and Forde's method exhibit smaller Mean Squared Error (MSE) values than those of Ghosh's three methods. While the naive method has a smaller MSE compared to Ghosh's method, it is important to note that this does not imply that the naive method is superior to Ghosh's method. The naive method has a much greater bias but a smaller variance than Ghosh's methods. This means that estimates from the naive method are closely clustered around an incorrect estimate of the causal effect (Tables 2.1-2.2).

Although the intervals may widen due to increased variability, the coverage is enhanced. Specifically, while the naive method may yield coverage under 90% in some cases, correction methods improve the coverage to around 94%.

It is worth noting that the correction has minimal impact on the power of the analysis. When employing a more stringent threshold, the variability increases due to a decreased number of selected variants. This trend is also observed in rank-based selection; a larger number of top-ranked variants leads to reduced variability and shorter intervals.

When dealing with a very small causal effect size (0.05), the pattern differs from that observed with a moderate causal effect. The relative bias is more pronounced compared to the case of a moderate causal effect. For instance, the BR-squared method exhibits an increase in relative bias from 2% to approximately 6%, while the MLE method shows an increase from 6% to 9%. However, the use of correction methods leads to a reduction of over fifty percent of the bias present in the naive estimator. The coverage sees a slight improvement, although not as noticeable as when the causal effect is 0.2, with values ranging from 92% to 93%.

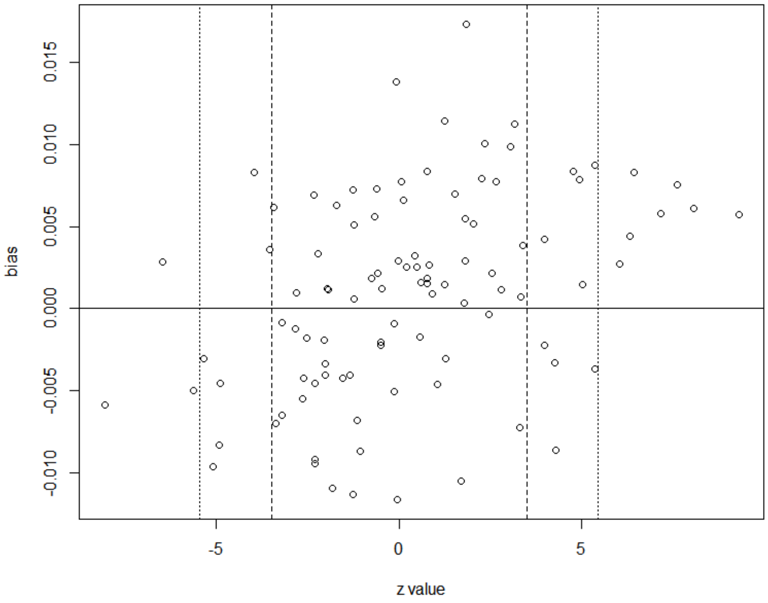


Figure 2.2: Plot of z-statistics against the bias (naive estimate - true effect) from one simulation. The dash lines correspond to the Bonferroni-corrected threshold ($5 \cdot 10^{-4}$), and the dotted lines represent the GWAS significance threshold ($5 \cdot 10^{-8}$).

Figure 2.2 illustrates the relationship between the z-statistic and bias of naive estimate. The bias becomes smaller as the the z-statistic is more significant. It is evident from the graph that variants with negative z-statistics that are significant typically have negative bias, while significant variants with positive z-statistics have positive bias.

In order to gain a deeper understanding of performance of each method in reducing bias for each naive association estimate, we explore the relationship between the proportion of corrected G-X estimate to the naive estimate and the absolute value of corresponding z-statistic for each IV. It has been observed that Ghosh's method tends to over-correct the estimate, especially when the associated z-statistic is close to the significant threshold. This tendency is evident in Figure 2.3, where Ghosh's three estimates are significantly reduced, amounting to only 20% of the original estimates when they are in proximity to the boundary. Such substantial reduction near the boundary could introduce bias in the causal estimate away from 0, particularly if a large proportion of variants have z-statistics that are close to this boundary.

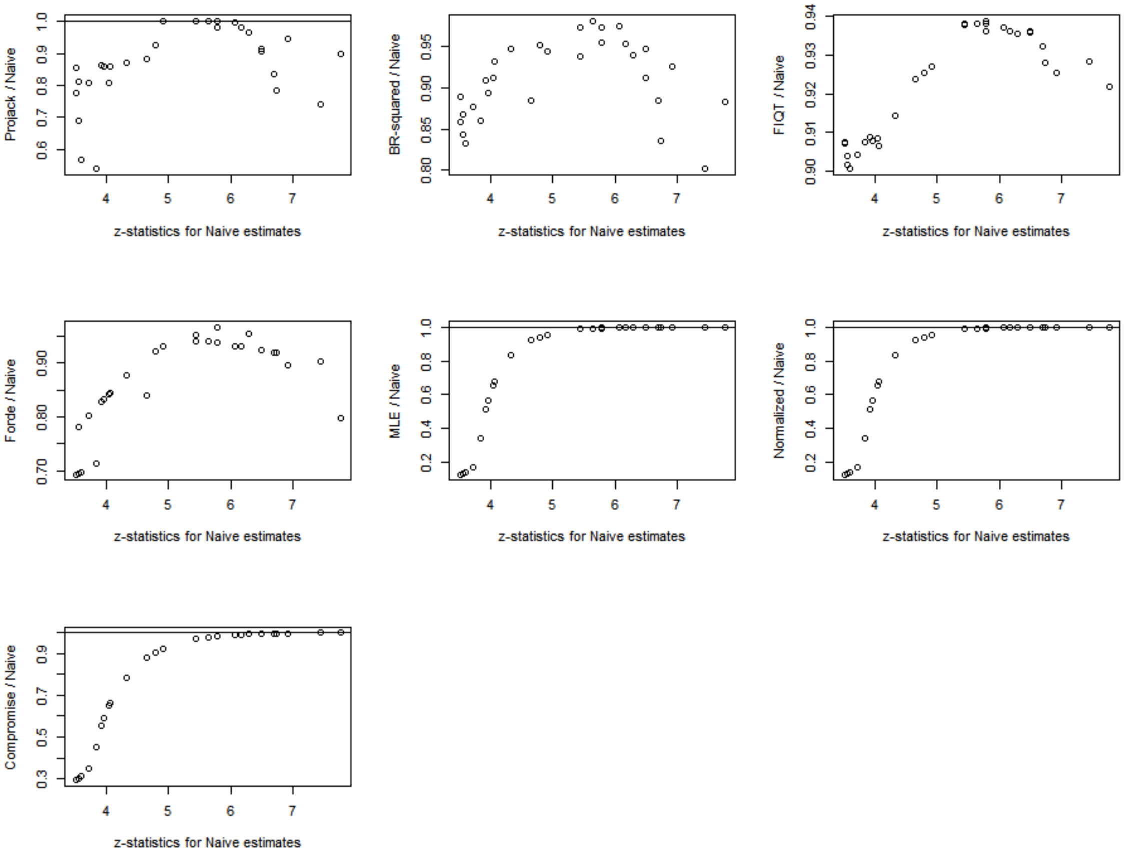


Figure 2.3: Plot the absolute z-statistic for the naive estimates against the ratio between the corrected estimate and naive estimate in one simulation. Scenario: threshold-based selection ($5 \cdot 10^{-4}$); true causal effect is 0.2.

There is minimal variation observed when employing different thresholds or varying the number of selected variants in the selection process. To explore this phenomenon further, we conduct additional simulations with more pronounced differences in either the threshold or the number of top variants selected, as detailed in Appendix A. The outcomes indicate a slight increase in statistical power with a higher number of variants or a less stringent threshold (Tables A.1-A.4). The reason is that the power is associated with the proportion of variance in the exposure explained by the genetic variants (Brion *et al.*, 2013). A larger number of variants results in a higher R^2 from the first-stage regression. Estimates from the naive method, projack, FIQT, BR-squared, and Forde's method in Tables A.1-A.4 exhibit negligible differences compared to the estimates in Tables 2.1 and 2.2. However, a more pronounced difference is observed in Ghosh's estimates between those obtained with a significance threshold of 0.1 (Table A.1) and those with smaller significance thresholds, as shown in Table 2.1. This discrepancy arises because a larger significance level threshold leads to a larger proportion of variants with z-statistics significantly deviating from the boundary, resulting in more z-statistics with minimal or no correction.

2.5 Applied Examples

To assess how winner's curse affects Mendelian randomization analysis and the effectiveness of the correction method, we examined two examples as discussed in Rees *et al.* (2019). We started by selecting instruments based on their genetic association with the exposure variable, and using the same data to estimate the effect size. We obtained estimates for the genetic association with the outcome variable from an independent GWAS dataset. Next, we adjusted the estimates for exposure association and used the IVW method to evaluate the presence of a causal effect and estimate its magnitude.

2.5.1 Causal Effect of Body Mass Index on Schizophrenia Risk

Individuals with schizophrenia tend to have a higher prevalence of obesity, but it is commonly believed that this association is due to the impact of antipsychotic medication on body composition (reverse causation) and genetic factors, rather than any causal effect of BMI on the risk of developing schizophrenia (Annamalai *et al.*, 2017).

The G-X association was obtained from the Genetic Investigation of Anthropometric Traits (GIANT) consortium, involving 339,225 individuals and 2,555,511 variants (Locke *et al.*, 2015). A set of independent variants was obtained by clumping the GIANT variants with a correlation threshold of $r^2 > 0.1$ and a minimum separation of 500 kilobases, leaving 122,802 SNPs. Within this set, 97 variants with p-values less than 5×10^{-8} were identified. The winner's curse correction was applied for these 97 variants. During the correction process, we used summary statistics for all 122,802 SNPs, except for Ghosh's method, which relies solely on the summary statistics for significant SNPs. The G-Y association was obtained from the Psychiatric Genomics Consortium (PGC), encompassing 35,476 cases and 46,839 controls (Pantelis *et al.*, 2014). Subsequently, the causal effect in MR was estimated using only the significant 97 variants.

The results are depicted in Figure 2.4 and Table 2.5. Although none of the estimates are considered statistically significant at a 5% significance level (in line with findings in Hartwig *et al.* (2016)), the corrected estimates are consistently larger than the naive estimate, especially the Ghosh's estimators (Table 2.5). This can be attributed to a relatively high proportion of naive estimates being significantly over-corrected, where approximately 36% of variants exhibit a corrected estimate that is less than 70% of the naive estimate. Ghosh's methods (MLE, compromise, normalized) demonstrate a tendency to over-correct the estimates when they are close to the significance threshold, as indicated in Figure 2.4. Additionally, the projack method exhibits more variability in the correction compared to BR-squared, FIQT, and Forde's methods.

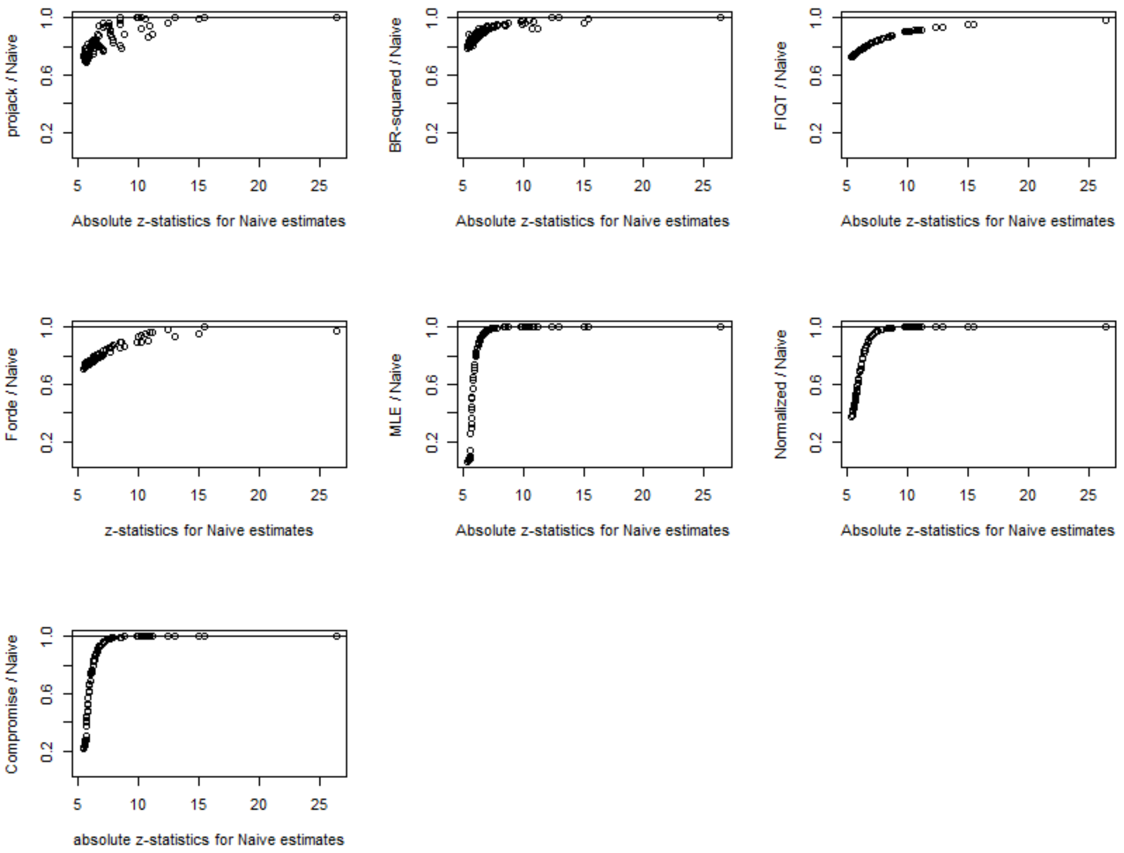


Figure 2.4: Plot the absolute z-statistic for the naive estimates against the ratio between the corrected estimate of exposure association and naive estimate of exposure association for the example of the impact of body mass index on schizophrenia. Abbreviations: Forde: Forde's bootstrap method; MLE: conditional MLE; Normalized: the mean of normalized likelihood estimator; Compromise: compromise estimator

Table 2.5: Estimates, standard errors and 95% confidence intervals of the causal effect of body mass index on schizophrenia risk (log odds ratio for schizophrenia per 1 standard deviation increase in body mass index) for naive estimator, projack estimator(k=5), BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of the normalized likelihood (Normalized), and Compromise estimator (Compromise).

Method	**Estimate**	**SE**	**95% Interval**
Naive	0.027	0.044	-0.060, 0.114
Projack $k = 5$	0.045	0.050	-0.053, 0.143
BR-squared	0.041	0.048	-0.053, 0.134
FIQT	0.043	0.052	-0.059, 0.144
Forde	0.043	0.050	-0.055, 0.140
MLE	0.071	0.050	-0.027, 0.169
Normalized	0.072	0.050	-0.027, 0.170
Compromise	0.071	0.050	-0.027, 0.170

2.5.2 Causal Effect of Low-density Lipoprotein Cholesterol on Alzheimer's Disease Risk

Numerous investigations have delved into the potential applications of lipid-lowering agents, particularly statins, to combat or prevent Alzheimer's disease, driven by the belief that elevated cholesterol levels might heighten the risk of developing the condition. Both epidemiological studies and preclinical research have suggested an inverse association between high cholesterol and Alzheimer's disease. However, human studies examining the effects of statins have yielded inconsistent outcomes, making it challenging to draw definitive conclusions (Shepardson *et al.*, 2011a,b).

The genetic association estimates with LDL-C were obtained from the Global Lipid Genetics Consortium (GLGC) with a sample size of up to 173,082 and 2,427,752 SNPs (Willer *et al.*, 2013). Variants were clumped with $r^2 < 0.1$, and they were separated by at least 1000 kb. After clumping, 122,802 SNPs remained, and among them, 180 SNPs were determined to be significant at the GWAS threshold (5×10^{-8}). For the genetic variant association on Alzheimer's disease (AD), the International Genomics of Alzheimer's Project (IGAP) data were used, involving 17,008 cases and 37,154 controls of European descent (Lambert *et al.*, 2013).

All causal estimates suggest a positive causal effect of LDL-C on the risk of AD, with FIQT showing the largest effect (Table 2.6). Once again, Ghosh's method performed the least favorably. Figure 2.5 illustrates that the majority of naive estimates were essentially left uncorrected by Ghosh's methods, while a few were over-corrected, resulting in the corrected causal estimate being similar to the unadjusted causal estimate. Specifically, there is around 56% of variants having their corrected estimates greater than 99% of the naive estimate while only 11% of variants having their corrected estimates less than 70% of the naive estimate. When compared to the FIQT method, Forde's method tended to produce smaller estimates for variants with a particularly large effect size. Additionally, Forde's method had a greater number of uncorrected variants compared to the BR-squared method.

It is also worth noting that we see a difference in the performance of Ghosh's MLE method in two applied examples. The reason is that there is a relatively great portion of variants with their z-statistics close to the boundary. Specifically, in the BMI- schizophrenia example, approximately 38% of variants have absolute z-statistics less than 6, wheras in this example, this percentage is only around 13%. Hence, in the first example, a greater proportion of variants undergo over-correction in their G-X associations, whereas in the second example, a higher proportion of variants experience only minimal correction in their G-X associations.

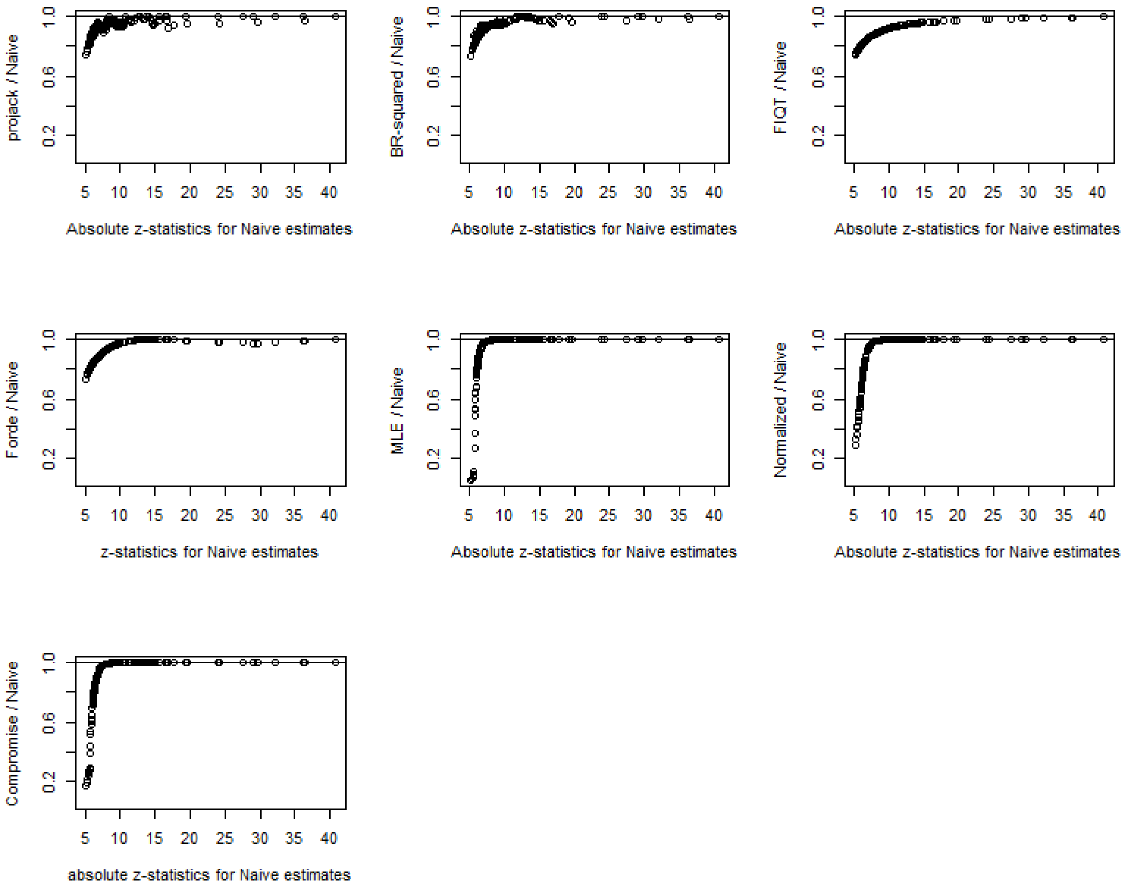


Figure 2.5: Plot the z-statistic for the naive estimates against the ratio between the corrected estimate of exposure association and naive estimate of exposure association for the example of the impact of low-density lipoprotein cholesterol on Alzheimer's disease risk. Abbreviations: Forde: Forde's bootstrap method; MLE: conditional MLE; Normalized: the mean of normalized likelihood estimator; Compromise: compromise estimator

Table 2.6: Estimates, standard errors and 95% confidence intervals of the causal effect of low-density lipoprotein cholesterol on Alzheimer's disease risk (log odds ratio for Alzheimer's per 1 standard deviationincrease in low-density lipoprotein cholesterol) from the IVW method with for naive estimator, projack estimator(k=5), BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of the normalized likelihood (Normalized), and compromise estimator (Compromise).

Method	**Estimate**	**SE**	**95% Interval**
Naive	0.586	0.028	0.531, 0.641
Projack $k = 5$	0.616	0.029	0.560, 0.674
BR-squared	0.611	0.029	0.554, 0.668
FIQT	0.644	0.030	0.589, 0.703
Forde	0.611	0.029	0.554, 0.668
MLE	0.599	0.029	0.543, 0.656
Normalized	0.612	0.029	0.556, 0.670
Compromise	0.606	0.029	0.549, 0.662

2.6 Summary

This chapter investigated the influence of the winner's curse on MR analysis and implements a correction method to mitigate it. The impact of the winner's curse was illustrated through simulations and two practical examples, and the correction method was explained in detail with its application in MR.

We find that when the causal effect size reduces, the magnitude of bias decreases, but the relative bias increases. Nonetheless, the winner's curse does not affect the overall conclusions, as evidenced by both the simulation and real data application. For instance, in the example of BMI's impact on schizophrenia risk, neither the naive nor the corrected causal estimate is significant. Furthermore, in the simulation with a small causal effect, the power is only minimally influenced by the winner's curse.

The winner's curse will bias the estimator of causal effect towards null. One way to mitigate this bias is by applying a correction. However, this correction comes at a cost of increased variability, resulting in a bias-variance trade-off. Consequently, the intervals are wider due to the increased variability, but coverage improves. Winner's curse can also be mitigated by using additional GWAS data which is separate from discovery GWAS. However, in reality, it can be challenging to find two distinct datasets.

In comparing various correction methods, it has been observed that Ghosh's methods yield more variable outcomes than other methods. The reason for this is that the estimate is excessively corrected when the associated z-statistic is near the boundary, whereas minimal correction is applied to variants with z-statistics far from the boundary. If a considerable proportion of significant variants are subject to heavy over-correction, the causal effect is likely to be overestimated. Conversely, the estimate will be similar to the naive estimate. For instance, in the case of the relationship between BMI and schizophrenia risk, Ghosh's estimates are three times as large as the naive estimate, whereas in the case of the association between LDL-C and Alzheimer's disease risk, the corrected estimate is similar to the initial estimate.

The type-I error is not examined in the main simulation. This is because we expect the winner's curse will not influence the type-I error when the instrument variable is valid. If the instrument variable meets the criteria for being a valid IV, then the genetic association with outcome (β_{gy}) will be zero. We also try to rectify this gap by re-performing the simulations with a null causal effect. The type-I errors consistently remain at 5% across all methods. Further details can be found in Table A.5 of Appendix A.

If researchers aim to obtain a point estimate, they can use a correction method to achieve a more accurate estimate. For those interested in obtaining confidence intervals, applying a correction method can lead to improved coverage but wider intervals. However, if the goal is to investigate the presence of a causal effect, the correction methods have minimal impact on the power to detect a causal effect. The R code for our extension of the BR-squared method to require only summary statistics is available in `https://github.com/Bianmj/WinnersCurse`.

Chapter 3

Addressing the Bias Due to Overlapping Samples in Mendelian Randomization

In the previous chapter, we explored the issue of the winner's curse, which arises when the dataset used for both selection and estimation exhibits full overlap. Specifically, our focus in this chapter is on the bias resulting from the overlap between two estimation datasets, namely the G-X association and G-Y association. In such cases, utilizing the traditional "first-order" weight to account for variability in the IVW estimate is no longer appropriate. This because that the traditional approach neglects the covariance between the G-X association and G-Y association, induced by correlation.

Burgess *et al.* (2016) has indicated the combination of the overlapping samples and weak instruments leads to substantial bias in causal estimation. Employing weak instruments biases the IVW estimator towards zero in the case of non-overlapping samples and towards the confounded (observational) association with one-sample overlapping. Consequently, the bias in cases with some degree of overlap lies between these two scenarios.

To address these challenges, we have developed a novel method building on the work of Bowden *et al.* (2019). Our approach has undergone extensive testing through various simulation scenarios, incorporating different strengths of IVs varying degrees of overlap, the effects of confounders, and different sample sizes for the exposure and outcome variables. Remarkably, our results consistently demonstrate the exceptional performance of the proposed method across all these scenarios, as it provides an unbiased causal estimator when compared with the IVW method and and while the Type-I error may be inflated, it is much better other methods when there is substantial overlap between samples.

3.1 Introduction

In recent years, the development of Mendelian randomization has been driven by the availability of summarized data. The two-sample MR method relies on the prerequisite that the exposure and outcome data from GWAS are acquired from independent samples. However, the presence of sample overlap in conjunction with weak instruments can introduce bias into the estimation of causal effects. The weak instrument bias arises when the weak association between the genetic variant and exposure leads to greater proportion of the difference in exposure being attributed to the chance difference in confounding factors, rather than being explained by the IVs themselves (Burgess and Thompson, 2011). When the two-sample MR estimator is obtained from non-overlapping samples, the bias is in the direction of the null. However, if there is sample overlap, it will be biased towards the confounded observational association (Burgess *et al.*, 2016).

The IVW method is widely recognized as one of the most commonly used methods in two-sample MR analyses (Burgess *et al.*, 2013). However, it is important to note that the IVW method that uses the "first-order" weight for each ratio estimate, does not account for the uncertainty associated with the genetic variant-exposure effect. The absence of measurement error in the effect of genetic variant-exposure association is referred to as the "NO Measurement Error" (NOME) assumption (Bowden *et al.*, 2017). The use of weak instrumental variables violates the NOME assumption. Suppose $\beta_{g_ix} = \hat{\beta}_{g_ix} + \eta$, where η denotes measurement error and follows $N(0, \sigma^2_\eta)$. A substantial residual standard error $\hat{\sigma}_\eta$ occurs when the genetic variants are weakly correlated with the exposure due to the small variability in exposure explained by genetic variants. This circumstance leads to regression dilution bias in the regression of $\hat{\beta}_{g_iy}$ on $\hat{\beta}_{g_ix}$, thereby shifting the slope (IVW causal estimator) towards zero. To address the limitations of the "first-order" weights, "second-order" weights have been introduced to better account for the complete uncertainty in the ratio estimate of the causal effect. However, it is important to highlight that, as the instrument strength decreases, the second-order IVW estimates become increasingly susceptible to regression dilution bias, resulting in under coverage (Bowden *et al.*, 2019). Bowden *et al.* (2019) used modified weights to mitigate this dilution bias while Zhao *et al.* (2020) proposed a robust adjusted profile score approach (MR-RAPS) to address this bias. It is worth mentioning that the magnitude of the bias introduced by weak IVs and sample overlap is less pronounced when the strength of IVs becomes stronger.

Lin and Sullivan (2009) were the first to tackle the challenge of combining the summary statistics from multiple GWAS with overlapping samples. They adjusted the inverse variance estimator for genetic effect from multiple GWAS while accounting for the overlapping among these studies. Their work mainly focused on the case-control studies. LeBlanc *et al.* (2018) extended the work of Lin and Sullivan (2009) for scenarios that involve overlap for both two quantitative phenotypes and a combination of a quantitative trait and a case-control study. Their methodology is applied within a polygenic pleiotropy-informed framework, where the two traits share a common genetic basis, termed pleiotropy.

In the context of MR, the issue of sample overlapping has not been extensively investigated. Burgess *et al.* (2016) have shown that for a continuous outcome, bias due to sample overlap is linearly proportional to the overlap between the samples. Zou *et al.* (2020) proposed a Bayesian method to convert a scenario involving two overlapping samples into a one-sample MR setting. This approach requires individual-level data and uses iterative imputation of missing data conditioned on the observed data and estimated parameters through MCMC. More recently, Mounier and Kutalik (2023) introduced a method called MR-lap, which takes into account weak instrument bias and winner's curse, while also addressing the potential occurrence of sample overlap. In their method, they used the intercept of cross-trait LD score regression (LDSC) (Bulik-Sullivan *et al.*, 2015) to estimate the phenotypic correlation and sample size for overlap. As we have discussed in Chapter 2, the winner's curse arises when association estimates are obtained from the same dataset used for discovery, which results in the MR estimator of the causal effect being biased. Our focus is primarily on the challenges of weak instrument bias and sample overlap, without explicit consideration of the winner's curse. Our proposed method uses modified weights to account for weak instrument bias and sample overlap. The approach builds upon the works of LeBlanc *et al.* (2018) and Bowden *et al.* (2019) to tackle these challenges. It is worth noting that, in our method, the estimated correlation between G-X and G-Y association estimators is equal to the intercept of cross-trait LDSC. In cases where individual-level data is unavailable, we can rely on cross-trait LDSC to obtain this estimated correlation. Subsequently, this estimated correlation is included in the estimation of variance for the ratio estimator of each IV, resulting in the modification of the weight for each ratio estimate. This differs from the approach proposed by Mounier and Kutalik (2023), wherein they did not explicitly modify the weight for each ratio estimate due to sample overlap. Instead, they proposed a corrected IVW estimator where a term proportional to the intercept of cross-trait LDSC is substracted from the naive estimator.

In this chapter, we initially introduce our method that incorporates modified weights for each genetic variant. Subsequently, we use simulations and real data examples to effectively demonstrate the performance and practical applicability of our method. Notably, our simulations are specifically based on real data examples, ensuring a realistic evaluation of the effectiveness of our proposed approach.

3.2 Method

Let $\hat{\beta}_c$ denote the estimate of causal effect, $\hat{\beta}_{gx}$ as the estimate of the effect of genetic variants on the exposure, and $\hat{\beta}_{gy}$ as the estimate of the effect of genetic variants on the outcome, then the ratio estimate of the causal effect is:

$$\hat{\beta}_c = \frac{\hat{\beta}_{gy}}{\hat{\beta}_{gx}} \tag{3.2.1}$$

By applying the delta method, we can easily get the estimated variance of the causal estimate:

$$\widehat{\text{var}}(\hat{\beta}_c) = \frac{1}{\hat{\beta}_{gx}^2}\text{se}^2(\hat{\beta}_{gy}) + \frac{\hat{\beta}_{gy}^2}{\hat{\beta}_{gx}^4}\text{se}^2(\hat{\beta}_{gx}) - 2\frac{\hat{\beta}_{gy}}{\hat{\beta}_{gx}^3}\widehat{\text{cov}}(\hat{\beta}_{gx}, \hat{\beta}_{gy}) \tag{3.2.2}$$

In the case of overlapping samples, $\hat{\beta}_{gx}$ and $\hat{\beta}_{gy}$ are correlated since the G-X study and G-Y study share some individuals and their genetic variants. Thus, $\hat{\beta}_{gx}$ and $\hat{\beta}_{gy}$ are not independent, and it is necessary to consider the covariance between them, as expressed in Equation (3.2.2). If the correlation is ignored, then the estimated variance is written as

$$\widehat{\text{var}}(\hat{\beta}_c) = \frac{1}{\hat{\beta}_{gx}^2}\text{se}^2(\hat{\beta}_{gy}) + \frac{\hat{\beta}_{gy}^2}{\hat{\beta}_{gx}^4}\text{se}^2(\hat{\beta}_{gx}) \tag{3.2.3}$$

The covariance of two estimates with two case-control studies was derived by Lin and Sullivan (2009), and their work was further expanded upon by LeBlanc *et al.* (2018). In the latter study, the authors obtained the correlation estimate for various scenarios, including cases involving two quantitative phenotypes or a combination of a quantitative phenotype and a case-control study. The deviation in the estimates was based on maximum likelihood (ML) estimation, assuming that the genetic variants in both the G-X study and G-Y study follow a binomial distribution with parameters $(2, p)$. The correlation between $\hat{\beta}_{gx}$ and $\hat{\beta}_{gy}$ for two quantitative traits can be expressed using Equation (3.2.4), as presented in the work by LeBlanc *et al.* (2018):

$$\text{corr}(\hat{\beta}_{gx}, \hat{\beta}_{gy}) = \frac{n_c}{\sqrt{n_x \cdot n_y}} r_{xy} \tag{3.2.4}$$

Here, r_{xy} represents the phenotypic correlation between the exposure X and outcome Y, n_c denotes the number of shared individuals between G-X study and G-Y study, n_x represents the number of individuals in G-X study, and n_y represents the number of individuals in G-Y study. Note that the right-hand side of Equation (3.2.4) represents the intercept term in the cross-trait LD score regression. This implies that even in cases where the overlap percentage or the phenotype information is not available, performing cross-trait LD score regression allows us to obtain the correlation between G-X and G-Y association estimates. Equation (3.2.5) provides the covariance between the G-X and G-Y association estimates. Then the covariance between G-X and G-Y association estimates can be expressed as follows:

$$\widehat{\text{cov}}(\hat{\beta}_{gx}, \hat{\beta}_{gy}) = \text{corr}(\hat{\beta}_{gx}, \hat{\beta}_{gy}) \cdot \text{se}(\hat{\beta}_{gx}) \cdot \text{se}(\hat{\beta}_{gy}) \tag{3.2.5}$$

In the IVW method, the ratio estimates for each genetic variant are combined to obtain an overall causal estimate:

$$\hat{\beta}_{IVW} = \frac{\sum_{j=1}^{M} w_j \hat{\beta}_{cj}}{\sum_{j=1}^{M} w_j} \tag{3.2.6}$$

where w_j is the inverse variance weight for the jth ratio estimate. The variance of IVW estimate is then

$$\text{var}(\hat{\beta}_{IVW}) = \frac{1}{\sum_{j=1}^{M} w_j} \tag{3.2.7}$$

Note that the value of w_j will vary based on whether the correlation between G-X and G-Y association estimates is taken into account. Box 1 provides the expressions of the two common first-order and second-order weights, as well as the modified weight. Box 2 outlines the specific weight used in the calculation of the causal estimate for different methods.

Box 1

- First-order weight:

$$w_j = \left(\frac{1}{\hat{\beta}^2_{g_j x}} \mathrm{se}^2(\hat{\beta}_{g_j y}) \right)^{-1}$$

- Second-order weight:

$$w_j = \left(\frac{1}{\hat{\beta}^2_{g_j x}} \mathrm{se}^2(\hat{\beta}_{g_j y}) + \frac{\hat{\beta}^2_{g_j y}}{\hat{\beta}^4_{g_j x}} \mathrm{se}^2(\hat{\beta}_{g_j x}) \right)^{-1}$$

- Modified weight:

$$w_j = \left(\frac{1}{\hat{\beta}^2_{g_j x}} \mathrm{se}^2(\hat{\beta}_{g_j y}) + \frac{\hat{\beta}^2_{g_j y}}{\hat{\beta}^4_{g_j x}} \mathrm{se}^2(\hat{\beta}_{g_j x}) - 2 \frac{\hat{\beta}_{g_j y}}{\hat{\beta}^3_{g_j x}} \widehat{\mathrm{cov}}(\hat{\beta}_{g_j x}, \hat{\beta}_{g_j y}) \right)^{-1}$$

Box 2

- The conventional IVW method considers only the first-order weight, given by:
$$w_j = \left(\frac{1}{\hat{\beta}_{g_jx}^2} \mathrm{se}^2(\hat{\beta}_{g_jy}) \right)^{-1}$$
- The modified IVW method incorporates the covariance term:
$$w_j = \left(\frac{1}{\hat{\beta}_{g_jx}^2} \mathrm{se}^2(\hat{\beta}_{g_jy}) + \frac{\hat{\beta}_{g_jy}^2}{\hat{\beta}_{g_jx}^4} \mathrm{se}^2(\hat{\beta}_{g_jx}) - 2\frac{\hat{\beta}_{g_jy}}{\hat{\beta}_{g_jx}^3} \widehat{\mathrm{cov}}(\hat{\beta}_{g_jx}, \hat{\beta}_{g_jy}) \right)^{-1}$$
where j indexes the jth ratio estimate.

The modified IVW method can be adapted to address the issue of weak instruments by extending the "Exact" method proposed by Bowden *et al.* (2019) to incorporate the correlation between the G-X association and G-Y association estimates. To implement this extension, we follow the methodology outlined in Bowden *et al.* (2019). Specifically, we formulate two models: the first model represents the data-generating process for the G-Y associations under the assumption of no pleiotropy. This model is defined as a function of the causal effect and the true G-X association. The second model represents the one that is used for fitting the data. Importantly, it should be noted that if the estimates of G-X and G-Y associations are obtained from two independent datasets, the covariance term becomes 0, resulting in the same expression as presented in Bowden *et al.* (2019).

$$\text{Underlying model:} \quad \hat{\beta}_{g_jy} = \beta_c \beta_{g_jx} + \mathrm{se}(\hat{\beta}_{g_jy})\varepsilon_j, \quad \varepsilon_j \sim N(0,1) \tag{3.2.8}$$

$$\begin{aligned}\text{Fitted model:}\quad & \hat{\beta}_{g_jy} = \beta_c\hat{\beta}_{g_jx} + \sqrt{\beta_c^2\text{se}^2(\hat{\beta}_{g_jx}) + \text{se}^2(\hat{\beta}_{g_jy}) - 2\beta_c\widehat{\text{cov}}(\hat{\beta}_{g_jx}, \hat{\beta}_{g_jy})}\varepsilon'_j \\ & \varepsilon'_j \sim N(0,1) \end{aligned} \tag{3.2.9}$$

By dividing both sides of the fitted model by the G-X association estimate, we can derive a model for the estimate of the jth ratio estimate, given by:

$$\hat{\beta}_{cj} = \beta_c + \sqrt{\frac{\beta_c^2\text{se}^2(\hat{\beta}_{g_jx}) + \text{se}^2(\hat{\beta}_{g_jy}) - 2\beta_c\widehat{\text{cov}}(\hat{\beta}_{g_jx}, \hat{\beta}_{g_jy})}{\hat{\beta}_{g_jx}^2}}\varepsilon'_j \tag{3.2.10}$$

From this model, we can express the variance of each ratio estimate as

$$\text{var}(\hat{\beta}_{c_j}) = \frac{\beta_c^2\text{se}^2(\hat{\beta}_{g_jx}) + \text{se}^2(\hat{\beta}_{g_jy}) - 2\beta_c\widehat{\text{cov}}(\hat{\beta}_{g_jx}, \hat{\beta}_{g_jy})}{\hat{\beta}_{g_jx}^2} \tag{3.2.11}$$

Based on Equation (3.2.11), it can be observed that the variance of each ratio estimate depends on the true causal effect. We can represent the weight assigned to each ratio estimate as the reciprocal inverse variance, denoted as $w_j(\beta_c) = 1/\text{var}(\hat{\beta}_{c_j})$. Using this weight, we introduce the modified Cochran's Q statistic:

$$Q_m(w(\beta_c), \beta_c) = \sum w_j(\beta_c)(\hat{\beta}_{c_j} - \beta_c)^2 \tag{3.2.12}$$

The modified exact IVW estimate $\hat{\beta}_{\text{ME,IVW}}$ is obtained by directly minimizing the generalized Q statistic Q_m with respect to β_c. By replacing β_c by $\hat{\beta}_{\text{ME,IVW}}$ in Equation (3.2.11), we can get the variance for the modified exact IVW estimate. It can be expressed as:

$$\mathrm{var}(\hat{\beta}_{\mathrm{ME,IVW}}) = \frac{1}{\sum_{j=1}^{M} w_j(\hat{\beta}_{\mathrm{ME,IVW}})} \tag{3.2.13}$$

Box 3 defines the weight assigned to each ratio estimate j for the original "exact" method and the modified "exact" method.

Box 3

- The weight used to get the exact IVW estimate is given by

$$w_j = \left(\frac{\beta_c^2 \mathrm{se}^2(\hat{\beta}_{g_j x}) + \mathrm{se}^2(\hat{\beta}_{g_j y})}{\hat{\beta}_{g_j x}^2} \right)^{-1}$$

- The weight used to get the modified exact IVW estimate is given by

$$w_j = \left(\frac{\beta_c^2 \mathrm{se}^2(\hat{\beta}_{g_j x}) + \mathrm{se}^2(\hat{\beta}_{g_j y}) - 2\beta_c \widehat{\mathrm{cov}}(\hat{\beta}_{g_j x}, \hat{\beta}_{g_j y})}{\hat{\beta}_{g_j x}^2} \right)^{-1}$$

3.3 Data Description

We present two illustrative examples in our study. The first example pertains to the investigation of the causal relationship between body mass index (BMI) and systolic blood pressure (SBP), which was previously examined in Zhao *et al.* (2020). To select the IVs, we employ data from the GIANT consortium (Locke *et al.*, 2015). The clumping of variants is performed using the "extract_instrument" function in the *TwoSampleMR* software (Hemani *et al.*, 2018), with three distinct significant thresholds ($5 \cdot 10^{-2}$, $5 \cdot 10^{-5}$, $5 \cdot 10^{-8}$) applied during the clumping procedure, while keeping the remaining parameters at their default values. The BMI and SBP data utilized in this analysis are sourced from the UK Biobank (Neale lab: `https://www.nealelab.is/uk-biobank`). Following the clumping process, the number of variants remaining for each significant threshold is reported as 1114, 249, and 78, respectively. The phenotypic correlation is available in `https://ukbb-rg.hail.is/rg_browser/`. It is obtained through cross-trait LD score regression, and the LD scores are computed based on data from the 1000 Genomes project (1000 Genomes Project Consortium, 2012).

In the second example, we investigate the causal association between BMI and HDL-C, which has been previously established as a negative causal association in the study conducted by He *et al.* (2022). The selection of genetic variants is based on the UK Biobank dataset, whereas the G-X association estimates for the associated variants are obtained from the GIANT consortium, and the G-Y association estimates for HDL-C are obtained from the GLGC consortium (Willer *et al.*, 2013). As reported by Burgess et al. (Burgess *et al.*, 2016), a total of 55 common studies, comprising approximately 71% of the participants in the GLGC, are referenced in the publications authored by these two consortia. The sample size for the genetic variant-exposure association encompasses up to 339,224 individuals, while for the genetic variant-outcome association, it includes up to 187,167 individuals. In other words, the GIANT consortium has nearly twice as many samples as the GLGC consortium. Thus, if we define the overlap with respect to the GIANT consortium, then the overlapping is around 39%. Similarly, we apply the clumping procedure (it selects the most significant variant within a specified distance and remove all other nearby variants that have a high LD with the selected variants) using three thresholds, resulting in 854, 464, and 222 variants remaining for the respective thresholds. The phenotypic correlation in this applied example is obtained from cross-trait LD score regression, with LD scores computed from 1000 Genomes (1000 Genomes Project Consortium, 2012).

The measure of IV strength is determined by calculating the mean of $\hat{\beta}_{gx}^2/\operatorname{se}^2(\hat{\beta}_{gx})$ across multiple variants, which follows an F-distribution. Table 3.1 presents the specific values of F-statistics for the two examples. Notably, as the threshold becomes more stringent, the average F-statistics exhibit an upward trend, indicating an increase in IV strength.

Table 3.1: The mean of F-statistics for varing thresholds in two examples.

Thresholds	The mean of F-statistics	
	BMI-SBP	BMI-HDL-C
$5 \cdot 10^{-2}$	13.69	7.56
$5 \cdot 10^{-5}$	39.72	12.50
$5 \cdot 10^{-8}$	77.56	20.72

3.4 Simulation

3.4.1 Design

To compare the performance of the four methods in the case of weak instrument bias more realisticly, we perform the following simulations based on the above two real data examples. The data-generating model is given below.

$$\begin{aligned}
&X_i = \sum_{k=1}^{M} \alpha_k G_{ik} + \alpha_u U_i + \varepsilon_{x_i} \\
&Y_i = \beta_c X_i + \beta_u U_i + \varepsilon_{y_i} \\
&U_i \sim N\left(0, \sigma_u^2\right); \varepsilon_{x_i} \sim N\left(0, \sigma_x^2\right); \varepsilon_{y_i} \sim N\left(0, \sigma_y^2\right)
\end{aligned}$$

The parameters, including the G-X association effect and the MAF, as well as the causal effect, are determined based on real data examples. The process involves generating X and Y variables using a predetermined number of variants corresponding to different significance thresholds. The number of variants is intentionally reduced compared to the real data examples to ensure the comparability of the resulting R^2 values with a different sample size.

To be more specific, in the case of BMI-SBP, the number of variants is constrained to 100 for the significant threshold of $5 \cdot 10^{-2}$, 25 for the threshold of $5 \cdot 10^{-5}$, and 8 for the threshold of $5 \cdot 10^{-8}$. Similarly, in the example involving BMI-HDL-C, 50, 25, and 12 variants are considered for the respective significance thresholds.

The genotype G_{ik} is generated from $Bin(n, p_k)$, where $n = 2$ and p_k represents MAF sampled from the real data example. The true G-X associations α_k are obtained from the same variants that were used to select the MAF values from the real data during the genotype generation process. Additionally, the estimates obtained from the real data are adjusted by a scaling factor to ensure the F-statistics remain comparable to those in the real data example, despite using a smaller sample size in the simulation.

To closely simulate the BMI-SBP example, we set $\beta_c = 0.1$ and $\sigma_u^2 = \sigma_x^2 = \sigma_y^2 = 1$

in our simulation. Additionally, for each of the three significance thresholds, we consider different strengths of the confounder effect, specifically $\alpha_u = \beta_u = 0.4, 0.6, \text{or } 0.8$. Our simulated dataset comprises 20,000 observations, with the number of observations for G-X and G-Y associations being half of the sample size. To estimate the G-X association, we use the first half of the observations, while varying the second half to create overlapping segments. For example, if we desire a 25% overlap, we utilize observations from 7,501 to 17,500 for G-Y association estimation, and so on. We explore cases with 0% overlap, and gradually increase in increments of 25% up to 100% overlap. We repeat the simulation 1000 times to obtain the results.

To simulate the BMI-HDL-C example, we set $\beta_c = -0.3$ to facilitate comparison. We also assume a unit variance for the confounder effect and random errors. Similar to the previous scenario, we explore different strengths of the confounder effect, namely $\alpha_u = \beta_u = -0.4, -0.6, \text{or } -0.8$, for each of the three significance thresholds. It is worth noting that the direction of confounded (observational) association depends on the product of the signs of the confounder effect on both the exposure and the outcome. In this case, the confounded (observational) association is positive, given that the confounder has a negative effect on both the exposure and the outcome. Given that the GIANT consortium's sample size is nearly twice as large as that of GLGC, we conduct simulations with 30,000 observations. In these simulations, the sample size for estimating the G-X association is twice that of the G-Y association. The estimation of the G-X association is performed using the first 20,000 observations, while the remaining observations are varied to create overlapping segments. The degree of overlap is defined with respect to the smaller study, meaning that for a 25% degree of overlap, we use observations from 17,501 to 27,500 for the G-Y association estimation. Similarly, for 50% overlap, we use observations from 15,001 to 25,000 for the G-Y association estimation, and so forth. We perform 1000 repetitions of the simulation.

In our simulations, we use the invidual-level data to estimate phenotypic correlation while in our applied examples, their phenotypic correlations are obtained from cross-trait LD score regression and LD scores are computed from 1000 Genomes (1000 Genomes Project Consortium, 2012). We expect this estimate would be different from that obtained from the cross-trait LD score regression. Bulik-Sullivan *et al.* (2015) and Lee *et al.* (2018) have shown cross-trait LD score regression yields accurate estimates of genetic correlation but the performance of phenotypic correlation remains uncertain. Lee *et al.* (2018) point out that the intercept term is influenced by the confounding affecting both traits and sample overlap.

It is important to highlight that we have also taken into account scenarios where the confounding association is negative, indicating that the direction of the confounding effect on the exposure and outcome differs. Furthermore, we conducted a simulation with a null causal effect to evaluate the type-I error for various methods.

3.4.2 Results

The results of Simulation 1 are presented in Figure 3.1, Tables 3.2-3.3, Tables B.1-B.4 and Figures B.1-B.2. Similarly, the results of simulation 2 are provided in Figure 3.4, Tables 3.4-3.5, Tables B.5-B.8 and Figures B.3-B.4. To assess the strength of IVs, the mean F value is calculated for each threshold.

To evaluate the performance of each approach, we compute the mean of the causal estimates, mean of the standard errors (SE), standard deviation (SD) of the estimates, and coverage of 95% Wald intervals. Notably, in both simulation studies, we observe significant discrepancies in the causal estimates for the IVW, modified IVW, and exact methods at varying degrees of sample overlap. Conversely, the modified exact IVW estimates align more closely with the true causal effects (Figures 3.1 and 3.4). As anticipated from Box 3, the modified exact IVW method yields the same estimate as the exact IVW method when there is no overlap.

For a positive causal effect, the IVW and modified IVW methods tend to overestimate the causal effect with 100% overlap and underestimate it with no overlap (Figure 3.1). Conversely, with a negative causal effect, the IVW and modified IVW methods overestimate the causal effect, irrespective of the extent of sample overlap (Figure 3.4). This phenomenon arises due to the weak instrument bias, leading the estimates in two-sample MR towards the null and in one-sample MR towards the direction of the confounded observational association.

The behavior of the exact IVW method differs slightly due to its ability to mitigate weak instrument bias. However, an increase in bias is still observable as the degree of overlap increases. When the direction of confounder is positive, the exact method tends to overestimate the causal effect for a positive causal effect and underestimate it for a negative causal effect. Conversely, when the direction of confounder is negative, the exact method tends to underestimate the causal effect regardless of the direction of the causal effect.

The simulation results involving negative confounders are presented in Appendix B, specifically in Figures B.7-B.12. When considering a positive causal effect along with a negative confounder, all the IVW, modified IVW, and exact IVW methods exhibit underestimation. Moreover, the extent of bias increases with the degree of overlap, whereas the modified exact IVW method maintains stability around the true effect. On the other hand, when dealing with a negative causal effect and a negative confounder, the pattern mirrors that observed with a positive causal effect and a positive confounder. Here, the IVW, modified IVW, and exact methods result in underestimation with complete overlap and overestimation with no overlap.

However, it is important to note that bias reduction does come at a cost. The exact-based IVW methods tend to underestimate the standard deviation, particularly in cases of non-overlapping samples or when the degree of overlap is low (Table 3.3). This underestimation in standard errors leads to under coverage, with the lowest coverage observed at around 90% when weak instruments are present. As the strength of the instrumental variables increases, the coverage improves.

While there is a slight underestimation of coverage in cases involving non-overlapping samples or low degrees of overlap, it is worth noting that the modified exact method demonstrates enhanced coverage compared to the conventional IVW method in some scenarios. This improvement is particularly noticeable when the instrumental variables are weak and a high degree of overlap is present (Table 3.2). For example, when there is 100% overlap and the causal effect is 0.1, along with a confounding factor effect of 0.6, the IVW method achieves a coverage of 88.6% with a significant threshold of 0.05. In contrast, the modified exact method achieves a higher coverage of 93.6% under the same conditions.

As the mean of F-statistics increases, the difference between causal estimates using different methods diminishes, even with a high degree of sample overlap (see Table 3.2 and Figures B.1-B.4). However, it is worth noting that the extent of bias is still relatively high in the modified IVW method compared to other methods. It may seem counterintuitive since the second-order weights for the weak IVs tend to be smaller than the first-order weights. One would anticipate a less biased estimator with second-order weights, but the results indicate the opposite. For instance, with fully overlapping samples, the mean IVW estimate is 0.126 with a mean F-statistics around 10, which decreases to 0.105 with a mean F value around 55. Similarly, the mean of the exact IVW estimate drops from 0.139 to 0.107, and the modified IVW estimate decreases from 0.145 to 0.113. In contrast, the mean of the modified exact estimates remains stable at the true causal effect.

The bias in causal estimation for the IVW and modified IVW methods is influenced by the strength of the confounder. For instance, in the case of fully overlapping samples and a significant threshold of $5 \cdot 10^{-5}$, the IVW estimate increases from 0.103 with a weak confounder effect (0.4) to 0.115 with a relatively strong confounder effect (0.8). Figures 3.1 and 3.4 suggest the presence of an interaction effect between the confounding factor and the degree of overlap. In other words, the impact of overlap on the extent of bias in causal estimation depends on the strength of the confounding factor. For example, with a significant threshold of $5 \cdot 10^{-2}$, the IVW estimate increases by 23% from non-overlapping to complete overlapping with a weak confounder, 41.6% with a moderate confounder, and 64.8% with a strong confounder (Table 3.2, Table B.1 and Table B.3). It has been observed that for a positive causal effect and a positive confounder, a strong confounder increases the bias at a faster rate than a weak confounder, while for a negative causal effect and a positive confounder, a weak confounder tends to decrease the bias at a faster rate than a strong confounder. The observed interaction effects can be attributed to the strength of confounder, as a strong confounder leads to a greater weak instrument bias. In cases of strong confounding, noticeable overlap would be observed in the distributions of exposure across genotype groups. This occurs because the genetic variation explains minimal differences in exposure.

The exact IVW method does not provide a clear direction of bias. Across all scenarios, the modified exact IVW estimates consistently exhibit consistency for all degrees of overlap, different thresholds, and various confounder effects.

In the context of a null causal effect (Figure 3.7), the estimates exhibit unbiased behavior when there is no overlap, but as the degree of overlap increases, bias becomes more pronounced for the IVW, modified IVW, and exact IVW methods. However, the modified exact IVW method remains unbiased throughout. Regarding the type-I error, the IVW method maintains a 5% level when there is null overlap or the confounding effect is weak (Table 3.6). However, a significant increase in type-I error is observed when the confounder becomes stronger or the degree of overlap rises. For instance, when the confounder is weak, the IVW type-I error increases from 5% to 6.2% as the degree of overlap goes from 0% to 100%, and for a strong confounder, the type-I error surges from 5.3% to 25.7%. Conversely, the modified exact method exhibits a slight decrease in type-I error as the overlap increases. For a weak confounder, the type-I error reduces from 8.5% to 6.5%, and for a strong confounder, it decreases from 9.3% to 7.0%. Notably, the modified exact method consistently maintains the type-I error below 10%.

Through the pairwise plots (Figures 3.2-3.3 and Figures 3.5-3.6), we can illustrate that the modified exact method consistently outperforms other methods in reducing overlapping bias and weak instrument bias, although it comes at the expense of underestimating the standard error. However, this underestimation has a minimal impact on coverage of 95% Wald-type confidence intervals. The proposed method shows a substantial advantage over the conventional IVW method when the degree of overlap is high and the instruments are relatively weak.

Table 3.2: Mean of causal estimate and coverage for 1000 simulations. Scenario: positive causal effect ($\beta_c = 0.1$) and moderate confounder ($\beta_u = \alpha_u = 0.6$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (Mean F)	IVW		Modified IVW		Exact		Modified Exact	
	Mean	Coverage	Mean	Coverage	Mean	Coverage	Mean	Coverage
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 0%								
$5 \cdot 10^{-2} (F = 10.77)$	0.089	0.933	0.082	0.938	0.100	0.909	0.100	0.909
$5 \cdot 10^{-5} (F = 29.04)$	0.096	0.954	0.090	0.966	0.099	0.950	0.099	0.950
$5 \cdot 10^{-8} (F = 54.67)$	0.099	0.945	0.096	0.954	0.101	0.942	0.101	0.942
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 25%								
$5 \cdot 10^{-2} (F = 10.77)$	0.099	0.948	0.098	0.972	0.110	0.899	0.100	0.912
$5 \cdot 10^{-5} (F = 29.04)$	0.100	0.950	0.099	0.965	0.104	0.949	0.100	0.943
$5 \cdot 10^{-8} (F = 54.67)$	0.101	0.955	0.100	0.963	0.103	0.955	0.101	0.950
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 50%								
$5 \cdot 10^{-2} (F = 10.77)$	0.106	0.955	0.112	0.958	0.118	0.900	0.099	0.930
$5 \cdot 10^{-5} (F = 29.04)$	0.102	0.941	0.106	0.949	0.105	0.932	0.099	0.931
$5 \cdot 10^{-8} (F = 54.67)$	0.101	0.948	0.103	0.957	0.103	0.946	0.099	0.943
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 75%								
$5 \cdot 10^{-2} (F = 10.77)$	0.116	0.925	0.128	0.867	0.128	0.834	0.099	0.922
$5 \cdot 10^{-5} (F = 29.04)$	0.105	0.944	0.114	0.941	0.109	0.938	0.099	0.936
$5 \cdot 10^{-8} (F = 54.67)$	0.104	0.944	0.109	0.947	0.106	0.942	0.100	0.936
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 100%								
$5 \cdot 10^{-2} (F = 10.77)$	0.126	0.886	0.145	0.720	0.139	0.759	0.101	0.936
$5 \cdot 10^{-5} (F = 29.04)$	0.108	0.952	0.122	0.913	0.113	0.934	0.099	0.941
$5 \cdot 10^{-8} (F = 54.67)$	0.105	0.946	0.113	0.949	0.107	0.946	0.100	0.942

Table 3.3: Mean of standard errors (SE) and standard deviation (SD) for 1000 simulations. Scenario: positive causal effect ($\beta_c = 0.1$) and moderate confounder ($\beta_u = \alpha_u = 0.6$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (Mean F)	IVW		Modified IVW		Exact		Modified Exact	
	SE	SD	SE	SD	SE	SD	SE	SD
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 0%								
$5 \cdot 10^{-2} (F = 10.77)$	0.032	0.033	0.033	0.030	0.032	0.037	0.032	0.037
$5 \cdot 10^{-5} (F = 29.04)$	0.039	0.039	0.040	0.037	0.040	0.040	0.040	0.040
$5 \cdot 10^{-8} (F = 54.67)$	0.053	0.054	0.053	0.052	0.053	0.055	0.053	0.055
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 25%								
$5 \cdot 10^{-2} (F = 10.77)$	0.032	0.033	0.033	0.030	0.032	0.037	0.032	0.037
$5 \cdot 10^{-5} (F = 29.04)$	0.039	0.039	0.040	0.037	0.040	0.040	0.039	0.040
$5 \cdot 10^{-8} (F = 54.67)$	0.053	0.052	0.053	0.050	0.053	0.053	0.052	0.053
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 50%								
$5 \cdot 10^{-2} (F = 10.77)$	0.032	0.031	0.032	0.029	0.032	0.035	0.031	0.035
$5 \cdot 10^{-5} (F = 29.04)$	0.039	0.039	0.040	0.037	0.040	0.040	0.039	0.041
$5 \cdot 10^{-8} (F = 54.67)$	0.053	0.052	0.052	0.051	0.053	0.053	0.052	0.053
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 75%								
$5 \cdot 10^{-2} (F = 10.77)$	0.032	0.031	0.032	0.030	0.032	0.035	0.031	0.035
$5 \cdot 10^{-5} (F = 29.04)$	0.039	0.039	0.039	0.037	0.040	0.041	0.039	0.041
$5 \cdot 10^{-8} (F = 54.67)$	0.053	0.052	0.052	0.050	0.053	0.053	0.052	0.053
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.6$, overlap = 100%								
$5 \cdot 10^{-2} (F = 10.77)$	0.032	0.030	0.032	0.029	0.032	0.034	0.031	0.033
$5 \cdot 10^{-5} (F = 29.04)$	0.039	0.039	0.039	0.038	0.040	0.041	0.038	0.041
$5 \cdot 10^{-8} (F = 54.67)$	0.053	0.053	0.051	0.052	0.053	0.054	0.051	0.054

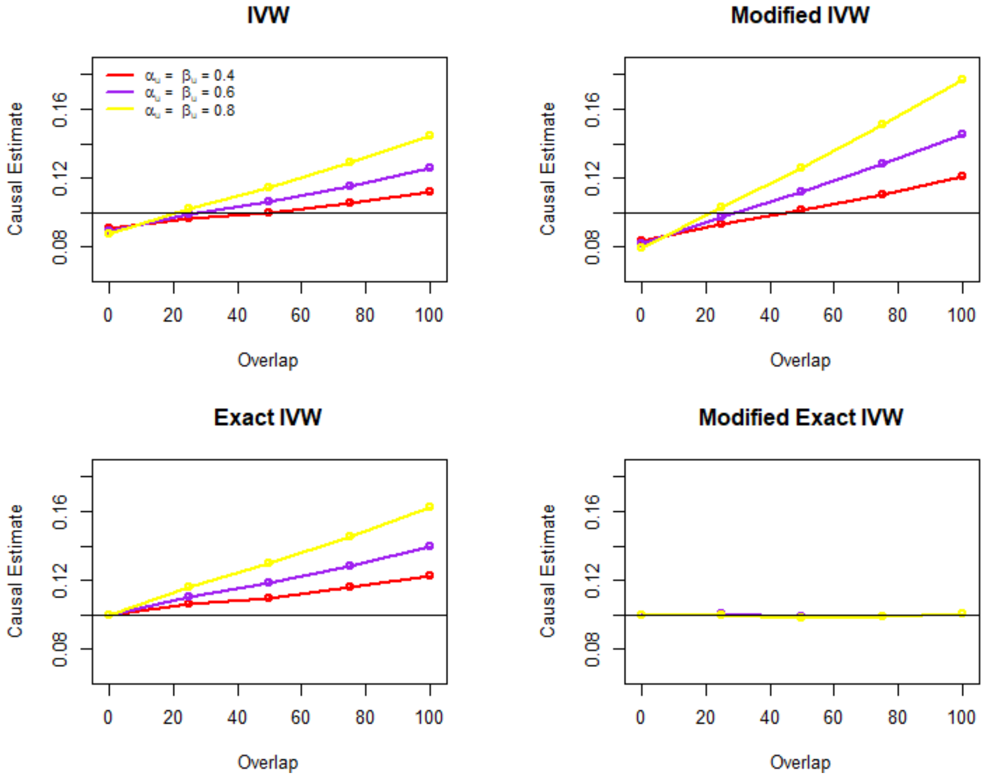


Figure 3.1: Causal estimates with different strengths of confounder. Scenario: Significant threshold $5 \cdot 10^{-2}$; Causal effect=0.1 (horizontal line). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

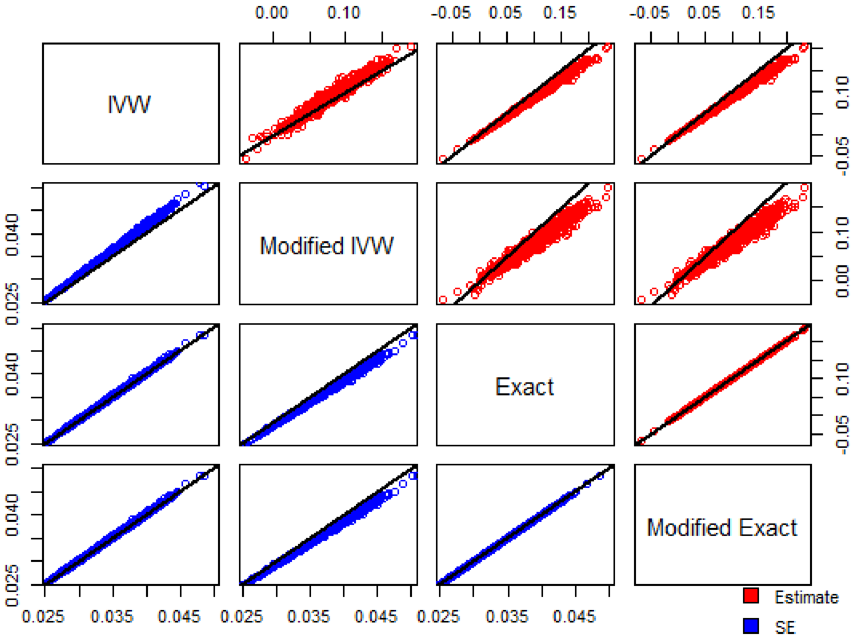


Figure 3.2: The pairwise comparison in standard error and causal estimate for different methods. Significant threshold $5 \cdot 10^{-2}$; causal effect=0.1; degree of overlap=0%; moderate confounder ($\beta_u = \alpha_u = 0.6$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

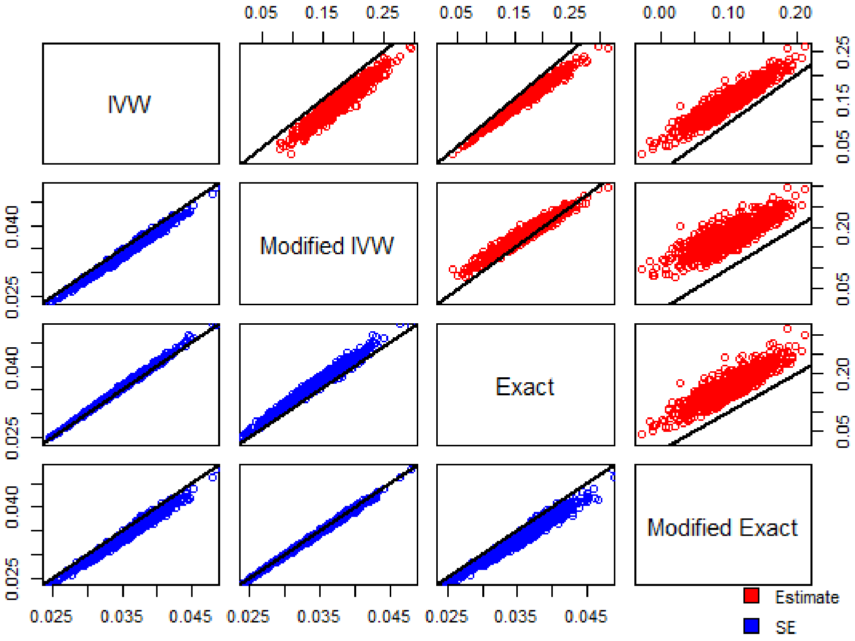


Figure 3.3: The pairwise comparison in standard error and causal estimate for different methods. Significant threshold $5 \cdot 10^{-2}$; causal effect=0.1; degree of overlap=100%; moderate confounder ($\beta_u = \alpha_u = 0.6$). Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Table 3.4: Mean of causal estimate and coverage of 95% Wald-type confidence interlvas for 1000 simulations. Scenario: positive causal effect ($\beta_c = -0.3$) and weak confounder ($\beta_u = \alpha_u = -0.4$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (Mean F)	IVW		Modified IVW		Exact		Modified Exact	
	Mean	Coverage	Mean	Coverage	Mean	Coverage	Mean	Coverage
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 0%								
$5 \cdot 10^{-2} (F = 10.07)$	-0.265	0.921	-0.240	0.889	-0.298	0.934	-0.298	0.934
$5 \cdot 10^{-5} (F = 15.61)$	-0.280	0.942	-0.257	0.943	-0.302	0.934	-0.302	0.934
$5 \cdot 10^{-8} (F = 24.58)$	-0.283	0.930	-0.265	0.931	-0.297	0.934	-0.297	0.934
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 25%								
$5 \cdot 10^{-2} (F = 10.07)$	-0.270	0.943	-0.249	0.924	-0.303	0.935	-0.298	0.932
$5 \cdot 10^{-5} (F = 15.61)$	-0.285	0.941	-0.265	0.947	-0.306	0.931	-0.304	0.933
$5 \cdot 10^{-8} (F = 24.58)$	-0.285	0.949	-0.269	0.954	-0.299	0.950	-0.297	0.948
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 50%								
$5 \cdot 10^{-2} (F = 10.07)$	-0.277	0.937	-0.259	0.937	-0.312	0.933	-0.302	0.927
$5 \cdot 10^{-5} (F = 15.61)$	-0.287	0.944	-0.270	0.949	-0.309	0.928	-0.303	0.926
$5 \cdot 10^{-8} (F = 24.58)$	-0.291	0.954	-0.278	0.960	-0.306	0.950	-0.302	0.948
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 75%								
$5 \cdot 10^{-2} (F = 10.07)$	-0.281	0.941	-0.267	0.943	-0.315	0.930	-0.301	0.925
$5 \cdot 10^{-5} (F = 15.61)$	-0.289	0.939	-0.276	0.956	-0.310	0.937	-0.302	0.928
$5 \cdot 10^{-8} (F = 24.58)$	-0.293	0.955	-0.282	0.969	-0.307	0.953	-0.302	0.947
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 100%								
$5 \cdot 10^{-2} (F = 10.07)$	-0.286	0.949	-0.275	0.959	-0.320	0.914	-0.301	0.915
$5 \cdot 10^{-5} (F = 15.61)$	-0.291	0.952	-0.281	0.966	-0.313	0.943	-0.301	0.938
$5 \cdot 10^{-8} (F = 24.58)$	-0.297	0.946	-0.288	0.959	-0.311	0.943	-0.303	0.935

Table 3.5: Mean of standard errors (SE) and standard deviation (SD) for 1000 simulations. Scenario: positive causal effect ($\beta_c = -0.3$) and weak confounder ($\beta_u = \alpha_u = -0.4$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (Mean F)	IVW		Modified IVW		Exact		Modified Exact	
	SE	SD	SE	SD	SE	SD	SE	SD
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 0%								
$5 \cdot 10^{-2} (F = 10.07)$	0.066	0.065	0.069	0.061	0.068	0.073	0.068	0.073
$5 \cdot 10^{-5} (F = 15.61)$	0.076	0.077	0.080	0.073	0.078	0.084	0.078	0.084
$5 \cdot 10^{-8} (F = 24.58)$	0.091	0.098	0.095	0.093	0.093	0.104	0.093	0.104
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 25%								
$5 \cdot 10^{-2} (F = 10.07)$	0.066	0.064	0.069	0.060	0.068	0.072	0.067	0.072
$5 \cdot 10^{-5} (F = 15.61)$	0.076	0.078	0.079	0.073	0.078	0.085	0.078	0.085
$5 \cdot 10^{-8} (F = 24.58)$	0.091	0.095	0.094	0.091	0.093	0.101	0.093	0.101
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 50%								
$5 \cdot 10^{-2} (F = 10.07)$	0.066	0.064	0.069	0.060	0.068	0.073	0.067	0.073
$5 \cdot 10^{-5} (F = 15.61)$	0.076	0.081	0.079	0.076	0.078	0.088	0.077	0.088
$5 \cdot 10^{-8} (F = 24.58)$	0.091	0.094	0.094	0.089	0.094	0.100	0.092	0.100
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 75%								
$5 \cdot 10^{-2} (F = 10.07)$	0.066	0.066	0.068	0.061	0.068	0.075	0.067	0.075
$5 \cdot 10^{-5} (F = 15.61)$	0.076	0.080	0.078	0.074	0.078	0.087	0.077	0.087
$5 \cdot 10^{-8} (F = 24.58)$	0.091	0.090	0.093	0.086	0.094	0.096	0.092	0.096
$\beta_c = -0.3$, $\beta_u = \alpha_u = -0.4$, overlap = 100%								
$5 \cdot 10^{-2} (F = 10.07)$	0.066	0.068	0.068	0.063	0.068	0.077	0.066	0.077
$5 \cdot 10^{-5} (F = 15.61)$	0.076	0.079	0.078	0.073	0.078	0.086	0.076	0.086
$5 \cdot 10^{-8} (F = 24.58)$	0.091	0.093	0.093	0.087	0.094	0.098	0.091	0.098

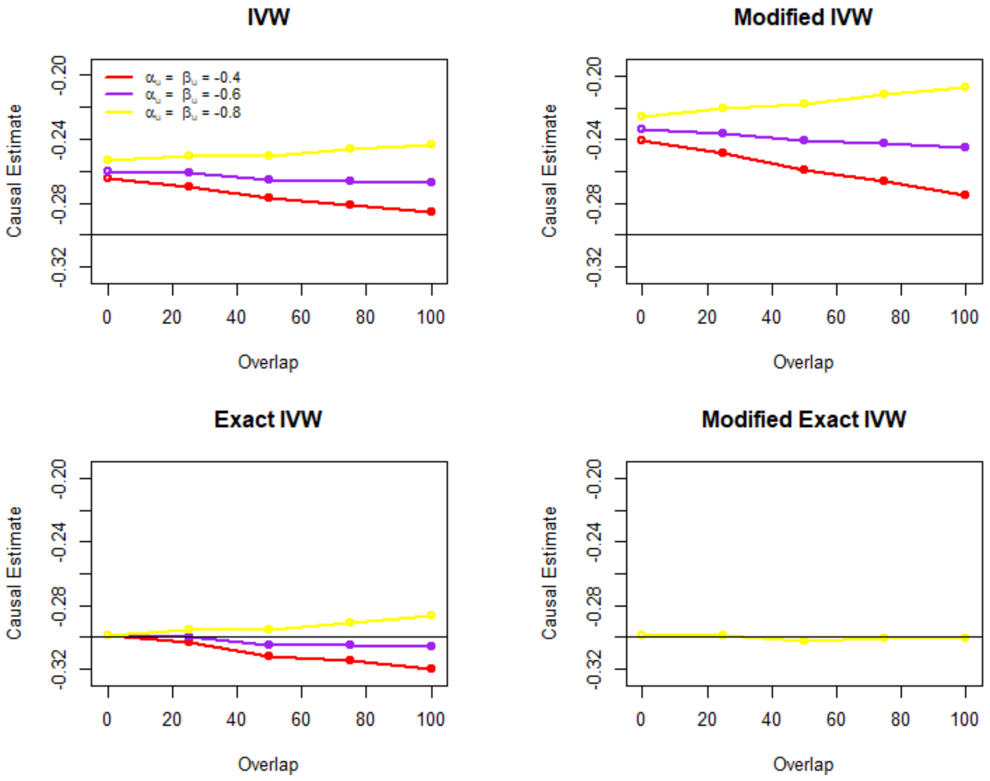


Figure 3.4: Causal estimates with different strengths of confounder. Scenario: Significant threshold $5 \cdot 10^{-2}$; causal effect=-0.3 (horizontal line). Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

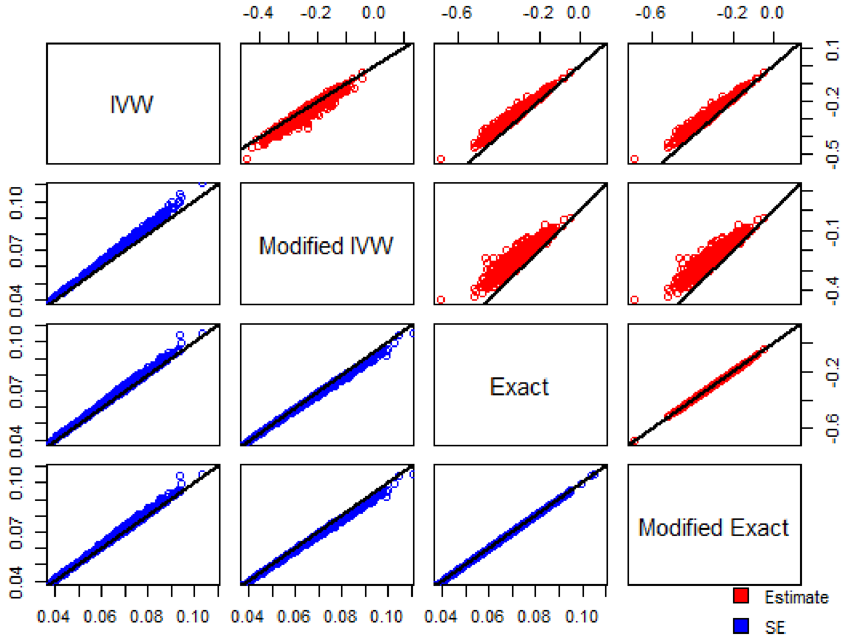


Figure 3.5: The pairwise comparison in standard error and causal estimate for different methods. Significant threshold $5 \cdot 10^{-2}$; causal effect=-0.3; degree of overlap=0%; weak confounder ($\beta_u = \alpha_u = -0.4$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

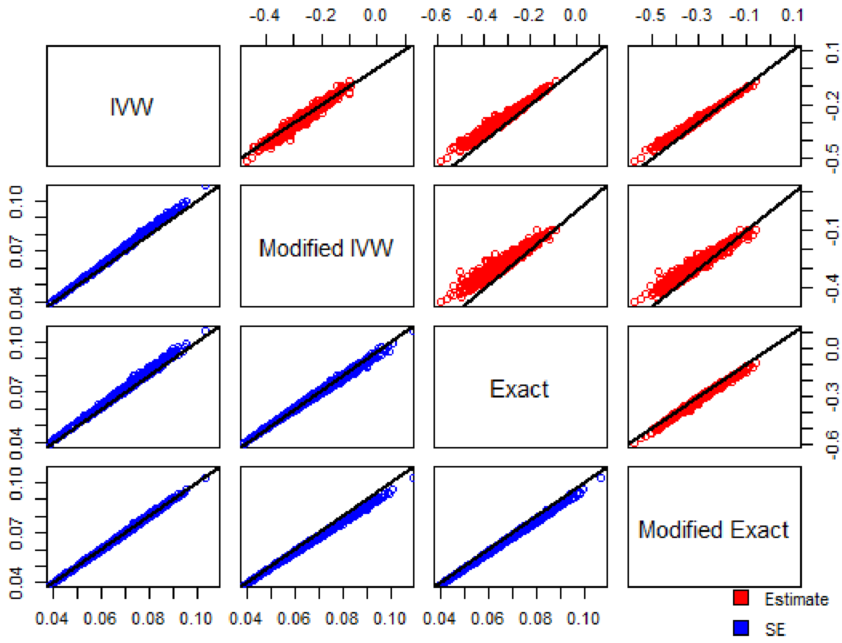


Figure 3.6: The pairwise comparison in standard error and causal estimate for different methods. Significant threshold $5 \cdot 10^{-2}$; causal effect=-0.3; degree of overlap=100%; weak confounder ($\beta_u = \alpha_u = -0.4$). Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Table 3.6: Type-I error with significance threshold of $5 \cdot 10^{-2}$. Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Percentage of overlap	0%	25%	50%	75%	100%
$\beta_c = 0$, $\beta_u = \alpha_u = 0.4$					
IVW	0.050	0.057	0.049	0.059	0.062
Modified IVW	0.029	0.036	0.035	0.054	0.086
Exact	0.085	0.081	0.071	0.086	0.093
Modified exact	0.085	0.077	0.066	0.075	0.065
$\beta_c = 0$, $\beta_u = \alpha_u = 0.6$					
IVW	0.052	0.060	0.056	0.102	0.131
Modified IVW	0.029	0.050	0.069	0.172	0.280
Exact	0.082	0.099	0.090	0.141	0.182
Modified exact	0.082	0.086	0.069	0.078	0.064
$\beta_c = 0$, $\beta_u = \alpha_u = 0.8$					
IVW	0.053	0.078	0.085	0.166	0.257
Modified IVW	0.025	0.067	0.159	0.354	0.645
Exact	0.093	0.107	0.133	0.220	0.346
Modified exact	0.093	0.100	0.071	0.082	0.070

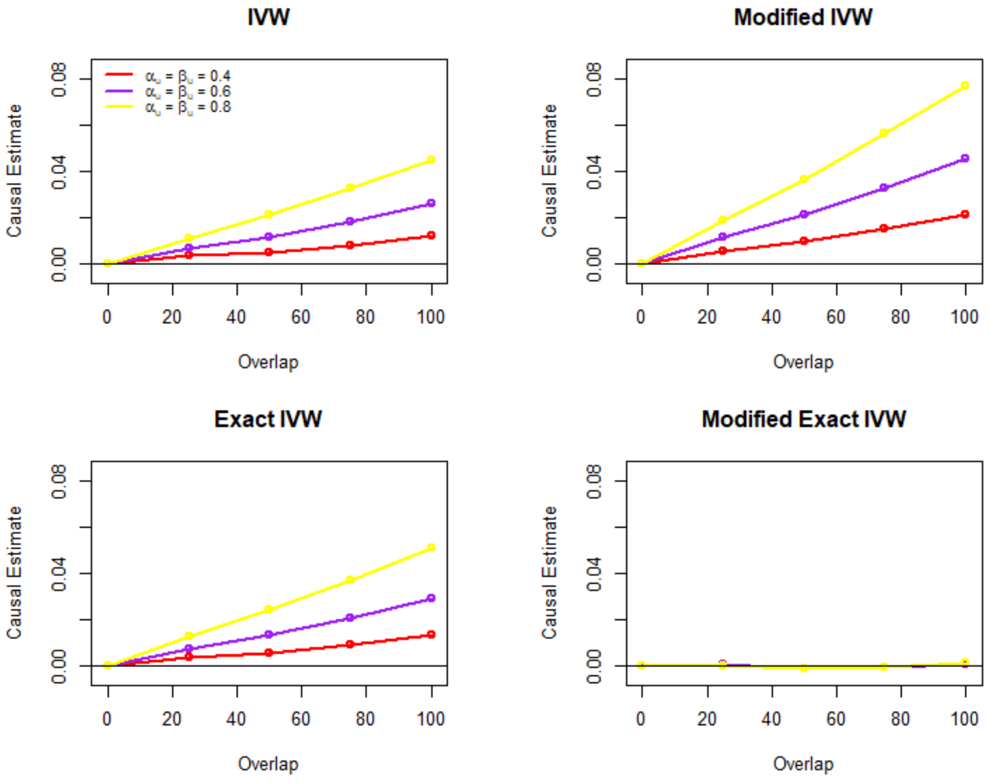


Figure 3.7: Estimate with significance threshould of $5 \cdot 10^{-2}$ and null causal effect. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

3.5 Application

In the first example involving BMI-SBP (Table 3.7), both traits exhibit a high degree of sample overlap, with only a small number of missing observations in either BMI or SBP. As the threshold becomes more strict, all methods, except for the modified exact method, present a decrease in the point estimate . Notably, the modified IVW and exact methods exhibit greater reduction compared to the IVW method. This consistent pattern aligns with the simulation results, suggesting a positive causal effect with a high degree of overlap. For instance, in Table 3.2, when 100% overlap is present, the IVW estimate decreases by 17% (from 0.126 to 0.105) with an increase in IV strength, while the modified IVW estimate decreases by 22% (from 0.145 to 0.113), and the exact estimate experiences a 23% drop (from 0.139 to 0.107).

In this real data example, the IVW, modified IVW, and exact estimates demonstrate reductions of 13%, 16%, and 34%, respectively, as a result of an increase in IV strength. The standard errors among different methods are similar, albeit slightly smaller in the modified exact method due to the inclusion of covariance in the variance term. All confidence intervals indicate a significant positive causal relationship between BMI and SBP. Notably, the modified exact method provides stable estimates across varying thresholds.

When evaluating the second example of BMI-HDL-C, we can use the simulation result featuring a 75% as a reference for the real data scenario. Table 3.8 shows that as the strength of IVs becomes stronger, IVW and modified IVW methods demonstrate a decrease in the estimate, whereas the exact method demonstrates an increase in the estimate. Specifically, the IVW and modified IVW methods yield estimates greater than those of the modified exact method, while the exact method produces estimates lower than the modified exact method. These findings align with the simulation results presented in Table 3.4. Importantly, all estimates demonstrate statistical significance at a 5% significance level, suggesting a negative causal relationship between BMI and HDL-C.

In summary, the modified method offers more stable estimates when the strength of IVs varies compared to other methods.

Table 3.7: Causal estimate, standard error (SE) and 95% confidence interval (CI) for the BMI-SBP example. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights.

	p-value $\leq 5 \cdot 10^{-2}$		
	Estimate	SE	CI
IVW	0.117	0.0083	(0.100,0.133)
Modified IVW	0.129	0.0087	(0.112,0.145)
Exact	0.141	0.0084	(0.124,0.157)
Modified Exact	0.098	0.0082	(0.082,0.114)
	p-value $\leq 5 \cdot 10^{-5}$		
	Estimate	SE	CI
IVW	0.107	0.0103	(0.087,0.127)
Modified IVW	0.119	0.0105	(0.099,0.140)
Exact	0.118	0.0104	(0.098,0.139)
Modified Exact	0.097	0.0101	(0.077,0.117)
	p-value $\leq 5 \cdot 10^{-8}$		
	Estimate	SE	CI
IVW	0.102	0.0132	(0.076,0.128)
Modified IVW	0.108	0.0133	(0.082,0.134)
Exact	0.107	0.0132	(0.081,0.133)
Modified Exact	0.096	0.0129	(0.071,0.122)

Table 3.8: Causal estimate, standard error (SE) and 95% confidence interval (CI) for the BMI-HDL-C example. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights.

p-value $\leq 5 \cdot 10^{-2}$			
	Estimate	SE	CI
IVW	-0.301	0.017	(-0.334, -0.268)
Modified IVW	-0.274	0.017	(-0.308, -0.240)
Exact	-0.375	0.017	(-0.409, -0.341)
Modified Exact	-0.341	0.017	(-0.374, -0.309)
p-value $\leq 5 \cdot 10^{-5}$			
	Estimate	SE	CI
IVW	-0.314	0.018	(-0.349, -0.280)
Modified IVW	-0.282	0.018	(-0.318, -0.246)
Exact	-0.368	0.018	(-0.404, -0.332)
Modified Exact	-0.345	0.018	(-0.380, -0.311)
p-value $\leq 5 \cdot 10^{-8}$			
	Estimate	SE	CI
IVW	-0.322	0.020	(-0.361, -0.283)
Modified IVW	-0.287	0.021	(-0.327, -0.246)
Exact	-0.356	0.021	(-0.396, -0.315)
Modified Exact	-0.342	0.020	(-0.381, -0.303)

3.6 Summary

In this study, we have adapted an existing exact method to accommodate overlapping samples and weak instrument variables. By utilizing real data examples and conducting simulations based on those examples, we have demonstrated that our proposed method offers distinct advantages in providing more accurate estimates when there are both overlap and weak instrument bias present.

We extensively tested our approach using various simulation scenarios, including different strengths of IVs, varying degrees of overlap, the effect of confounders, and different sample sizes for the exposure and outcome variables. Our results consistently showed that the modified exact method performed exceptionally well across all these scenarios.

The conventional first-order IVW method and the second-order modified IVW method were increasingly influenced by weak instrument bias as the strength of the instruments decreased. Although the exact method could mitigate weak instrument bias in non-overlapping scenarios, it still exhibited increased bias as the degree of overlap increased. In contrast, the modified exact method effectively handled weak instrument bias and overlapping samples by incorporating a covariance term in the weight assigned to each ratio estimate.

Moreover, the simulation results indicated that the IVW method suffers from inflated type-I error under certain conditions, while the modified exact IVW method offers a more reliable approach, keeping the type-I error not much affected with increased overlap and strong confounders.

The direction and magnitude of bias were found to depend on the sample overlap, the direction of the confounded observational association, the direction of causal effect, and the strength of the confounder. As expected, when the samples did not overlap, the bias in causal estimator is in the direction of null, but when the samples are completely overlapping, bias is direction of the confounder. Additionally, we observed an interaction effect between the strength of the confounder and the degree of overlap. For instance, in the scenarios where the direction of confounder and causal effect are both positive, a strong confounder increased the bias at a higher rate compared to a weak confounder. Conversely, for a negative causal effect and a positive confounder, the bias decreased at a higher rate with a weak confounder.

Although there was slight underestimation of standard error due to the inclusion of covariance, the impact was not substantial. The coverage of 95% CIs was around 90% to 95% in different scenarios and outperformed the IVW method when F-statistic was around 10 with complete overlapping.

This method can be easy to be implemented because it only requires summary statistics, which are often publicly available. To obtain the correlation term between G-X and G-Y association estimates, one can utilize the cross-trait LD score regression. This can be accomplished using the "ldsc" function within the *GenomicSEM* package (Grotzinger *et al.*, 2019). Notably, there is no requirement to have information about the degree of overlap by running LDSC, further simplifying the process.

The scope of our approach is limited because we solely focus on quantitative traits. In order to examine the overlap between a quantitative study of exposure and a case-control study for outcome, we might need to consider the specific sources of overlap: whether they originate from the case samples, the control samples, or both. In a study conducted by Burgess *et al.* (2016), they demonstrated that in a simulation analysis with no actual causal effect, bias was not observed when estimating the association between G-X solely in the control group. However, when estimating the G-X association across all participants, the relative bias was close to the inverse of the mean of F-statistics. Further investigation is required, particularly in scenarios where a causal effect exists and there are varying levels of overlap. In addition, the method to estimate of correlation between quantitative and case-control studies would be different from that used for two quantitative traits. A point biserial correlation coefficient might be used as a correlation coefficient in this case.

The degree of bias decreases when the instrument variables used to estimate causal effects are obtained from an independent study that strictly considers the p-value threshold. We suggest that researchers evaluate the presence of weak instrument bias by calculating F-statistic and examine the correlation between the associations of variables G-X and G-Y using LDSC regression. If F-statistic is much greater than 10 and the correlation between G-X and G-Y association is close to 0, then it is safe to use the conventional IVW method. However, if a correlation is observed and there is an indication of weak instrument bias, we recommend using our modified exact method. The R code for the proposed method is available in `https://github.com/Bianmj/Overlap`.

Chapter 4

Identifying Invalid Instruments in Mendelian Randomization

As discussed in Section 1.6.3, Assumption IV3 states that valid IVs should not exhibit horizontal pleiotropy on the outcome. In this chapter, our focus is on identifying potential invalid IVs that may display horizontal pleiotropy on the outcome. Prior to conducting MR analysis, researchers typically subject potential instrument variables to multiple sensitivity tests to identify the invalid IVs. Any identified invalid IVs are then excluded from the MR analysis.

However, when testing the causal effect hypothesis, the data-driven selection process for instrumental variables is often overlooked. In this chapter, we assess the consequences of disregarding this selection process by utilizing existing selection methods and proposing a novel approach based on the shrinkage of coefficient method. Our proposed method for selecting invalid IVs is conditional on exposure. Additionally, we introduce the use of the bootstrap method to account for the IV selection process. Simulation studies demonstrate that the bootstrap intervals approach the nominal level of coverage rate.

4.1 Introduction

Assumption IV3, which implies “no horizontal pleiotropy,” is crucial in Mendelian randomization analysis. Including pleiotropic variants in such analysis can introduce bias in causal effect estimates and elevate the type I error rates when testing the causal null hypothesis (Burgess and Thompson, 2013). Studies have revealed that genetic variants can influence multiple diseases and traits identified through GWAS studies, indicating the presence of pleiotropy (Sivakumaran *et al.*, 2011; Pickrell *et al.*, 2016).

To address the issue of pleiotropy, researchers have developed various methods that can be categorized into two groups (Slob and Burgess, 2020). The first group involves down-weighting or removing outliers, such as MR-PRESSO, MR-Lasso, MR-Robust, and MR-RAPS. The second group attempts to model the distribution of estimates from invalid IVs, including methods like MR-Egger, contamination mixture, and MR-Mix. It is worth noting that all these methods are designed for summary-level data. If individual-level data is available, one can also employ the sisVIVE method (Kang *et al.*, 2016) to identify invalid genetic variants and estimate the causal effect. Similar to Lasso, sisVIVE employs L_1 penalized methods to identify invalid IVs. The above methods are introduced in Section 1.8.5 and we thoroughly discussed MR-PRESSO, MR-Lasso and sisVIVE in Appendix C. After detecting the invalid IVs based on the three assumptions, researchers can proceed to estimate the causal effect and confidence intervals. Following the sisVIVE method, a median-type estimator based on Lasso is proposed (Windmeijer *et al.*, 2019), which remains consistent regardless of the IV correlation structure, provided that the proportion of invalid IVs is less than 50%.

Unfortunately, current methods for selecting valid IVs, such as MR-Lasso and MR-PRESSO, do not consider the instrument selection process when computing p-values or confidence intervals for testing causal effects. Moreover, several recent studies use the same data for both IV selection and causal effect estimation. For instance, Rees *et al.* (2019) employed all samples to estimate genetic associations in a one-sample setting using MR-Lasso. Bao *et al.* (2019) utilized the sisVIVE method to exclude the invalid IVs and estimate causal effects without considering the selection process.

Classical inference assumes that the model is specified before data collection, while post-selection inference involves selecting the model based on the data. Using classical inference following selection is problematic as the fact we have searched the model. The conventional inference ignores the prior selection process and it assumes a fixed hypothesis testing. However, since the model is selected based on the data, it is not fixed as we search through the model and choose the best model. Sample splitting is one way to address this issue, where samples are divided into two parts: one for selection and the rest for hypothesis testing. This approach ensures that the conditional null distribution given the selection model becomes the unconditional one, as all observations are independent.

To the best of our knowledge, only a few studies in MR have taken into account instrument selection. For instance, Bi *et al.* (2019) proposed a sampling method to generate the conditional null distribution of test statistics given the selected instruments. Their approach involves solving a randomized version of the sisVIVE selection procedure and reparametrizing the conditional null distribution of test statistics for

MR.

In this chapter, we aim to highlight the problem of ignoring the instrument selection process through a simulation study. We propose a method based on coefficient shrinkage to detect invalid IVs and introduce a bootstrap method to account for the selection process. It is worth noting that, similar to sisVIVE, our new method requires individual-level data so it does not apply to MR based on summary statistics. Our simulation study illustrates the issue of neglecting selection when applying conventional inference to the proposed and existing selection methods (MR-Lasso and sisVIVE). For computational feasibility, MR-PRESSO is excluded due to its limitations with large numbers of variants in a reasonable running time. Additionally, we conduct a simulation study to evaluate the performance of our proposed bootstrap method in accounting for model selection.

4.2 Model and Methods

4.2.1 Model

We assume the following Model (4.2.1) to generate data.

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{G}\boldsymbol{\phi} + \boldsymbol{X}\beta_c + \boldsymbol{U}\beta_u + \boldsymbol{\varepsilon_y} \\ \boldsymbol{X} &= \boldsymbol{G}\boldsymbol{\alpha} + \boldsymbol{U}\alpha_u + \boldsymbol{\varepsilon_x} \end{aligned} \tag{4.2.1}$$

The causal effect is denoted by β_c. The exposure $\boldsymbol{X}$ is a linear combination of genetic variants $\boldsymbol{G}$, a confounder $\boldsymbol{U}$, and random errors $\boldsymbol{\varepsilon_x}$. The outcome $\boldsymbol{Y}$ is a linear function of the genetic variants, exposure, confounder and random errors $\boldsymbol{\varepsilon_y}$. Genetic variants that are chosen to be the instrument variables in MR analysis are assumed to be independent.

If genetic variant i is a valid IV that satisfies Assumptions IV1-IV3, then there would be no direct effect of a genetic variant on the outcome given the exposure (i.e., $\phi_i = 0$). If genetic variant i is an invalid IV that violates Assumption IV3, then there remains an effect of a genetic variant on outcome even after conditioning on the exposure (i.e., $\phi_i \neq 0$).

4.2.2 Proposed Methods for Identifying Invalid Instruments

We explore several approaches for identifying invalid IVs. In the initial stage, a selection method is utilized to identify invalid IVs, which are subsequently excluded in estimating the causal effect. The choice of estimation method can be made by investigators, with Two Stage Least Squares (TSLS) being the most commonly used approach for individual data (section 1.7). In cases where the instruments are weak, methods like LIML (Anderson and Rubin, 1949) and Fuller (Fuller, 1977) are recommended.

The proposed method is based on the idea that conditional on exposure, a valid IV should not exhibit any effect on the outcome. If there is an effect, it indicates that the genetic variant is invalid. We investigate two strategies for identifying variants with pleiotropy effects ($\phi_i \neq 0$):

(i) Identifying variants that reject the null hypothesis (H_0: the pleiotropy effect $\phi_i = 0$ for variant i) at a pre-determined significance level is performed using multiple regression with these variants.

(ii) Identifying the variants whose coefficients are not shrunk to zero using a L_1

penalty term in the regression model. The objective function is given in Equation (4.2.5).

Method (i) is based on traditional hypothesis testing, whereas method (ii) incorporates L_1 regularization into the regression model of method (i). However, since the Lasso regression approach outperforms the traditional testing approach, we present only the method based on Lasso regression in this chapter, while the method of traditional hypothesis testing is presented in Appendix C.

Lasso

Consider the Model 4.2.1 with $\boldsymbol{Y}$ as the outcome variable, $\boldsymbol{X}$ as the exposure, $\boldsymbol{U}$ as the unobserved confounder, and $\boldsymbol{G}$ as the matrix for potential IVs. We define $\boldsymbol{v_1} = \beta_u \boldsymbol{U} + \boldsymbol{\varepsilon_y}$ and $\boldsymbol{v_2} = \alpha_u \boldsymbol{U} + \boldsymbol{\varepsilon_x}$, and express the linear projection of $\boldsymbol{v_1}$ on $\boldsymbol{v_2}$ in error form as $\boldsymbol{v_1} = \beta_u(\frac{\boldsymbol{v_2} - \boldsymbol{\varepsilon_x}}{\alpha_u}) + \boldsymbol{\varepsilon_y} = \rho \boldsymbol{v_2} + \boldsymbol{e}$, where $\rho = \frac{\beta_u}{\alpha_u}$ and $\boldsymbol{e} = -\rho\boldsymbol{\varepsilon_x} + \boldsymbol{\varepsilon_y}$ which has mean 0 and variance $\rho^2 + 1$. This allows us to rewrite $\boldsymbol{Y}$ as:

$$\boldsymbol{Y} = \boldsymbol{G\phi} + \boldsymbol{X}\beta_c + \boldsymbol{v_2}\rho + \boldsymbol{e} \tag{4.2.2}$$

Although $\boldsymbol{v_2}$ is not observed, we can estimate $\boldsymbol{\hat{v}_2}$ by taking the OLS residuals from the first-stage regression of $\boldsymbol{X}$ on $\boldsymbol{G}$:

$$\boldsymbol{\hat{v}_2} = \boldsymbol{X} - \boldsymbol{G}\hat{\gamma} \tag{4.2.3}$$

By substituting $\boldsymbol{\hat{v}_2}$ in Equation (4.2.3) into Equation (4.2.2), we obtain:

$$\boldsymbol{Y} = \boldsymbol{G\phi} + \boldsymbol{X}\beta_c + \boldsymbol{\hat{v}_2}\rho + \boldsymbol{e} \tag{4.2.4}$$

To detect the invalid IVs, we use the Lasso regression method by minimizing the sum of squared residuals from Equation (4.2.4) and adding a penalty term for the coefficient terms of genetic variants:

$$\arg\min_{\boldsymbol{\phi}} \|(\boldsymbol{Y} - \boldsymbol{G}\boldsymbol{\phi} - \hat{\boldsymbol{v}}_{\boldsymbol{2}}\rho - \boldsymbol{X}\beta_c)\|_2^2 + \lambda\|\boldsymbol{\phi}\|_1 \tag{4.2.5}$$

We define the squared L_2 norm for the vector $\boldsymbol{X}$ ($\|\boldsymbol{X}\|_2^2$) as the sum of the squares of its elements, and L_1 norm $\|\boldsymbol{X}\|_1$ as the sum of the absolute values of its elements. In the above objective function, only $\boldsymbol{\phi}$ is penalized. If ϕ_i for variant i is not penalized to zero, then variant i is considered "invalid".

The Lasso regression in Equation (4.2.5) includes all the variants in the objective function. The stopping rule uses Cochran's Q test from Bowden *et al.* (2018), which can be written as $Q = \sum_j Q_j = \sum_j w_j(\hat{\beta}_j - \hat{\beta}_{IVW})^2$. The assumption of this test is that each genetic variant is a valid IV. Cochran's Q test provides evidence for heterogeneity across instrument variables and the presence of invalid IVs. The tuning parameter λ is sorted in increasing order. For each λ, we identify "invalid" variants, fit IVW regression excluding those invalid variants, and calculate Cochran's Q statistic for each model. If there is no pleiotropy effect, then Cochran's Q statistic should follow a χ^2 distribution with $M - 1$ degrees of freedom (M is the number of variants in the model). We then select a model with the largest number of valid variants where Cochran's Q statistic is not significant at a pre-determined level of significance. In our simulations, we use a significance level of 5%. We have not explored the influence of varying the significance level used for the stopping rule on the results. We would anticipate that a less stringent threshold would lead to the identification of more IVs as invalid.

4.2.3 Methods to Account for Selection Process

In conventional inference, the null hypothesis of the causal effect is $H_0 : \beta_c = \beta_0$. The $1-\alpha$ confidence interval for β_c is constructed by inverting the pivotal null distribution of T, where $T = (\hat{\beta}_c - \beta_0)/\hat{\sigma}$, and $\hat{\sigma}$ is the standard error of $\hat{\beta}_c$:

$$P_{H_0:\beta_c=\beta_0}(T \geq t) \leq \alpha \tag{4.2.6}$$

However, Equation (4.2.6) does not taken into account the selection process and it assumes that the model is pre-specified while in reality the selected model is based on data. The goal is to obtain the conditional null distribution

$$P_{H_0:\beta_c=\beta_0}(T \geq t|\text{selected model}) \leq \alpha \tag{4.2.7}$$

We consider three ways to account for the selection process: (i) Case resampling bootstrap method, (ii) Sample splitting, and (iii) Sample splitting with case resampling bootstrap. In the bootstrap method, exposure x_i, outcome y_i, and genetic variants $\boldsymbol{G_i}$ are sampled with replacement from the original data. The selection method is then used to remove the invalid IVs, the remaining markers are then used to estimate the causal effect $\hat{\beta}_c^*$. Bootstrap estimates from R bootstrap samples are then used to construct bootstrap intervals.

Sample splitting is an alternative way to deal with this issue. Basically, the data is divided into two parts. One part is used in selection and the rest is used for estimation. In this case, the conditional null distribution given the selection model becomes the unconditional one since all the observations are independent.

If the original dataset is divided in half, this splitting procedure is repeated in each bootstrap sample, resulting in a subset of resampled observations used for the detection of invalid IVs and the remaining used to calculate the bootstrap estimates.

4.2.4 Confidence Intervals

We consider three types of confidence intervals, one is the Wald-type confidence interval and the other two are bootstrap confidence intervals - percentile and normal intervals. They are all based on the TSLS causal estimate. The following shows these three confidence intervals.

- Wald-type confidence intervals:

$$\left[\hat{\beta}_c - z_{1-\alpha/2}\ \hat{\sigma}, \hat{\beta}_c - z_{\alpha/2}\ \hat{\sigma}\right] \tag{4.2.8}$$

  where $\hat{\beta}_c$ is the causal estimate, $\hat{\sigma}$ is the standard error of causal estimate, and $z_{\alpha/2}$ is $\alpha/2$ percentile of standard normal distribution.

- Percentile confidence intervals:

$$\left[\hat{\beta}^*_{(\alpha/2)}, \hat{\beta}^*_{(1-\alpha/2)}\right] \tag{4.2.9}$$

  where $\hat{\beta}^*_{(\alpha/2)}$ and $\hat{\beta}^*_{(1-\alpha/2)}$ denote $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap estimates $\hat{\beta}^*_c$, respectively.

- Normal confidence intervals:

$$\left[\hat{\beta}_c - z_{1-\alpha/2}\ \hat{\sigma}_b, \hat{\beta}_c - z_{\alpha/2}\hat{\sigma}_b\right] \tag{4.2.10}$$

where $\hat{\sigma}_b = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\beta}_r^* - \frac{1}{R}\sum_{b=1}^{R}\hat{\beta}_b^*)^2}$, and R is the number of bootstrap samples. Note that the normal confidence interval is analogous to the Wald-type confidence interval, but replaces the standard error of estimate by the bootstrap standard error $\hat{\sigma}_b$.

We also explore alternative bootstrap intervals, including the basic interval, studentized interval, and normal interval, which includes the bootstrap estimate of bias. We assess their performances based on our proposed Lasso method. Further details can be found in Appendix C and Table C.12.

We observed that studentized intervals consistently exhibit over-coverage and have the widest interval lengths. On the other hand, basic, percentile, and normal intervals with the bootstrap estimate of bias show similar performances in terms of coverage and width of intervals. Notably, Normal intervals exhibit a higher coverage at 95% confidence intervals and a smaller average width.

4.3 Simulation Framework

We conducted a simulation study to compare the performance of MR-Lasso and sisVIVE with our two proposed methods in a realistic setting. The data were generated using Model (4.2.1). To mimic real-world scenarios, we generated the G-X association from a normal distribution $N(0, 0.02)$, but conditioned on the absolute value being greater than a threshold (such as 0.15), resulting in a standard deviation of around 0.03. This approach ensures that the G-X association is significantly different from zero, as MR typically selects IVs that pass a significance threshold, like $5 \cdot 10^{-8}$.

We considered two scenarios:

- Balanced pleiotropy: The mean of pleiotropy effects (ϕ_i) is set to zero. In this case, invalid IVs were generated from $N(0, 0.03)$, while valid IVs have $\phi_i = 0$.
- Directional pleiotropy: The mean of pleiotropy effects is away from zero. Here, the pleiotropy effect ϕ_i is generated from $N(0.03, 0.03)$, and $\phi_i = 0$ for valid IVs.

The causal effect (β_c) was set to 0.1, and the effects of the confounder on $\boldsymbol{X}$ and $\boldsymbol{Y}$ were denoted by α_u and β_u, respectively, both equal to 0.3. Each genetic variant $\boldsymbol{G_j}$ ($j = 1, 2, ..., M$) was independently generated from $Binomial(2, p_j)$, with p_j sampled from Unif$(0.1, 0.5)$. We standardized $\boldsymbol{G_j}$ to have a mean of 0 and a variance of 1. The error terms $\boldsymbol{\varepsilon_x}$ and $\boldsymbol{\varepsilon_y}$ were independent and normally distributed with a mean of 0. The variances of the error terms were chosen such that $\boldsymbol{U}$, $\boldsymbol{X}$, and $\boldsymbol{Y}$ all had unit variance.

The simulation was repeated 1000 times with a sample size of $n = 100,000$ ($100k$) for each run. To highlight the importance of accounting for the instrument selection process, we used the entire data for both selection and estimation. When accounting for the selection process, we split the data into two subsets (n_1 and n_2), with one subset used for selection (n_1) and the rest for estimation ($n_2 = n - n_1$). We employed the TSLS method for estimating the causal effect and considered different sizes of observations for selection and estimation ($n_1 = 10k, 30k$ or $50k$). We simulated data for 120 genetic variants, with a proportion of them being invalid (1/3: 80 valid and 40 invalid; 1/2: 60 valid and 60 invalid; 2/3: 40 valid and 80 invalid).

4.4 Simulation Results

Here, we only present the results for three shrinkage-based methods (MR-Lasso, our proposed Lasso method, and sisVIVE). Results concerning hypothesis testing methods are detailed in Appendix C, with qualitative discussions provided in this section.

4.4.1 Simulation result without accounting for selection

Initially, we conducted the analysis using the entire dataset (100k) observations) for both identifying valid IVs and computing the test statistic. This step was crucial to emphasize the issue of disregarding the instrument selection process.

The simulation results, without accounting for the instrument selection process, are presented in Tables 4.1, 4.2, and 4.3. We considered scenarios of balanced and directional pleiotropy, varying the proportion of valid IVs. For each scenario, we provided key metrics, such as the true positive rate (TPR), false positive rate (FPR), mean, median, standard error, and standard deviation of causal estimates across simulations. We also reported the bootstrap standard error (B.SE), coverage rate of a 95% Wald-type confidence level, and the average width of these confidence intervals over 1000 simulations. TPR is calculated as the probability that true invalid IVs are correctly identified and FPR is calculated as the probability that true valid IVs are wrongly identified as invalid IVs.

Table 4.1: True positive rate (TPR) and false positive rate (FPR) over 1000 simulations for MR-Lasso, Lasso and sisVIVE. True positive rate is the probability that the invalid IVs are truly identified and false positive rate is the probability that valid IVs are wrongly classified as "invalid" ones. The entire data (100k) is used for both selection and estimation.

	80 valid IVs		60 valid IVs		40 valid IVs	
Method	TPR	FPR	TPR	FPR	TPR	FPR
Scenario: balanced pleiotropy						
MR-Lasso	0.800	0.024	0.819	0.035	0.828	0.053
Lasso	0.755	0.006	0.784	0.012	0.802	0.026
sisVIVE	0.746	0.004	0.775	0.009	0.793	0.019
Scenario: directional pleiotropy						
MR-Lasso	0.869	0.020	0.877	0.028	0.885	0.046
Lasso	0.835	0.004	0.854	0.008	0.869	0.020
sisVIVE	0.829	0.003	0.848	0.005	0.864	0.014

Table 4.2: Mean,standard deviation (SD),mean of standard error (SE) and mean of bootstrap standard error (B.SE) for TSLS estimates. The true causal effect is 0.1. Results are reported over 1000 simulations. The entire data (100k) is used for both selection and estimation.

	80 valid IVs				60 valid IVs				40 valid IVs			
Method	Mean	SD	SE	B.SE	Mean	SD	SE	B.SE	Mean	SD	SE	B.SE
Scenario: balanced pleiotropy												
MR-Lasso	0.100	0.014	0.012	0.017	0.100	0.019	0.013	0.019	0.101	0.026	0.015	0.023
Lasso	0.100	0.014	0.012	0.016	0.100	0.019	0.013	0.018	0.101	0.027	0.015	0.021
sisVIVE	0.101	0.015	0.012	0.016	0.100	0.019	0.013	0.018	0.101	0.027	0.015	0.022
Scenario: directional pleiotropy												
MR-Lasso	0.100	0.014	0.012	0.017	0.101	0.018	0.014	0.020	0.101	0.026	0.016	0.024
Lasso	0.101	0.014	0.012	0.016	0.100	0.017	0.013	0.018	0.101	0.026	0.016	0.022
sisVIVE	0.101	0.015	0.012	0.016	0.100	0.018	0.013	0.018	0.102	0.025	0.016	0.023

We found that TPR remains unaffected by the number of invalid IVs for all approaches. For threshold-based methods, such as false discovery rate (FDR), significance level and Šidák correction (SID), the FPR noticeably increases as the number of invalid IVs increases. In contrast, shrinkage methods (MR-Lasso, Lasso, and sisVIVE) maintain a low FPR (Table C.1 and Table 4.1). Furthermore, the less stringent the significance level, the higher both TPR and FPR. Among the shrinkage methods, Lasso and sisVIVE also have significantly higher true negative counts (the number of valid IVs correctly identified) than MR-Lasso. However, their false negatives counts (the number of invalid IVs wrongly identified as valid) are slighter greater than those of MR-Lasso (Figures 4.1-4.2).

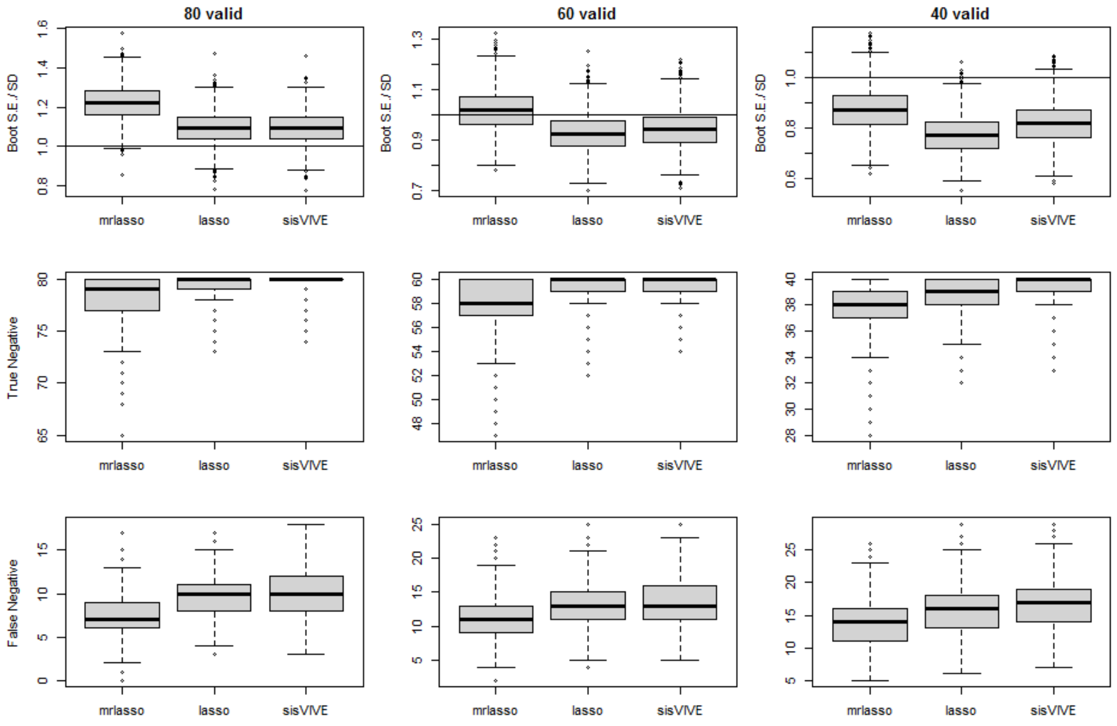


Figure 4.1: Scenario: balanced pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations for MR-Lasso, Lasso and sisVIVE methods. The entire data (100k) is used for selection and estimation.

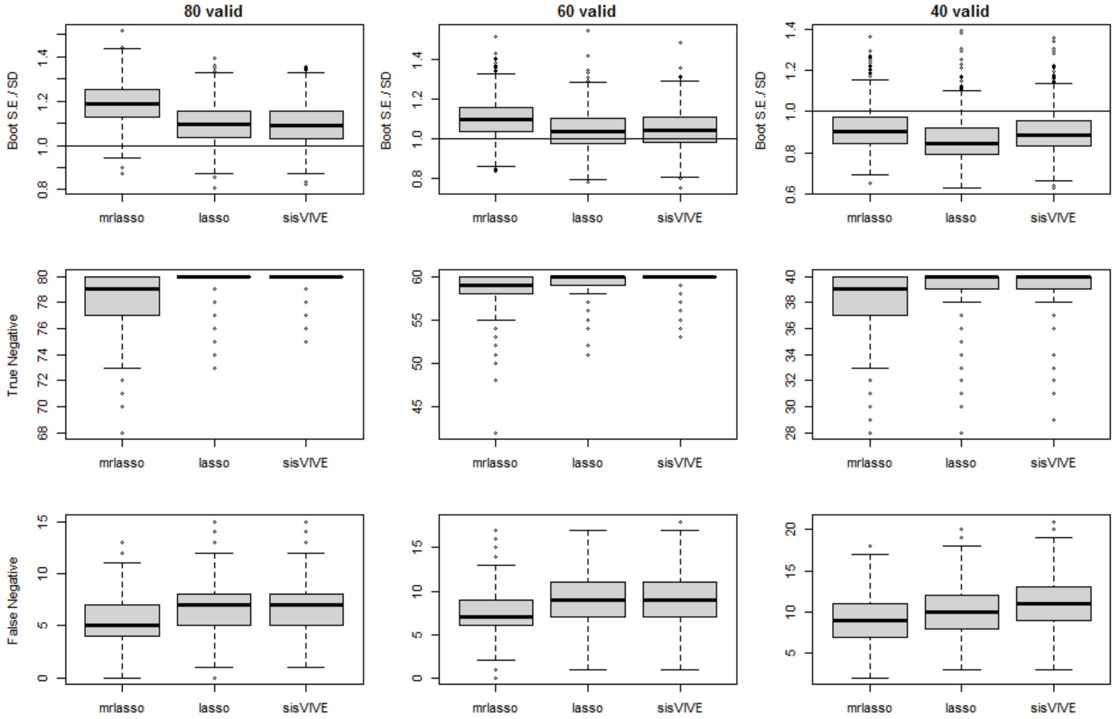


Figure 4.2: Scenario: directional pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations for MR-Lasso, Lasso and sisVIVE methods. The entire data (100k) is used for selection and estimation.

We observed that mean of estimates are all close to the true causal effect in all scenarios, as the bias depends on the direct genetic effect on the outcome and the genetic effect on exposure. When one of these effects has a mean of zero, the bias is close to zero (Windmeijer *et al.*, 2019). The variability (standard deviation) of estimates increases with a greater number of invalid IVs, as TPR tends to remain at a similar level while more invalid IVs are not identified. The standard deviation is consistently greater than the mean of the standard error of estimate (SE), as shown in Table 4.2. As the proportion of invalid IVs increases, the difference between them becomes larger. SE is underestimated due to the lack of consideration for uncertainty caused by IV selection. When the proportion of invalid IVs does not exceed 50%, SE is reasonably approximated using shrinkage methods for selection. However, when the proportion of invalid IVs reaches or exceeds 50%, SE is significantly underestimated, leading to confidence intervals with coverage below the desired 95% nominal level. In such cases, the bootstrap method is used to estimate the standard error, and Table 4.2 shows that the difference between SD and B.SE is much smaller than the difference between SD and SE, particularly with larger proportions of invalid IVs. This indicates that bootstrap can effectively account for the selection process.

Coverage rates are influenced by TPR, true negative rate (TNR: the probability that true valid IVs are correctly identified, which equals to 1-FPR), and the proportion of invalid IVs. Shrinkage methods generally have a lower TPR, a higher coverage and a shorter confidence interval width compared to threshold-based methods (Table C.2 and Table 4.3). For example, with 80 valid IVs, the coverage for the threshold-based method is around 80% in the balanced pleiotropy scenario, while it is around 90% for the shrinkage-based method in the same scenario. Shrinkage methods consistently include the majority of true valid IVs, leading to reduced estimate variability. Conversely, as more invalid IVs remain unidentified and more valid IVs are incorrectly removed, the estimate becomes more variable, and coverage moves away from the nominal 95% level. For instance, in the Lasso method, coverage drops from 90% to 74% as the proportion of invalid IVs increases from 1/3 to 2/3 in the case of balanced pleiotropy. Similarly, for the 5% FDR method, coverage drops from 80% to 58% with an increase in invalid IVs.

To summarize, disregarding the instrument selection process can lead to undercoverage. As the proportion of invalid IVs increases, the impact of selection on coverage estimation becomes more pronounced. While shrinkage methods can achieve high TNR (low FPR), no single perfect selection method exists. Therefore, to improve interval coverage, it is essential to incorporate approaches that account for the selection process.

4.4.2 Simulation Result by Use of Bootstrap

Table 4.2 demonstrates that the bootstrap standard error serves as a reliable estimate of the standard deviation, as it closely aligns with the standard deviation. Consequently, we will employ the bootstrap method to construct confidence intervals that account for the selection process. The results based on the bootstrap approach are presented in Table 4.4. Overall, the coverage rates are substantially improved compared to those in Table 4.3, where selection was not considered, particularly in scenarios with a high proportion of invalid IVs. For instance, when half of the IVs are invalid in the case of balanced pleiotropy, the Wald-type intervals using the Lasso method exhibit 81% coverage, whereas the bootstrap normal intervals achieve a coverage of 93%.

When the proportion of invalid IVs is less than 50%, the bootstrap method yields conservative coverages for shrinkage-based methods (MR-Lasso, Lasso, and sisVIVE), while for proportions exceeding 50%, the coverages are slightly below the nominal level. On the other hand, using the threshold-based technique, the coverage remains significantly underestimated at approximately 70% to 89% across various percentages of valid IVs (Table C.3).

Table 4.3: Coverage and average width of 95% Wald-type intervals over 1000 simulations. The entire data ($100k$) is used for both selection and estimation.

	80 valid IVs		60 valid IVs		40 valid IVs	
Method	Coverage	Width	Coverage	Width	Coverage	Width
Scenario: balanced pleiotropy;						
MR-Lasso	0.913	0.047	0.838	0.052	0.768	0.060
Lasso	0.896	0.046	0.813	0.051	0.740	0.058
sisVIVE	0.893	0.046	0.812	0.050	0.734	0.058
Scenario: directional pleiotropy						
MR-Lasso	0.908	0.047	0.865	0.053	0.786	0.063
Lasso	0.900	0.046	0.862	0.052	0.773	0.061
sisVIVE	0.892	0.046	0.858	0.052	0.778	0.062

4.4.3 Simulation Result Using Sample Splitting

Here, we explore two approaches for constructing the intervals: (i) sample splitting for the original data, and (ii) sample splitting for each bootstrap sample from the entire dataset. Simple sample splitting is utilized to compute the test statistic for Wald-type intervals. When calculating bootstrap intervals, sample splitting is applied to each bootstrap sample. Our aim is to investigate whether the coverage rates of intervals improve compared to the Wald-type intervals without considering the selection process, as presented in Table 4.3.

Table 4.4: Coverage and average width of 95% bootstrap normal intervals over 1000 simulations for MR-Lasso, Lasso and sisVIVE. The entire data (100k) is used for both selection and estimation. 200 bootstrap replicates are generated for each dataset.

	Balanced pleoitropy						Directional pleiotropy					
	80 valid IVs		60 valid IVs		40 valid IVs		80 valid IVs		60 valid IVs		40 valid IVs	
Method	Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width
MR-Lasso	0.984	0.066	0.954	0.075	0.915	0.089	0.976	0.066	0.971	0.076	0.931	0.093
Lasso	0.971	0.061	0.928	0.070	0.878	0.083	0.970	0.061	0.964	0.071	0.912	0.086
sisVIVE	0.969	0.062	0.922	0.071	0.897	0.085	0.972	0.062	0.969	0.072	0.921	0.087

Table 4.5: True positive rate (TPR) and false positive rate (FPR) over 1000 simulations with sample-splitting. True positive rate is the probability that the invalid IVs are truly identified and false positive rate is the probability that valid IVs are wrongly classified as "invalid" ones. The sample size is 100,000; n_1 is the number of observation used in selection.

	Balanced pleoitropy						Directional pleiotropy					
	80 valid IVs		60 valid IVs		40 valid IVs		80 valid IVs		60 valid IVs		40 valid IVs	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
						$n_1 = 10,000$						
MR-Lasso	0.497	0.034	0.527	0.047	0.549	0.064	0.634	0.032	0.672	0.046	0.694	0.066
Lasso	0.418	0.012	0.465	0.020	0.495	0.032	0.585	0.010	0.623	0.017	0.653	0.027
sisVIVE	0.399	0.008	0.445	0.015	0.475	0.024	0.568	0.007	0.608	0.012	0.639	0.021
						$n_1 = 30,000$						
MR-Lasso	0.670	0.030	0.697	0.044	0.710	0.062	0.780	0.026	0.797	0.038	0.808	0.054
Lasso	0.619	0.010	0.748	0.018	0.670	0.028	0.732	0.006	0.762	0.012	0.782	0.023
sisVIVE	0.595	0.007	0.634	0.013	0.657	0.022	0.722	0.005	0.752	0.009	0.773	0.018
						$n_1 = 50,000$						
MR-Lasso	0.732	0.027	0.754	0.040	0.769	0.058	0.823	0.022	0.836	0.034	0.846	0.051
Lasso	0.681	0.007	0.713	0.015	0.734	0.027	0.782	0.005	0.807	0.010	0.824	0.020
sisVIVE	0.690	0.005	0.701	0.011	0.723	0.020	0.774	0.004	0.799	0.007	0.817	0.015

We observe that the TPR is affected when fewer observations are used in the selection process compared to when the entire dataset is used (Tables 4.1 and 4.5).

However, the difference in TPR between the subset of data and the complete data is significantly reduced when half of the data ($50k$) is used for selection. For instance, under directional pleiotropy with half the data used in selection, the TPR rate for MR-Lasso is 0.823, while under the same scenario with all data used in selection, it is 0.869. A reduced sample size in selection, particularly for the threshold-based methods, leads to a lower rate of false positives (Table C.4). The FPR for the shrinkage methods shows only slight variation when different numbers of observations are used in selection (Tables 4.5).

Table 4.6: Mean, standard deviation (SD), mean of standard error (SE), and mean of bootstrap standard error (B.SE) for TSLS estimates with sample-splitting. The true causal effect is 0.1. Results are reported over 1000 simulations. The sample size is 100,000; n_1 is the number of observations used in selection and the remaining is used for estimation.

	80 valid IVs				60 valid IVs				40 valid IVs			
Method	Mean	SD	SE	B.SE	Mean	SD	SE	B.SE	Mean	SD	SE	B.SE
Scenario: balanced pleiotropy; $n_1 = 10,000$												
MR-Lasso	0.101	0.029	0.012	0.021	0.102	0.041	0.013	0.026	0.104	0.055	0.014	0.033
Lasso	0.100	0.032	0.011	0.021	0.102	0.044	0.012	0.026	0.105	0.058	0.013	0.032
sisVIVE	0.100	0.033	0.011	0.022	0.102	0.045	0.012	0.026	0.105	0.058	0.013	0.033
Scenario: directional pleiotropy; $n_1 = 10,000$												
MR-Lasso	0.102	0.030	0.012	0.021	0.101	0.040	0.028	0.013	0.105	0.061	0.015	0.038
Lasso	0.103	0.033	0.012	0.022	0.102	0.044	0.013	0.028	0.106	0.064	0.014	0.036
sisVIVE	0.102	0.034	0.012	0.022	0.102	0.046	0.013	0.028	0.106	0.066	0.014	0.037
Scenario: balanced pleiotropy; $n_1 = 30,000$												
MR-Lasso	0.101	0.020	0.017	0.017	0.101	0.026	0.015	0.021	0.103	0.038	0.017	0.026
Lasso	0.101	0.022	0.013	0.017	0.100	0.028	0.015	0.020	0.101	0.041	0.016	0.025
sisVIVE	0.101	0.022	0.013	0.017	0.101	0.029	0.015	0.021	0.101	0.041	0.016	0.025
Scenario: directional pleiotropy; $n_1 = 30,000$												
MR-Lasso	0.101	0.019	0.014	0.017	0.101	0.025	0.016	0.021	0.102	0.037	0.018	0.027
Lasso	0.101	0.022	0.014	0.017	0.101	0.027	0.015	0.020	0.102	0.040	0.018	0.026
sisVIVE	0.101	0.022	0.014	0.017	0.101	0.027	0.015	0.020	0.103	0.041	0.017	0.026
Scenario: balanced pleiotropy; $n_1 = 50,000$												
MR-Lasso	0.101	0.021	0.016	0.019	0.102	0.024	0.018	0.022	0.102	0.033	0.021	0.027
Lasso	0.101	0.021	0.016	0.019	0.103	0.026	0.018	0.022	0.103	0.036	0.020	0.026
sisVIVE	0.101	0.021	0.016	0.018	0.103	0.029	0.018	0.022	0.103	0.035	0.020	0.026
Scenario: directional pleiotropy; $n_1 = 50,000$												
MR-Lasso	0.101	0.019	0.017	0.019	0.103	0.023	0.019	0.022	0.103	0.031	0.022	0.028
Lasso	0.101	0.020	0.016	0.018	0.102	0.024	0.018	0.022	0.103	0.034	0.021	0.026
sisVIVE	0.101	0.021	0.016	0.018	0.102	0.024	0.018	0.022	0.103	0.034	0.021	0.026

A portion of the data is used to obtain the causal estimates in Table 4.6. When more observations are used in the selection process, the causal estimates become less variable, and the gap between the sample standard deviation and standard error narrows. However, this discrepancy becomes more noticeable as the percentage of invalid IVs reaches 50% or greater.

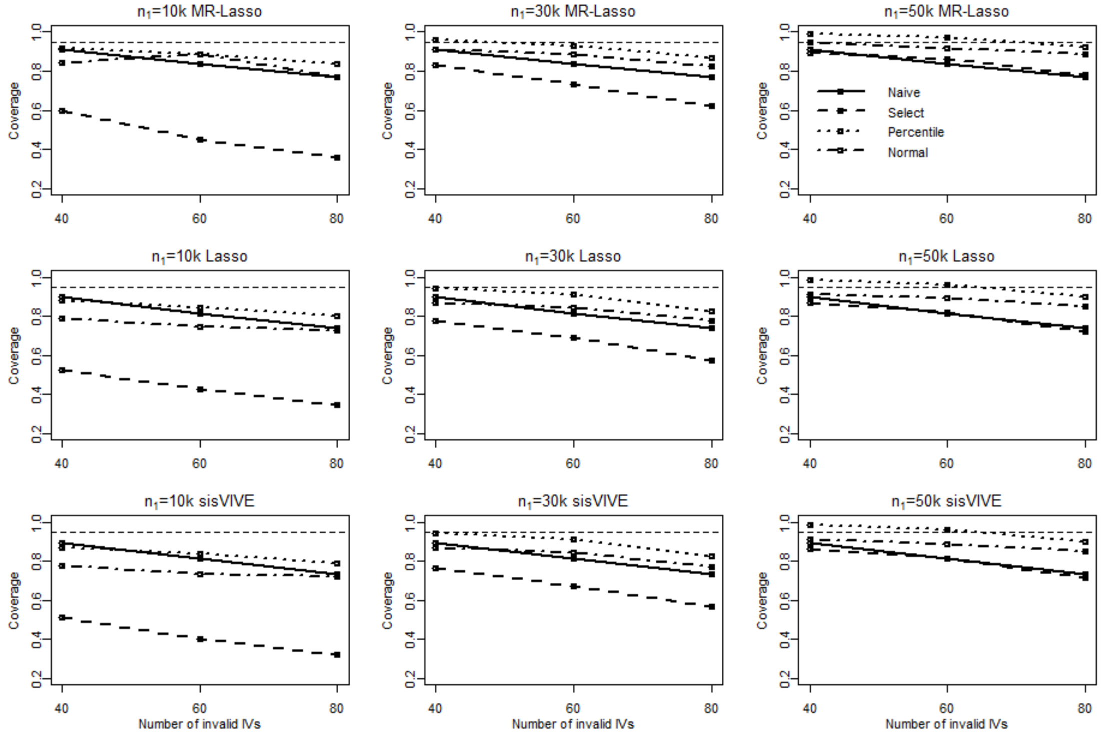


Figure 4.3: Scenario: balanced pleiotropy. Comparing bootstrap and Wald-type intervals for selection with different data splits, including Wald-type intervals using the full dataset. Abbreviations: Naive, Wald-type intervals using the entire data for selection and estimation; Select, Wald-type intervals using part of data; Percentile, Percentile intervals; Normal, Normal intervals. For Select, Percentile and Normal CIs, part of data ($n_1 = 10k$, $30k$ or $50k$) are used for selecting valid IVs and the remaining for estimation

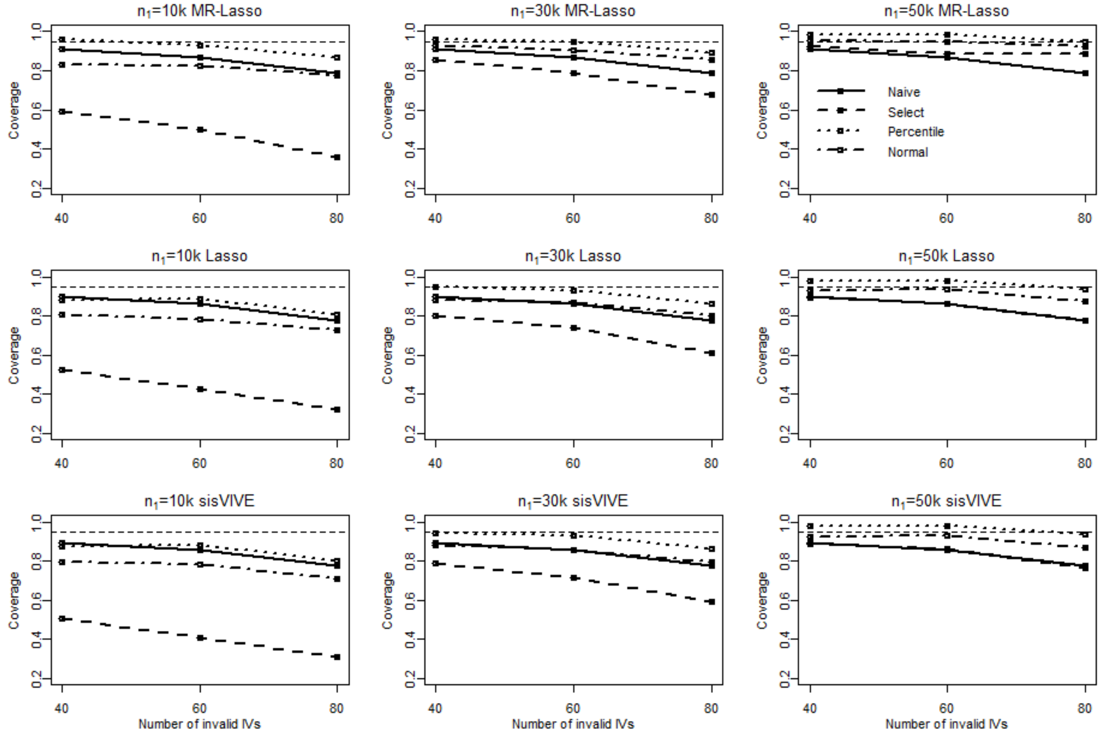


Figure 4.4: Scenario: directional pleiotropy. Comparing bootstrap and Wald-type intervals for selection with different data splits, including Wald-type intervals using the full dataset. Abbreviations: Naive, Wald-type intervals using the entire data for selection and estimation; Select, Wald-type intervals using part of data; Percentile, Percentile intervals; Normal, Normal intervals. For Select, Percentile and Normal CIs, part of data ($n_1 = 10k$, $30k$ or $50k$) are used for selecting valid IVs and the remaining for estimation

When the selection size used for selection is relatively small, such as $10k$ or $30k$, the Wald-type intervals that use sample splitting (select) for shrinkage methods exhibit poorer coverage compared to the naive Wald-type intervals, which utilize all observations for both selection and estimation (Figure 4.3-4.4). However, as the sample size for selection approaches $50k$, the coverage of these two Wald-type intervals becomes comparable. This is not surprising, as the number of invalid IVs left unidentified (false negative counts) increases significantly with a smaller size in selection (Tables C.1, C.2, 4.1), leading to more variable causal estimates. On the other hand, bootstrap percentile intervals and normal intervals generally provide better coverage than the Wald-type intervals, whether or not sample splitting is employed (Figures 4.3-4.4 and Table 4.4). For instance, when the selection size approaches 50k and the percentage of invalidity is less than 50%, the coverage of bootstrap intervals ranges from 88% to 98%, with percentile intervals being slightly conservative (Tables C.10-C.11). In comparison, the coverage for Wald-type intervals that use sample splitting is around 82% to 90% in the same scenario. When employing the entire data for selection, the coverage for Wald-type intervals is around 81% to 91%. As for the comparison of the performance of bootstrap intervals with different sizes in selection, the coverage of bootstrap intervals with sample splitting is not comparable to the coverage of bootstrap intervals with the entire data for selection and estimation until the size for selection approaches 50k (Figures 4.3-4.4 and Table 4.4).

For the threshold-based method, the Wald-type intervals (select) show improved coverage compared to the Wald-type intervals (naive) without accounting for selection as the size for selection approaches 50k, although they still have a low coverage when the proportion of invalid IVs is greater than 50% (Figures C.7-C.8). The coverage of bootstrap intervals with sample splitting improves compared to that of the bootstrap intervals with the entire data when the size for selection is $50k$ (Tables C.3 and C.10-C.11).

Table 4.7: Average time (in minutes) for different methods over 5 simulations. Scenario: a total of 120 potential IVs with 80 valid; directional pleiotropy. Simulations are performed on a multi-core server. The hardware is 20 cores @ 2.67Ghz with 40GB RAM and the operating System is CentOS Linux release 7. n_1 is the number of observations used in selection and the remaining is used for estimation.

Method	entire data	$n_1 = 10k$	$n_1 = 30k$	$n_1 = 50k$
MR-Lasso	9.76	4.69	4.96	5.34
Lasso	21.09	5.75	7.94	10.46
sisVIVE	26.31	6.43	9.73	13.21

Regarding computing time (Table 4.7 and Table C.5), MR-Lasso is the quickest as it is based on summary statistics, while sisVIVE is the slowest. Our new Lasso method is faster than sisVIVE, and threshold-based methods are less time-consuming than Lasso and sisVIVE. When sample splitting is used, the difference in computing time between methods becomes smaller.

In conclusion, reducing the sample size for selection would make it more challenging to detect invalid IVs and lead to decreased coverage. Therefore, solely employing sample splitting to account for instrument selection is insufficient. The proposed bootstrap method mimics the selection process by repeating the selection procedure for each bootstrap sample. When the proportion of invalid IVs is less than half, the coverage of bootstrap intervals is close to the nominal level. Overall, shrinkage-based methods outperform threshold-based methods in terms of coverage probability. For the three shrinkage methods, sample splitting is unnecessary as their coverage is already satisfactory, and the intervals are narrower compared to those with sample splitting. Nevertheless, sample splitting offers the advantage of requiring less computational time than using the complete dataset.

4.4.4 Summary

This chapter has addressed the challenges in coverage estimation when disregarding the selection process for instrumental variables and proposed a novel approach using the Lasso method to identify invalid IVs. However, both the proposed and existing methods for identifying invalid IVs are not flawless, even when dealing with a small proportion of invalid variants. In cases where the correct model was not chosen by the selection method, the coverage could never attain the nominal level without considering the selection process. Moreover, the impact of selection on coverage rates becomes more pronounced when the proportion of invalid IVs exceeds 50%.

The coverage of Wald-type intervals can be enhanced by using 50k observations for selection, but it may worsen with a relatively small sample size for selection, as this comes at the expense of power due to the loss in sample size. On the other hand, bootstrap methods have shown improvement in coverage compared to Wald-type intervals. The key concept behind using bootstrap is to better estimate the standard error, as original standard error estimation fails to consider the selection process. By mimicking the selection process, the proposed bootstrap method accounts for selection variability, leading to bootstrap standard errors that are closer to the standard deviation than the original standard errors (Tables 4.2 and 4.6). The coverage of bootstrap intervals is close to the nominal level when the proportion of invalid IVs is less than or equal to 50%, and it remains around 88%-93% when the proportion of invalid IVs exceeds 50%.

Among our proposed selection methods, Lasso demonstrates superior performance compared to threshold-based methods in terms of coverage accuracy. Additionally, when compared to existing methods, Lasso is less time-consuming than sisVIVE and outperforms MR-LASSO in terms of true negative counts.

We believe that our selection and bootstrap methods can be useful in selecting valid IVs and correcting the selection effect in inference. The R code for the proposed method is available in `https://github.com/Bianmj/InvalidIVs`.

Chapter 5

Discussion

5.1 Summary

In this thesis, we delve into three important challenges within Mendelian randomization that can introduce bias into causal inference. Initially, we introduce innovative approaches to address the “winner’s curse” problem in Mendelian randomization. This issue arises when the same dataset is used both for selecting IVs and estimating the associations between IVs and exposure. Secondly, we address the issue of overlapping samples between exposure and outcome. Thirdly, we put forth a methodology for identifying invalid IVs in the presence of horizontal pleiotropy, and we also incorporate the bootstrap method to account for this selection of IV procedure. Each of these proposed methods has its own set of limitations and strengths. In this concluding chapter, we will highlight the key components of each chapter and subsequently outline potential avenues for future research, building upon the contributions made in this work.

This thesis begins by introducing the various key concepts related to causal inference, genetics, and the application of Mendelian randomization in estimating causal effects in the presence of unmeasured confounding, which cannot be addressed by observational studies, highlighting its advantages over randomized controlled trials. We also provide a concise overview of robust methods for MR analysis, with a particular focus on two-sample MR when Assumption IV3 (no horizontal pleiotropy) is not perfectly met. Furthermore, the initial chapter introduces the three-sample GWAS design, which proves to be valuable in mitigating bias arising from complete overlapping between discovery and estimation GWAS studies. It lays the groundwork and establishes the motivation for the development of causal inference approaches, taking into account specific challenges such as the winner's curse, overlapping between exposure and outcome association, and the presence of pleiotropy effects.

We initially focus on tackling the issue of the winner's curse, a phenomenon that distorts exposure association estimates and subsequently affects the estimation of causal effect. To address this, we modify the original BR-squared approach, which was originally tailored for individual-level data, to make it suitable for summary-level data. Additionally, we customize Ghosh's method to function with ranking-based selection, wherein the significance threshold is established based on observed test statistics among the top-ranked variants. Through simulations and real data examples, we thoroughly evaluate the impact of the winner's curse on MR estimation. We find that the winner's curse introduces a bias towards the null in the causal estimator, although it does not significantly affect the overall conclusion, as observed in both simulated and real data applications. Applying the correction significantly reduces this bias, albeit at the cost of increased variability in the estimates. As a result, the confidence intervals become wider, but the coverage of the intervals is improved. Moreover, we note that Ghosh's methods tend to produce more variable outcomes compared to other methods. This is because the estimate is excessively corrected for variants with z-values close to the boundary, while minimal correction is applied to variants with z-values far from the boundary. Consequently, when a considerable proportion of significant variants undergo heavy over-correction, the Ghosh's method is at risk of overestimating the causal effect, while it closely resembles the naive estimate when the percentage of such variants is not substantial. Our approach would be beneficial for researchers seeking a more precise point estimate. However, if the primary goal is to examine the existence of a causal effect, the correction method may not provide significant advantages, as its impact on statistical power is minimal.

Next, we deal with the challenge of overlapping in two estimation studies for the genetic association with exposure and outcome, particularly in the presence of weak instruments. To tackle these issues, we propose a novel method inspired by the work of Bowden *et al.* (2019). Our results from various simulation settings and real data applications align with existing research on analysis with overlapping samples, which indicate that overlap can bias the causal estimator towards the null with non-overlapping samples and towards the confounded association with complete overlapping. Our proposed method allows us to obtain an unbiased estimator with varying degrees of overlap or strength of IVs. Type-I error may be inflated, it is much better other methods when there is substantial overlap between samples. Another strength of our proposed method is its ease of implementation, as it only requires summary statistics, and the phenotypic correlation between exposure and outcome can be calculated using cross-trait LD score regression. Nevertheless, our approach comes with certain drawbacks. For instance, it tends to underestimate the standard error because it incorporates the covariance term in variability estimation, leading to intervals with lower coverage. Therefore, in certain cases, such as when the IVs are not very weak or the degree of overlap is low, the IVW method may provide better coverage than our proposed method. As a result, we recommend that researchers evaluate the strength of IVs and the degree of overlap (by calculating the phenotype correlation) before selecting the appropriate method for MR analysis.

In this thesis, we address overlap and weak instrument bias as distinct challenges from the winner's curse. Certain studies have examined how the interplay between the winner's curse, overlap bias, and weak instrument bias impacts the estimation of causal effect. These investigations have demonstrated that the winner's curse substantially amplifies the degree of weak instrument bias (Sadreev *et al.*, 2021; Mounier and Kutalik, 2023). The extent and direction of this bias depend on factors such as the degree of sample overlap, the causal effect, and confounding variables, which aligns with our own findings. When the confounder and the causal effect share the same sign, the causal effect is overestimated in cases of complete sample overlap and underestimated in scenarios of non-overlapping samples. However, when their signs are opposite, the causal effect is consistently underestimated across all levels of overlap (Mounier and Kutalik, 2023).

Apart from the winner's curse, overlap, and weak instrument bias, another notable concern in MR stems from the bias resulting from horizontal pleiotropy. We introduce a Lasso method to identify and subsequently remove these invalid IVs that are susceptible to horizontal pleiotropy in our analysis. Additionally, we take the selection of IVs process into account by utilizing the bootstrap method, a step that has been overlooked by existing robust methods that involve the removal of invalid IVs in MR. Through simulations using both the sample-splitting and bootstrap approaches, we observe that the coverage of conventional Wald-type intervals is underestimated, especially as the proportion of invalid IVs increases. Surprisingly, the sample-splitting approach does not offer any advantage over the naive method when the size for selection is insufficient, while the bootstrap method significantly improves the coverage compared to the naive method. Based on our findings, we recommend researchers refrain from using traditional inference when some invalid IVs are removed after conducting a sensitivity analysis.

5.2 Limitations and Future Work

In this section, our main emphasis will be on the limitations of the methods proposed in Chapters 2-4. We will then explore potential future research directions to enhance these methods for more robust applications.

5.2.1 Linkage Disequilibrium

The methods discussed in this thesis all assume that the genetic variants used as IVs are uncorrelated (linkage equilibrium). This assumption is made to ensure the validity of the MR framework. Despite this concern, there are some scenarios in which including genetic variants in linkage disequilibrium (LD) could be advantageous for a Mendelian randomization study. One such scenario is in *cis*-MR, where variants in close proximity to the gene of interest are used as instrumental variables. In *cis*-MR, the focus is on studying the associations of variants in the region with a specific disease outcome, which provides insights into whether the encoded protein can be a potential drug target for the outcome. Since *cis*-MR studies typically rely on a single gene region, researchers have to select genetic instruments from a pool of potentially correlated variants. Additionally, including correlated genetic variants can improve the statistical power of the study, especially when only a limited number of SNPs remain after sensitivity analysis and exclusion procedures.

An intriguing direction for further research would involve evaluating the impact of correlated genetic variants on the performance of existing Winner's curse correction methods. Investigating how these methods handle the presence of correlated SNPs and understanding their limitations and strengths in this context would be valuable. Additionally, it may be worth considering an extension of the BR-squared method that takes into account the genetic correlation between variants. It is also might be interesting to explore the impact of varying LD thresholds on causal estimation.

5.2.2 Case-control Studies

Our approach has inherent limitations as it exclusively focuses on quantitative traits. In Chapter 3, to investigate the overlap between a quantitative study and a case-control study, we may need to consider the specific sources of overlap: whether they arise from the case samples, the control samples, or both.

In a previous study by Burgess *et al.* (2016), they observed that in the absence of a causal effect, no bias was detected when estimating the association of G-X solely in the control group and G-Y association in all participants. However, when estimating the G-X association across all participants, the relative bias closely approached the inverse of the mean F-statistics. Potential future research could explore scenarios involving the existence of a causal effect and varying degrees of overlap. Furthermore, when it comes to estimating the correlation between quantitative and case-control studies, the method employed will differ from that used for two quantitative traits. In such cases, a point biserial correlation coefficient could serve as a suitable correlation coefficient.

5.2.3 A Larger Sample Size

Our simulation study in Chapter 3 has limitations related to the sample size used in the data-generating model. Given that GWAS are increasingly conducted on large study populations, the choice of a sample size of 10,000 for our simulation study may be somewhat conservative. This reduction in sample size compared to real data is driven by computational considerations, as we aim to simulate individual-level data to control the degree of overlap, which requires a substantial amount of memory compared to working with summary statistics alone. Despite these challenges, we make efforts to keep the F-statistic and R^2 values as close to reality as possible, even with a smaller sample size and fewer genetic variants, to maintain practicality. However, it would be more practical and insightful if we could base our simulations on a larger sample size in the future. Furthermore, exploring the development of a simulation design that exclusively utilizes summary statistics would eliminate computational difficulties, making it an interesting avenue for future research.

5.2.4 Bootstrap for Summary Statistics

In Chapter 4, the bootstrap procedure for addressing selection bias is developed using individual-level data. However, it is common for MR analyses to rely on summary-level data. While there are robust methods available for conducting MR studies with summary-level data, there are currently few methods that specifically account for instrument selection in this context. Hence, future research focused on extending the bootstrap method to incorporate summary statistics would be both promising and valuable.

5.2.5 Selection Bias

Throughout this thesis, we do not consider the selection bias, which occurs when the selection of participants depends on the exposure or the outcome. Then conditioning on selection induces an association between the genetic variant and confounder in both cases. For instance, selection bias in MR studies can arise when the original outcome GWAS is chosen based on survival until recruitment (Yang *et al.*, 2021). This happens due to the time lag between genetic randomization and GWAS recruitment, potentially causing the bias in MR estimates. Simulation studies have shown that when selection into a study is strongly influenced by the exposure, selection bias can markedly impact the accuracy of causal effect estimates (Gkatzionis and Burgess, 2019).

Appendix A

Supplementary material of Chapter 2

A.1 Derivation of BR-Squared by summary statistics

Assume G_i and x_i are all standardized to have mean 0 and variance 1, then the within bootstrap estimate and out-of-sample bootstrap estimate can be written as below:

$$\beta_D^* = \frac{\sum f_i G_i x_i}{\sum f_i G_i^2} \tag{A.1a}$$

$$\beta_E^* = \frac{\sum I_i G_i x_i}{\sum I_i G_i^2} \tag{A.1b}$$

where f_i represents the frequency for (G_i, x_i) of observation i in the bootstrap sample, I_i as the indicator function which takes value of 1 when f_i equals 0 and n is the sample size.

Then

$$\begin{aligned}
\operatorname{cov}(\beta_D^*, \beta_E^*) &= \operatorname{cov}\left(\frac{\sum f_i G_i x_i}{\sum f_i G_i^2}, \frac{\sum I_i G_i x_i}{\sum I_i G_i^2}\right) \\
&= \mathrm{E}\left(\frac{\sum f_i G_i x_i}{\sum f_i G_i^2} \times \frac{\sum I_i G_i x_i}{\sum I_i G_i^2}\right) - \mathrm{E}\left(\frac{\sum f_i G_i x_i}{\sum f_i G_i^2}\right)\mathrm{E}\left(\frac{\sum I_i G_i x_i}{\sum I_i G_i^2}\right) \\
&= \mathrm{E}\left(\frac{\sum_i \sum_j f_i G_i x_i I_j G_j x_j}{\sum_i \sum_j f_i G_i^2 I_j G_j^2}\right) - \mathrm{E}\left(\frac{\sum f_i G_i x_i}{\sum f_i G_i^2}\right)\mathrm{E}\left(\frac{\sum I_i G_i x_i}{\sum I_i G_i^2}\right) \qquad \text{(A.2)}
\end{aligned}$$

According to the first-order Taylor expansion, the second terms in equation A.2 can be simplified as

$$\mathrm{E}\left(\frac{\sum f_i G_i x_i}{\sum f_i G_i^2}\right) \approx \frac{\mathrm{E}(\sum f_i G_i x_i)}{E(\sum f_i G_i^2)} = \frac{\sum G_i x_i}{\sum G_i^2} \qquad \text{(A.3a)}$$

$$\mathrm{E}\left(\frac{\sum I_i G_i x_i}{\sum I_i G_i^2}\right) \approx \frac{E(\sum I_i G_i x_i)}{E(\sum I_i G_i^2)} = \frac{\sum_i^n G_i x_i}{\sum_i^n G_i^2} \qquad \text{(A.3b)}$$

where $\mathrm{E}(f_i) = n \cdot \frac{1}{n} = 1$

For the first term in Equation A.2 $\left(E\left(\frac{\sum_i \sum_j f_i G_i x_i I_j G_j x_j}{\sum_i \sum_j f_i G_i^2 I_j G_j^2}\right)\right)$

$\mathrm{E}(f_i I_j) = 0$ when $i = j$; when $i \neq j$:

$$\begin{aligned}
\mathrm{E}(f_i I_j) &= \mathrm{E}(f_i I_j | f_j = 0)\ p(f_j = 0) + \mathrm{E}(f_i I_j | f_j \neq 0)\ p(f_j \neq 0) \\
&= \mathrm{E}(f_i | f_j = 0)\ p(f_j = 0) + 0 \\
&= \frac{n}{n-1} * m \qquad \text{(A.4)}
\end{aligned}$$

where $f_i | f_j = 0 \sim Bin\left(n - 0 = n, \frac{1/n}{1-1/n} = \frac{1}{n-1}\right)$, $m = p(f_j = 0) \approx e^{-1} = 0.368$

Therefore, the numerator for $\mathrm{E}\left(\frac{\sum_i \sum_j f_i G_i x_i I_j G_j x_j}{\sum_i \sum_j f_i G_i^2 I_j G_j^2}\right)$ is

$$\begin{aligned}\mathrm{E}\left(\sum_i \sum_j f_i G_i x_i I_j G_j x_j\right) &= \frac{nm}{n-1}\sum_{i\neq j}\sum_{i\neq j}(G_i x_i)(G_j x_j)\\ &= \frac{nm}{n-1}\sum_i\sum_j (G_i x_i)(G_j x_j) - \frac{nm}{n-1}\sum_i (G_i x_i)^2\\ &= \frac{nm}{n-1}\left(\sum_i (G_i x_i)\right)^2 - \frac{nm}{n-1}\sum_i (G_i x_i)^2 \qquad \text{(A.5)}\end{aligned}$$

The denominator is $\mathrm{E}\left(\sum_i \sum_j f_i G_i^2 I_j G_j^2\right) = \frac{nm}{n-1}\left(\sum\limits_i G_i^2\right)^2 - \frac{nm}{n-1}\sum\limits_i G_i^4$

$\Longrightarrow$

$$\mathrm{cov}(\beta_D^*, \beta_E^*) \approx \frac{(\sum G_i x_i)^2 - \sum (G_i x_i)^2}{(\sum G_i^2)^2 - \sum G_i^4} - \frac{(\sum G_i x_i)^2}{(\sum G_i^2)^2}$$

since $(\sum\limits_i G_i^2)^2 \approx 3n + n(n-1) \propto n^2$ and $\sum\limits_i G_i^4 \approx 3n$, when n is big enough, $(\sum\limits_i G_i^2)^2 \gg \sum\limits_i G_i^4$.

$\Longrightarrow$

$$\mathrm{cov}(\beta_D^*, \beta_E^*) \approx \frac{(\sum G_i x_i)^2 - \sum (G_i x_i)^2 - (\sum G_i x_i)^2}{(\sum G_i^2)^2} = \frac{-\sum (G_i x_i)^2}{(\sum G_i^2)^2} \qquad \text{(A.6)}$$

And the numerator for Equation A.6 can be simplified as

$$\begin{aligned}\sum (G_i x_i)^2 \approx \sum G_i^2 (G_i^2 \hat{\beta}_N^2 + \varepsilon_i^2) &= \sum G_i^4 \hat{\beta}_N^2 + \sum G_i^2 \varepsilon_i^2\\ &\approx 3n\hat{\beta}_N^2 + n(1-R^2)\\ &\approx 3n\hat{\beta}_N^2 + n(1-\hat{\beta}_N^2) \qquad \text{(A.7)}\end{aligned}$$

where $\hat{\beta}_N$ is the naive estimate for the effect of G on x, and R^2 is the coefficient of determination from the regression of x on G. Therefore, the covariance between β_D^*

and β_E^* is eventually simplified as

$$\mathrm{cov}(\beta_D^*, \beta_E^*) \approx \frac{-(1+2\hat{\beta}_N^2)}{n} \tag{A.8}$$

The variances for the within-bootstrap estimates and out-of-sample bootstrap estimates are proportional to the sample size:

$$\sigma_D^2 = \frac{1-R^2}{n} \approx \frac{1-\hat{\beta}_N^2}{n} \tag{A.9a}$$

$$\sigma_E^2 = \frac{1-R^2}{mn} \approx \frac{1-\hat{\beta}_N^2}{mn} \tag{A.9b}$$

Proof of $R^2 = \hat{\beta}_N^2$:

The slope of a simple linear regression without intercept term can be written as:

$$\begin{aligned}
\hat{\beta}_N = \frac{\sum G_i x_i}{\sum\limits_i G_i^2} &= \frac{\sum(G_i-0)(x-0)}{\sum(G_i-0)^2} \\
&= \frac{\sum(G_i-\overline{G})(x_i-\overline{x})}{\sum(G_i-\overline{G})^2} \\
&= \frac{\sum(G_i-\overline{G})(x_i-\overline{x})/n}{\sum(G_i-\overline{G})^2/n} \\
&= \frac{s_{G,x}}{s_G^2} \\
&= r_{G,x}\frac{s_x}{s_G} \qquad \text{(A.10)}
\end{aligned}$$

where $r_{G,x}$ is the sample correlation coefficient between G and x, s_x and s_G are the uncorrected sample standard deviations for x and G,. s_G^2 and $s_{G,x}$ are the sample variance and sample covariance. Because s_G and s_x are all equal to 1, we can get

$$\hat{\beta}_N^2 = R^2$$

A.2 Supplementary tables

Table A.1: Mean, relative bias, sqaure root of mean of variance (SE), standard deviation (SD), coverage probability, and mean of interval length, mean squared error (MSE) by naive method, projack method, BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of normalized likelihood estimator (Normalized), and Compromise estimator (Compromise). We repeated 250 simulations for threshould-based selection ($1 \cdot 10^{-1}$). True causal effect=0.2.

Method	Mean	Relative bias	SE	SD	MSE	CI length	Coverage	Power
threshold for p-value is $1 \cdot 10^{-1}$								
Naive	0.177	-0.117	0.032	0.032	0.0017	0.124	0.888	1
Projack	0.196	-0.020	0.036	0.037	0.0014	0.138	0.948	1
BR-squared	0.196	-0.022	0.035	0.037	0.0013	0.138	0.948	1
FIQT	0.194	-0.030	0.035	0.036	0.0013	0.136	0.94	1
Forde	0.194	-0.029	0.034	0.037	0.0014	0.136	0.944	1
MLE	0.183	-0.086	0.033	0.035	0.0015	0.130	0.904	1
Normalized	0.184	-0.081	0.033	0.037	0.0016	0.130	0.90	1
Compromise	0.183	-0.085	0.033	0.037	0.0016	0.130	0.90	1

Table A.2: Mean, relative bias, sqaure root of mean of variance (SE), standard deviation (SD), coverage probability, and mean of interval length, mean squared error (MSE) by naive method, projack method, BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of normalized likelihood estimator (Normalized), and Compromise estimator (Compromise). We repeated 250 simulations for threshould-based selection ($1 \cdot 10^{-6}$). True causal effect=0.2.

Method	Mean	Relative bias	SE	SD	MSE	CI length	Coverage	Power
		threshold for p-value is $1 \cdot 10^{-6}$						
Naive	0.179	-0.105	0.046	0.048	0.0027	0.174	0.912	0.948
Projack	0.197	-0.017	0.051	0.053	0.0029	0.191	0.96	0.948
BR-squared	0.197	-0.012	0.051	0.054	0.0029	0.191	0.96	0.948
FIQT	0.194	-0.033	0.050	0.052	0.0027	0.187	0.948	0.948
Forde	0.196	-0.018	0.051	0.053	0.0028	0.190	0.96	0.948
MLE	0.194	-0.028	0.054	0.057	0.0033	0.195	0.936	0.924
Normalized	0.205	0.027	0.055	0.059	0.0035	0.199	0.948	0.944
Compromise	0.201	0.0047	0.055	0.058	0.0034	0.198	0.936	0.94

Table A.3: Mean, relative bias, sqaure root of mean of variance (SE), standard deviation (SD), coverage probability, and mean of interval length, mean squared error (MSE) by naive method, projack method, BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of normalized likelihood estimator (Normalized), and Compromise estimator (Compromise). We repeated 250 simulations for rank-based selection (top 70 variants). True causal effect=0.2.

Method	Mean	Relative bias	SE	SD	MSE	CI length	Coverage	Power
			top 70 variants					
Naive	0.180	-0.0993	0.032	0.034	0.0016	0.123	0.876	1
Projack	0.200	0.0014	0.036	0.039	0.0015	0.137	0.940	1
BR-squared	0.200	-0.0001	0.035	0.039	0.0015	0.136	0.944	1
FIQT	0.198	-0.0090	0.035	0.038	0.0015	0.135	0.932	1
Forde	0.198	-0.0119	0.035	0.037	0.0015	0.134	0.920	1
MLE	0.185	-0.073	0.033	0.035	0.0015	0.127	0.900	1
Normalized	0.186	-0.068	0.033	0.036	0.0015	0.127	0.896	1
Compromise	0.186	-0.070	0.033	0.036	0.0015	0.127	0.896	1

Table A.4: Mean, relative bias, sqaure root of mean of variance (SE), standard deviation (SD), coverage probability, and mean of interval length, mean squared error (MSE) by naive method, projack method, BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of normalized likelihood estimator (Normalized), and Compromise estimator (Compromise). We repeated 250 simulations for rank-based selection (top 10 variants). True causal effect=0.2.

Method	Mean	Relative bias	SE	SD	MSE	CI length	Coverage	Power
			top 10 variants					
Naive	0.178	-0.109	0.048	0.048	0.0028	0.187	0.928	0.960
Projack	0.197	-0.015	0.054	0.054	0.0029	0.207	0.968	0.960
BR-squared	0.197	-0.013	0.053	0.054	0.0029	0.207	0.968	0.968
FIQT	0.192	-0.038	0.052	0.052	0.0027	0.202	0.972	0.960
Forde	0.195	-0.026	0.053	0.052	0.0027	0.203	0.964	0.964
MLE	0.196	-0.018	0.059	0.062	0.0038	0.225	0.948	0.912
Normalized	0.212	0.062	0.060	0.063	0.0042	0.230	0.948	0.936
Compromise	0.207	0.031	0.060	0.063	0.0040	0.228	0.936	0.928

Table A.5: Type-I errors for naive method, projack method, BR-squared, FIQT, Forde, conditional MLE (MLE), the mean of normalized likelihood estimator (Normalized), and Compromise estimator (Compromise). We repeated 250 simulations for threshould-based selection ($5 \cdot 10^{-4}$).

Method	Type-I error
Naive	0.056
Projack	0.056
BR-squared	0.056
FIQT	0.056
Forde	0.056
MLE	0.056
Normalized	0.056
Compromise	0.056

Appendix B

Supplementary material of Chapter 3

Table B.1: Mean of causal estimate and coverage for 1000 simulations. Scenario: positive causal effect ($\beta_c = 0.1$) and weak confounder ($\beta_u = \alpha_u = 0.4$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	Mean	Coverage	Mean	Coverage	Mean	Coverage	Mean	Coverage
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 0%								
$5 \cdot 10^{-2} (F = 12.45)$	0.091	0.936	0.084	0.940	0.100	0.910	0.100	0.910
$5 \cdot 10^{-5} (F = 33.89)$	0.096	0.950	0.091	0.963	0.099	0.947	0.099	0.947
$5 \cdot 10^{-8} (F = 63.92)$	0.099	0.948	0.096	0.959	0.101	0.945	0.101	0.945
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 25%								
$5 \cdot 10^{-2} (F = 12.45)$	0.096	0.947	0.093	0.962	0.106	0.919	0.101	0.918
$5 \cdot 10^{-5} (F = 33.89)$	0.099	0.950	0.097	0.962	0.102	0.944	0.100	0.943
$5 \cdot 10^{-8} (F = 63.92)$	0.100	0.949	0.099	0.959	0.102	0.948	0.101	0.947
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 50%								
$5 \cdot 10^{-2} (F = 12.45)$	0.100	0.955	0.101	0.971	0.110	0.925	0.099	0.937
$5 \cdot 10^{-5} (F = 33.89)$	0.099	0.942	0.100	0.954	0.102	0.931	0.099	0.929
$5 \cdot 10^{-8} (F = 63.92)$	0.100	0.946	0.100	0.954	0.102	0.946	0.100	0.944
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 75%								
$5 \cdot 10^{-2} (F = 12.45)4$	0.105	0.950	0.111	0.962	0.116	0.901	0.099	0.925
$5 \cdot 10^{-5} (F = 33.89)$	0.102	0.938	0.105	0.946	0.105	0.931	0.099	0.931
$5 \cdot 10^{-8} (F = 63.92)$	0.102	0.940	0.104	0.944	0.104	0.940	0.101	0.937
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 100%								
$5 \cdot 10^{-2} (F = 12.45)$	0.112	0.945	0.121	0.914	0.123	0.873	0.100	0.936
$5 \cdot 10^{-5} (F = 33.89)$	0.103	0.946	0.110	0.947	0.107	0.937	0.099	0.936
$5 \cdot 10^{-8} (F = 63.92)$	0.102	0.947	0.106	0.948	0.104	0.947	0.100	0.943

Table B.2: Mean of standard errors (SE) and standard deviation (SD) for 1000 simulations. Scenario: positive causal effect ($\beta_c = 0.1$) and weak confounder ($\beta_u = \alpha_u = 0.4$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	SE	SD	SE	SD	SE	SD	SE	SD
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 0%								
$5 * 10^{-2}(F = 12.45)$	0.029	0.030	0.030	0.028	0.029	0.033	0.029	0.033
$5 * 10^{-5}(F = 33.89)$	0.036	0.036	0.037	0.034	0.036	0.037	0.036	0.037
$5 * 10^{-8}(F = 63.92)$	0.048	0.049	0.049	0.048	0.048	0.050	0.048	0.050
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 25%								
$5 * 10^{-2}(F = 12.45)$	0.029	0.030	0.030	0.028	0.029	0.033	0.029	0.033
$5 * 10^{-5}(F = 33.89)$	0.036	0.036	0.037	0.034	0.036	0.037	0.036	0.037
$5 * 10^{-8}(F = 63.92)$	0.048	0.048	0.048	0.046	0.048	0.049	0.048	0.049
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 50%								
$5 * 10^{-2}(F = 12.45)$	0.029	0.029	0.030	0.027	0.029	0.032	0.029	0.032
$5 * 10^{-5}(F = 33.89)$	0.036	0.036	0.036	0.034	0.036	0.037	0.036	0.037
$5 * 10^{-8}(F = 63.92)$	0.048	0.048	0.048	0.047	0.048	0.049	0.048	0.049
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 75%								
$5 * 10^{-2}(F = 12.45)$	0.029	0.029	0.030	0.027	0.030	0.032	0.029	0.032
$5 * 10^{-5}(F = 33.89)$	0.036	0.036	0.036	0.034	0.036	0.037	0.036	0.038
$5 * 10^{-8}(F = 63.92)$	0.048	0.048	0.048	0.047	0.048	0.049	0.047	0.049
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.4$, overlap = 100%								
$5 * 10^{-2}(F = 12.45)$	0.029	0.028	0.029	0.027	0.030	0.031	0.029	0.031
$5 * 10^{-5}(F = 33.89)$	0.036	0.036	0.036	0.035	0.036	0.037	0.035	0.037
$5 * 10^{-8}(F = 63.92)$	0.048	0.049	0.048	0.048	0.048	0.050	0.047	0.050

Table B.3: Mean of causal estimate and coverage for 1000 simulations. Scenario: positive causal effect ($\beta_c = 0.1$) and strong confounder ($\beta_u = \alpha_u = 0.8$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	Mean	Coverage	Mean	Coverage	Mean	Coverage	Mean	Coverage
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 0%								
$5 \cdot 10^{-2}(F = 9.11)$	0.088	0.929	0.079	0.939	0.100	0.901	0.100	0.901
$5 \cdot 10^{-5}(F = 24.25)$	0.095	0.953	0.089	0.964	0.100	0.947	0.100	0.947
$5 \cdot 10^{-8}(F = 45.51)$	0.099	0.943	0.095	0.957	0.101	0.942	0.101	0.942
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 25%								
$5 \cdot 10^{-2}(F = 9.11)$	0.102	0.945	0.103	0.969	0.116	0.886	0.100	0.902
$5 \cdot 10^{-5}(F = 24.25)$	0.101	0.957	0.102	0.968	0.106	0.952	0.100	0.941
$5 \cdot 10^{-8}(F = 45.51)$	0.101	0.961	0.102	0.967	0.104	0.957	0.101	0.956
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 50%								
$5 \cdot 10^{-2}(F = 9.11)$	0.115	0.935	0.126	0.923	0.130	0.854	0.098	0.924
$5 \cdot 10^{-5}(F = 24.25)$	0.105	0.942	0.113	0.944	0.109	0.933	0.098	0.931
$5 \cdot 10^{-8}(F = 45.51)$	0.102	0.949	0.108	0.955	0.105	0.944	0.099	0.938
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 75%								
$5 \cdot 10^{-2}(F = 9.11)$	0.129	0.862	0.151	0.714	0.146	0.746	0.099	0.917
$5 \cdot 10^{-5}(F = 24.25)$	0.111	0.944	0.126	0.919	0.115	0.933	0.099	0.936
$5 \cdot 10^{-8}(F = 45.51)$	0.106	0.949	0.116	0.940	0.108	0.945	0.100	0.938
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 100%								
$5 \cdot 10^{-2}(F = 9.11)$	0.145	0.764	0.177	0.355	0.162	0.576	0.101	0.930
$5 \cdot 10^{-5}(F = 24.25)$	0.115	0.939	0.139	0.841	0.120	0.924	0.098	0.934
$5 \cdot 10^{-8}(F = 45.51)$	0.108	0.951	0.123	0.933	0.111	0.948	0.099	0.943

Table B.4: Mean of standard errors (SE) and standard deviation (SD) for 1000 simulations. Scenario: positive causal effect ($\beta_c = 0.1$) and strong confounder ($\beta_u = \alpha_u = 0.8$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	SE	SD	SE	SD	SE	SD	SE	SD
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 0%								
$5 \cdot 10^{-2}(F = 9.11)$	0.035	0.036	0.036	0.033	0.035	0.041	0.035	0.041
$5 \cdot 10^{-5}(F = 24.25)$	0.044	0.042	0.045	0.040	0.044	0.044	0.044	0.044
$5 \cdot 10^{-8}(F = 45.51)$	0.058	0.060	0.059	0.058	0.059	0.062	0.059	0.062
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 25%								
$5 \cdot 10^{-2}(F = 9.11)$	0.035	0.036	0.036	0.033	0.035	0.041	0.035	0.041
$5 \cdot 10^{-5}(F = 24.25)$	0.044	0.043	0.044	0.040	0.044	0.045	0.043	0.045
$5 \cdot 10^{-8}(F = 45.51)$	0.058	0.058 2	0.059	0.056	0.059	0.060	0.058	0.060
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 50%								
$5 \cdot 10^{-2}(F = 9.11)$	0.035	0.034	0.036	0.032	0.035	0.040	0.034	0.039
$5 \cdot 10^{-5}(F = 24.25)$	0.044	0.043	0.044	0.040	0.044	0.045	0.043	0.045
$5 \cdot 10^{-8}(F = 45.51)$	0.058	0.057	0.058	0.056	0.059	0.059	0.057	0.059
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 75%								
$5 \cdot 10^{-2}(F = 9.11)$	0.035	0.035	0.035	0.033	0.035	0.040	0.034	0.039
$5 \cdot 10^{-5}(F = 24.25)$	0.044	0.043	0.043	0.041	0.044	0.045	0.042	0.045
$5 \cdot 10^{-8}(F = 45.51)$	0.058	0.057	0.057	0.055	0.059	0.059	0.057	0.059
$\beta_c = 0.1$, $\beta_u = \alpha_u = 0.8$, overlap = 100%								
$5 \cdot 10^{-2}(F = 9.11)$	0.035	0.033	0.034	0.033	0.036	0.038	0.034	0.037
$5 \cdot 10^{-5}(F = 24.25)$	0.044	0.043	0.042	0.041	0.044	0.045	0.042	0.045
$5 \cdot 10^{-8}(F = 45.51)$	0.058	0.058	0.056	0.057	0.059	0.059	0.056	0.060

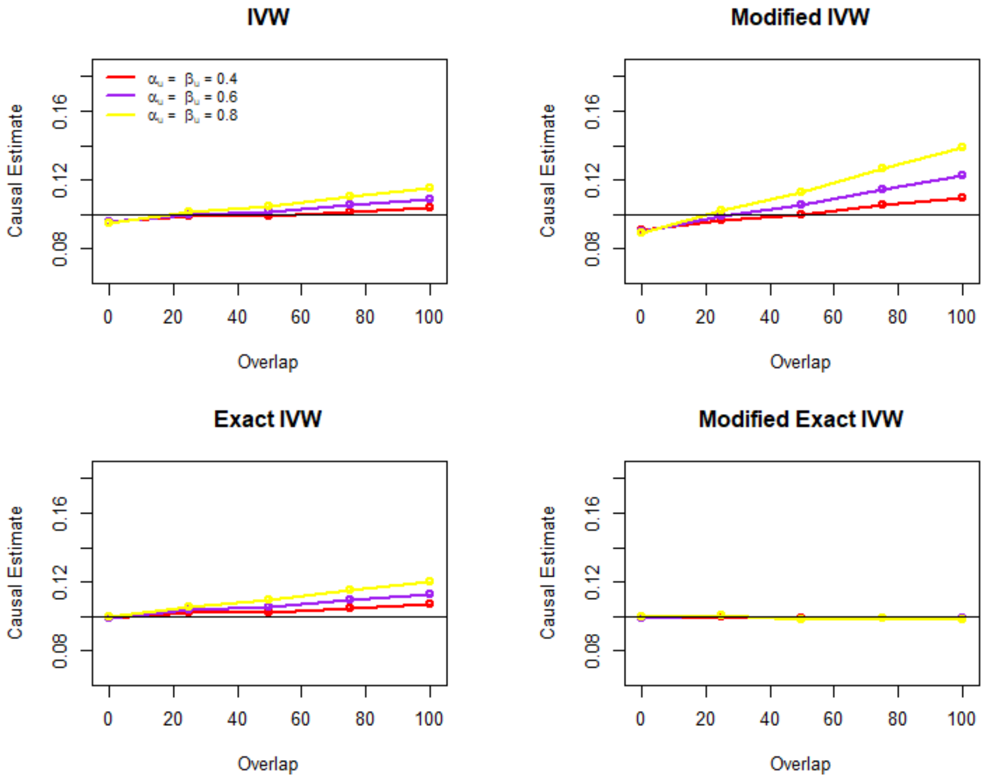


Figure B.1: Causal estimates with varing effect of confounder. Scenario: Significant threshold $5 \cdot 10^{-5}$; causal effect=0.1 (horizontal line). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

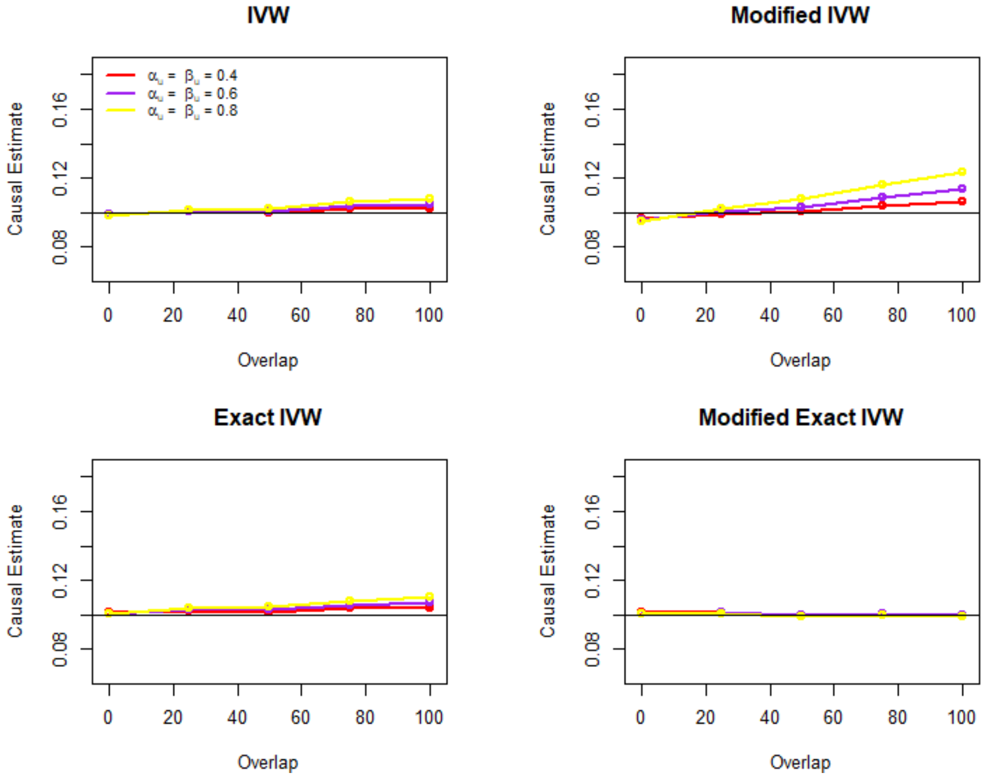


Figure B.2: Causal estimates with varing effect of confounder. Scenario: Significant threshold $5 \cdot 10^{-8}$; causal effect=0.1 (horizontal line). Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Table B.5: Mean of causal estimate and coverage for 1000 simulations. Scenario: positive causal effect ($\beta_c = -0.3$) and moderate confounder ($\beta_u = \alpha_u = -0.6$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	Mean	Coverage	Mean	Coverage	Mean	Coverage	Mean	Coverage
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 0%								
$5 \cdot 10^{-2} (F = 8.74)$	-0.260	0.916	-0.234	0.879	-0.298	0.927	-0.298	0.927
$5 \cdot 10^{-5} (F = 13.45)$	-0.277	0.940	-0.253	0.944	-0.303	0.931	-0.303	0.931
$5 \cdot 10^{-8} (F = 21.11)$	-0.281	0.925	-0.261	0.928	-0.297	0.927	-0.297	0.927
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 25%								
$5 \cdot 10^{-2} (F = 8.74)$	-0.261	0.926	-0.236	0.896	-0.300	0.927	-0.298	0.926
$5 \cdot 10^{-5} (F = 13.45)$	-0.280	0.938	-0.255	0.935	-0.305	0.936	-0.304	0.936
$5 \cdot 10^{-8} (F = 21.11)$	-0.282	0.942	-0.262	0.952	-0.298	0.941	-0.298	0.941
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 50%								
$5 \cdot 10^{-2} (F = 8.74)$	-0.266	0.911	-0.241	0.896	-0.305	0.925	-0.302	0.925
$5 \cdot 10^{-5} (F = 13.45)$	-0.280	0.932	-0.256	0.930	-0.305	0.927	-0.304	0.924
$5 \cdot 10^{-8} (F = 21.11)$	-0.287	0.942	-0.267	0.952	-0.304	0.942	-0.303	0.938
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 75%								
$5 \cdot 10^{-2} (F = 8.74)$	-0.266	0.918	-0.243	0.889	-0.305	0.921	-0.301	0.921
$5 \cdot 10^{-5} (F = 13.45)$	-0.280	0.929	-0.257	0.925	-0.305	0.928	-0.303	0.925
$5 \cdot 10^{-8} (F = 21.11)$	-0.288	0.954	-0.269	0.957	-0.305	0.947	-0.303	0.945
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 100%								
$5 \cdot 10^{-2} (F = 8.74)$	-0.267	0.917	-0.245	0.899	-0.306	0.918	-0.300	0.924
$5 \cdot 10^{-5} (F = 13.45)$	-0.280	0.941	-0.259	0.937	-0.306	0.932	-0.303	0.928
$5 \cdot 10^{-8} (F = 21.11)$	-0.290	0.947	-0.271	0.954	-0.307	0.938	-0.305	0.938

Table B.6: Mean of standard errors (SE) and standard deviation (SD) for 1000 simulations. Scenario: positive causal effect ($\beta_c = -0.3$) and moderate confounder ($\beta_u = \alpha_u = -0.6$). Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	SE	SD	SE	SD	SE	SD	SE	SD
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 0%								
$5 \cdot 10^{-2} (F = 8.74)$	0.068	0.067	0.072	0.063	0.070	0.077	0.070	0.077
$5 \cdot 10^{-5} (F = 13.45)$	0.079	0.080	0.083	0.075	0.081	0.088	0.081	0.088
$5 \cdot 10^{-8} (F = 21.11)$	0.095	0.102	0.099	0.097	0.097	0.109	0.097	0.109
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 25%								
$5 \cdot 10^{-2} (F = 8.74)$	0.068	0.067	0.072	0.063	0.070	0.076	0.070	0.077
$5 \cdot 10^{-5} (F = 13.45)$	0.079	0.081	0.083	0.076	0.081	0.090	0.081	0.090
$5 \cdot 10^{-8} (F = 21.11)$	0.095	0.099	0.098	0.094	0.097	0.106	0.097	0.106
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 50%								
$5 \cdot 10^{-2} (F = 8.74)$	0.068	0.067	0.072	0.063	0.070	0.078	0.070	0.078
$5 \cdot 10^{-5} (F = 13.45)$	0.079	0.085	0.083	0.079	0.081	0.094	0.081	0.094
$5 \cdot 10^{-8} (F = 21.11)$	0.095	0.098	0.098	0.093	0.097	0.106	0.097	0.106
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 75%								
$5 \cdot 10^{-2} (F = 8.74)$	0.068	0.070	0.072	0.065	0.070	0.080	0.070	0.080
$5 \cdot 10^{-5} (F = 13.45)$	0.079	0.084	0.082	0.078	0.081	0.093	0.081	0.093
$5 \cdot 10^{-8} (F = 21.11)$	0.095	0.096	0.098	0.091	0.097	0.103	0.097	0.103
$\beta_c = -0.3$, $\beta_u = -0.6 = \alpha_u = -0.6$, overlap = 100%								
$5 \cdot 10^{-2} (F = 8.74)$	0.068	0.072	0.072	0.066	0.070	0.083	0.070	0.083
$5 \cdot 10^{-5} (F = 13.45)$	0.079	0.083	0.082	0.077	0.081	0.092	0.081	0.092
$5 \cdot 10^{-8} (F = 21.11)$	0.095	0.099	0.098	0.092	0.097	0.105	0.097	0.105

Table B.7: Mean of causal estimate and coverage for 1000 simulations. Scenario: positive causal effect ($\beta_c = -0.3$) and strong confounder ($\beta_u = \alpha_u = -0.8$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	Mean	Coverage	Mean	Coverage	Mean	Coverage	Mean	Coverage
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 0%								
$5 \cdot 10^{-2} (F = 7.43)$	-0.254	0.909	-0.226	0.871	-0.299	0.919	-0.299	0.919
$5 \cdot 10^{-5} (F = 11.31)$	-0.273	0.944	-0.246	0.943	-0.304	0.927	-0.304	0.927
$5 \cdot 10^{-8} (F = 17.67)$	-0.277	0.922	-0.255	0.925	-0.297	0.922	-0.297	0.922
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 25%								
$5 \cdot 10^{-2} (F = 7.43)$	-0.251	0.904	-0.221	0.853	-0.295	0.919	-0.299	0.923
$5 \cdot 10^{-5} (F = 11.31)$	-0.273	0.928	-0.243	0.923	-0.303	0.934	-0.305	0.934
$5 \cdot 10^{-8} (F = 17.67)$	-0.277	0.937	-0.253	0.940	-0.297	0.935	-0.298	0.937
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 50%								
$5 \cdot 10^{-2} (F = 7.43)$	-0.250	0.882	-0.217	0.834	-0.295	0.908	-0.302	0.918
$5 \cdot 10^{-5} (F = 11.31)$	-0.270	0.916	-0.238	0.902	-0.300	0.920	-0.304	0.920
$5 \cdot 10^{-8} (F = 17.67)$	-0.280	0.936	-0.254	0.942	-0.301	0.933	-0.303	0.934
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 75%								
$5 \cdot 10^{-2} (F = 7.43)$	-0.247	0.856	-0.212	0.800	-0.290	0.896	-0.301	0.911
$5 \cdot 10^{-5} (F = 11.31)$	-0.267	0.916	-0.234	0.890	-0.297	0.915	-0.304	0.924
$5 \cdot 10^{-8} (F = 17.67)$	-0.280	0.943	-0.252	0.944	-0.301	0.941	-0.305	0.944
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 100%								
$5 \cdot 10^{-2} (F = 7.43)$	-0.243	0.862	-0.207	0.780	-0.287	0.885	-0.300	0.907
$5 \cdot 10^{-5} (F = 11.31)$	-0.265	0.910	-0.230	0.877	-0.295	0.916	-0.304	0.926
$5 \cdot 10^{-8} (F = 17.67)$	-0.280	0.936	-0.250	0.936	-0.301	0.929	-0.306	0.937

Table B.8: Mean of standard errors (SE) and standard deviation (SD) for 1000 simulations. Scenario: positive causal effect ($\beta_c = -0.3$) and strong confounder ($\beta_u = \alpha_u = -0.8$). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Threshold (F-stats)	IVW		Modified IVW		Exact		Modified Exact	
	SE	SD	SE	SD	SE	SD	SE	SD
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 0%								
$5 \cdot 10^{-2}(F = 7.43)$	0.071	0.070	0.075	0.065	0.073	0.083	0.073	0.083
$5 \cdot 10^{-5}(F = 11.31)$	0.082	0.083	0.087	0.078	0.085	0.094	0.085	0.094
$5 \cdot 10^{-8}(F = 17.67)$	0.099	0.107	0.104	0.101	0.102	0.115	0.102	0.115
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 25%								
$5 \cdot 10^{-2}(F = 7.43)$	0.071	0.071	0.075	0.066	0.073	0.083	0.073	0.083
$5 \cdot 10^{-5}(F = 11.31)$	0.082	0.085	0.087	0.079	0.085	0.097	0.085	0.097
$5 \cdot 10^{-8}(F = 17.67)$	0.099	0.105	0.104	0.099	0.102	0.114	0.102	0.114
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 50%								
$5 \cdot 10^{-2}(F = 7.43)$	0.071	0.072	0.075	0.067	0.073	0.084	0.073	0.084
$5 \cdot 10^{-5}(F = 11.31)$	0.082	0.089	0.087	0.083	0.085	0.101	0.086	0.101
$5 \cdot 10^{-8}(F = 17.67)$	0.099	0.104	0.104	0.098	0.102	0.114	0.103	0.114
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 75%								
$5 \cdot 10^{-2}(F = 7.43)$	0.071	0.075	0.075	0.070	0.073	0.088	0.074	0.087
$5 \cdot 10^{-5}(F = 11.31)$	0.082	0.089	0.087	0.084	0.085	0.101	0.086	0.101
$5 \cdot 10^{-8}(F = 17.67)$	0.099	0.103	0.105	0.097	0.102	0.112	0.103	0.112
$\beta = -0.3$, $\beta_u = \alpha_u = -0.8$, overlap = 100%								
$5 \cdot 10^{-2}(F = 7.43)$	0.071	0.076	0.076	0.071	0.073	0.091	0.074	0.090
$5 \cdot 10^{-5}(F = 11.31)$	0.082	0.088	0.088	0.082	0.085	0.100	0.086	0.099
$5 \cdot 10^{-8}(F = 17.67)$	0.099	0.106	0.105	0.099	0.102	0.115	0.104	0.115

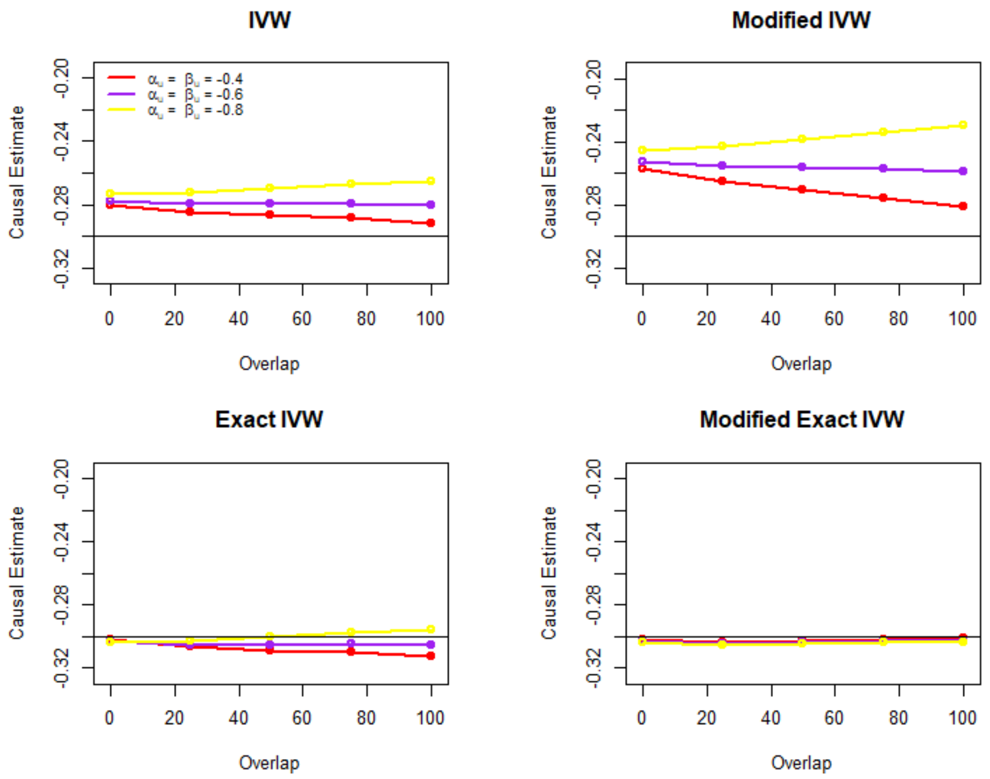


Figure B.3: Causal estimates with varing effect of confounder. Scenario: Significant threshold $5 \cdot 10^{-5}$; causal effect=-0.3 (horizontal line). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

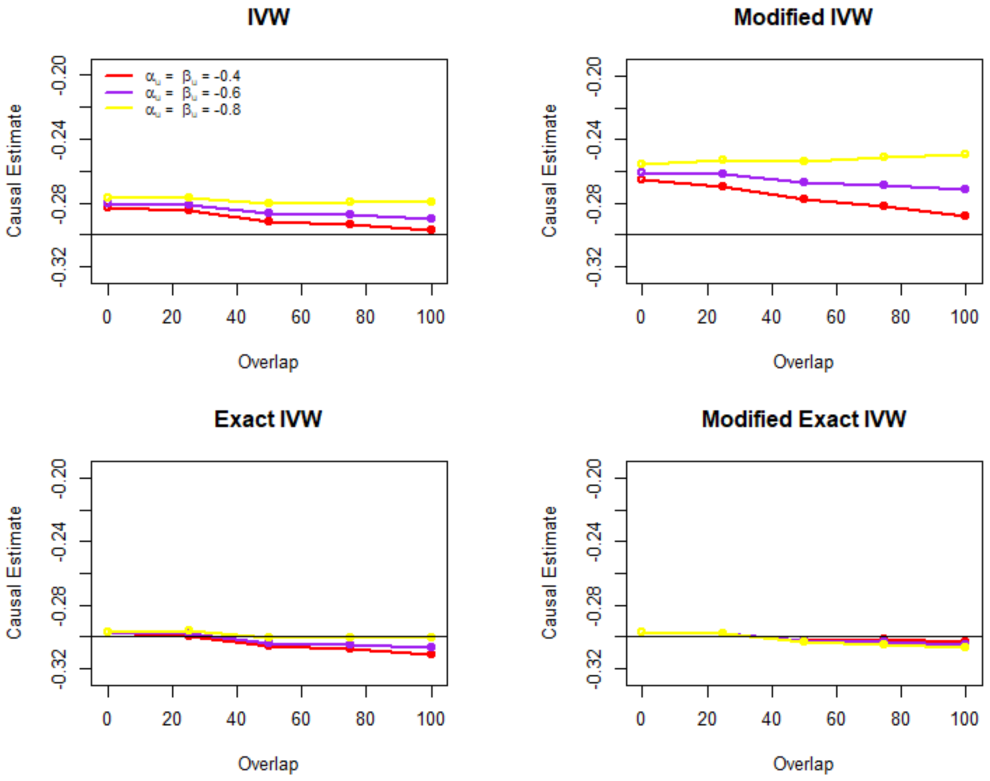


Figure B.4: Causal estimates with varing effect of confounder. Scenario: Significant threshold $5 \cdot 10^{-8}$; causal effect=-0.3 (horizontal line). Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Table B.9: Type I error with significance threshold of $5 \cdot 10^{-5}$. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Percentage of overlap	0%	25%	50%	75%	100%
$\beta_c = 0,\ \beta_u = \alpha_u = 0.4$					
IVW	0.045	0.045	0.059	0.067	0.059
Modified IVW	0.037	0.036	0.046	0.054	0.053
Exact	0.053	0.054	0.065	0.070	0.062
Modified exact	0.053	0.054	0.068	0.068	0.064
$\beta_c = 0,\ \beta_u = \alpha_u = 0.6$					
IVW	0.044	0.043	0.064	0.058	0.061
Modified IVW	0.035	0.035	0.052	0.062	0.086
Exact	0.049	0.052	0.068	0.066	0.071
Modified exact	0.049	0.051	0.066	0.064	0.059
$\beta_c = 0,\ \beta_u = \alpha_u = 0.8$					
IVW	0.040	0.040	0.060	0.058	0.071
Modified IVW	0.028	0.037	0.068	0.095	0.159
Exact	0.051	0.051	0.074	0.071	0.086
Modified exact	0.051	0.055	0.064	0.059	0.066

Table B.10: Type I error with significance threshold of $5 \cdot 10^{-8}$. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_c represents the causal effect, β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Percentage of overlap	0%	25%	50%	75%	100%
$\beta_c = 0$, $\beta_u = \alpha_u = 0.4$					
IVW	0.049	0.053	0.052	0.057	0.057
Modified IVW	0.039	0.042	0.047	0.055	0.052
Exact	0.050	0.055	0.054	0.059	0.057
Modified exact	0.050	0.055	0.054	0.059	0.057
$\beta_c = 0$, $\beta_u = \alpha_u = 0.6$					
IVW	0.051	0.056	0.059	0.059	0.046
Modified IVW	0.035	0.048	0.063	0.071	0.072
Exact	0.069	0.068	0.070	0.076	0.060
Modified exact	0.064	0.065	0.071	0.077	0.052
$\beta_c = 0$, $\beta_u = \alpha_u = 0.8$					
IVW	0.048	0.042	0.058	0.056	0.056
Modified IVW	0.039	0.036	0.047	0.063	0.067
Exact	0.052	0.046	0.062	0.064	0.059
Modified exact	0.052	0.046	0.061	0.057	0.057

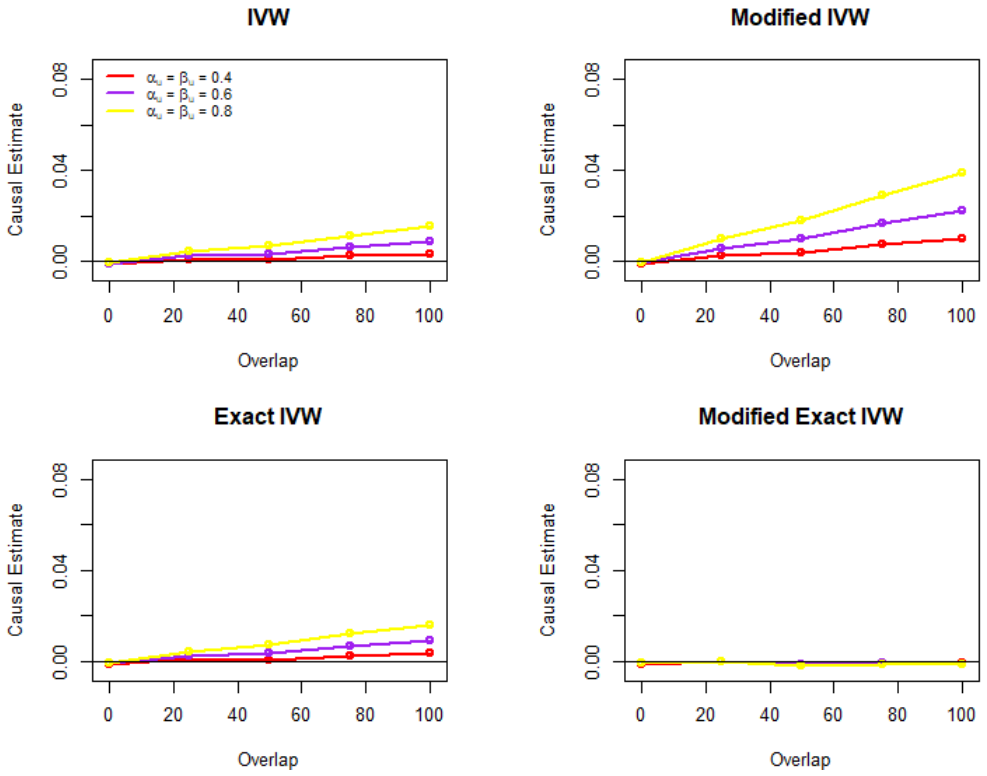


Figure B.5: Causal estimates with varing effect of confounder. Scenario: Significant threshold $5 \cdot 10^{-5}$; null causal effect. Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

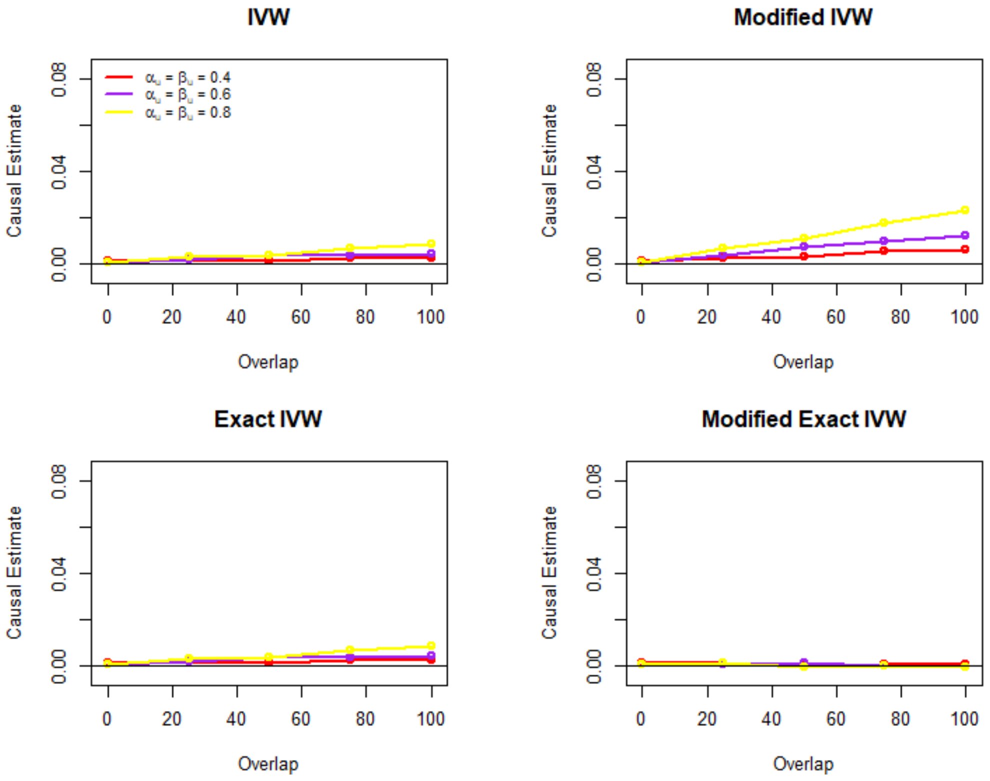


Figure B.6: Causal estimates with varing effect of confounder. Scenario: Significant threshold $5 \cdot 10^{-8}$; null causal effect. Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

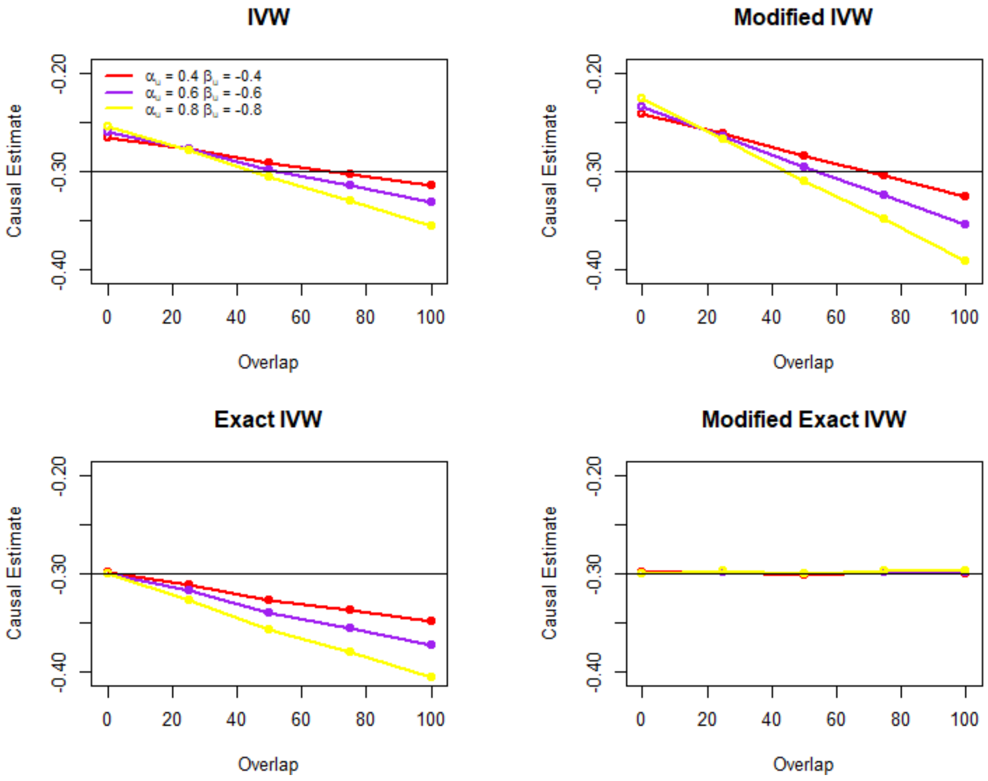


Figure B.7: Estimate with significance threshould of $5 \cdot 10^{-2}$ and negataive causal effect ($\beta_c = -0.3$) with negative confounder. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

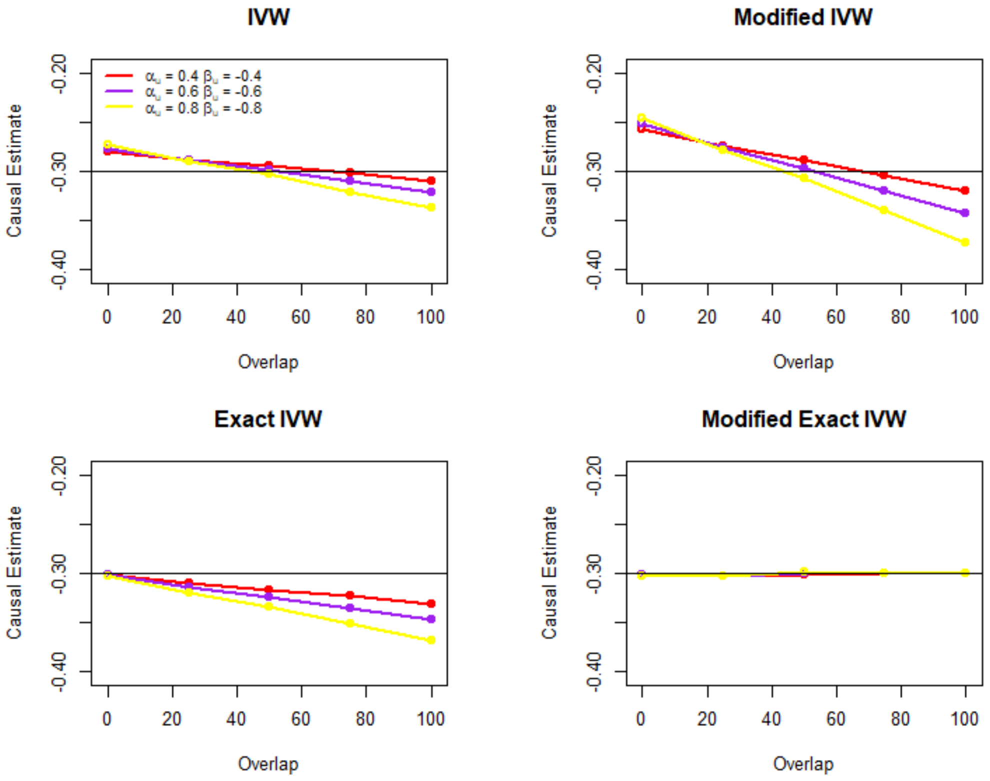


Figure B.8: Estimate with significance threshould of $5 \cdot 10^{-5}$ and negative causal effect ($\beta_c = -0.3$; horizontal line) with negative confounder. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

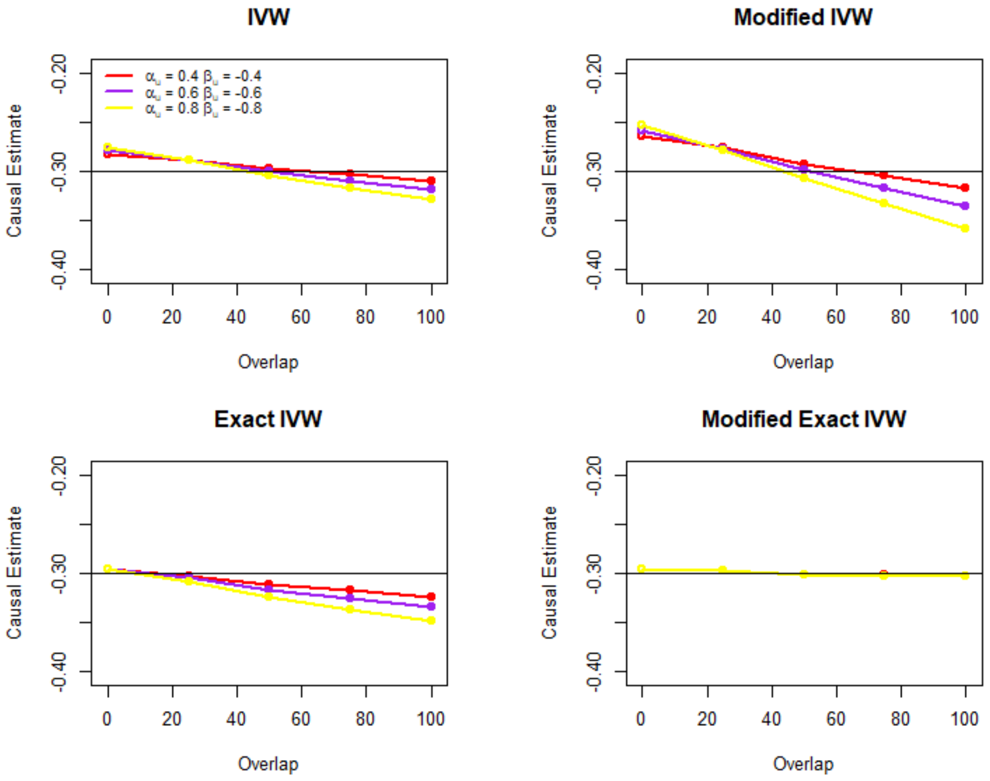


Figure B.9: Estimate with significance threshould of $5 \cdot 10^{-8}$ and negative causal effect ($\beta_c = -0.3$; horizontal line) with negative confounder. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

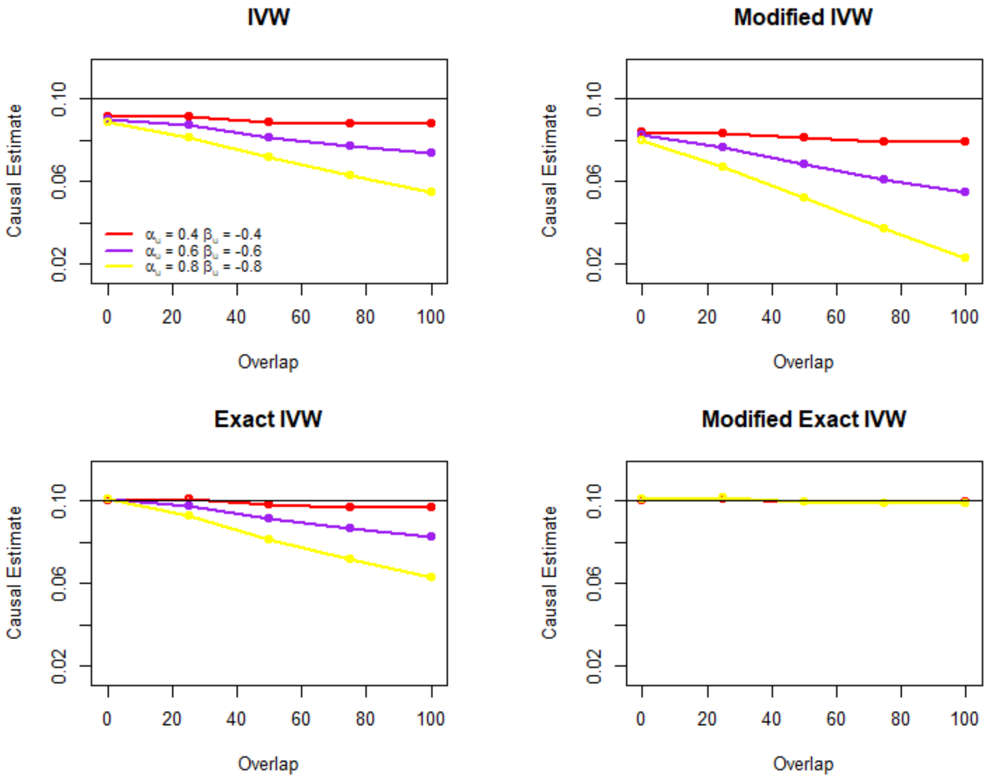


Figure B.10: Estimate with significance threshould of $5 \cdot 10^{-2}$ and positive causal effect ($\beta_c = 0.1$; horizontal line) with negative confounder. Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

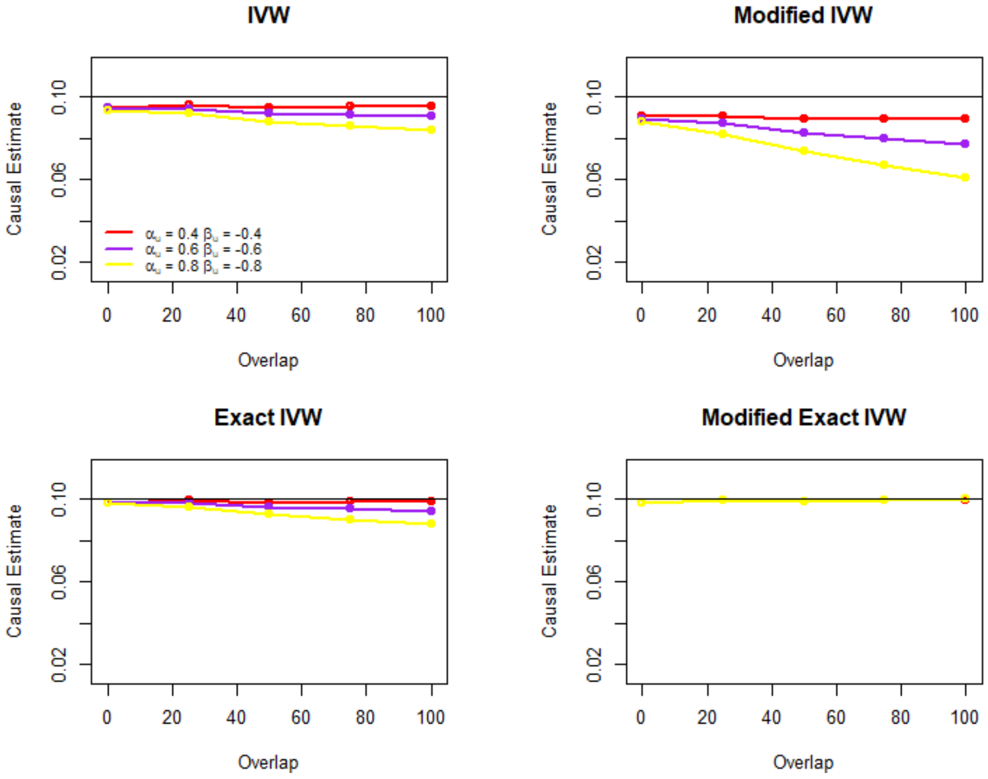


Figure B.11: Estimate with significance threshould of $5 \cdot 10^{-5}$ and positive causal effect ($\beta_c = 0.1$; horizontal line) with negative confounder. Abbreviations: IVW: "first order" inverse variance weight. Modified IVW: "second order" inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

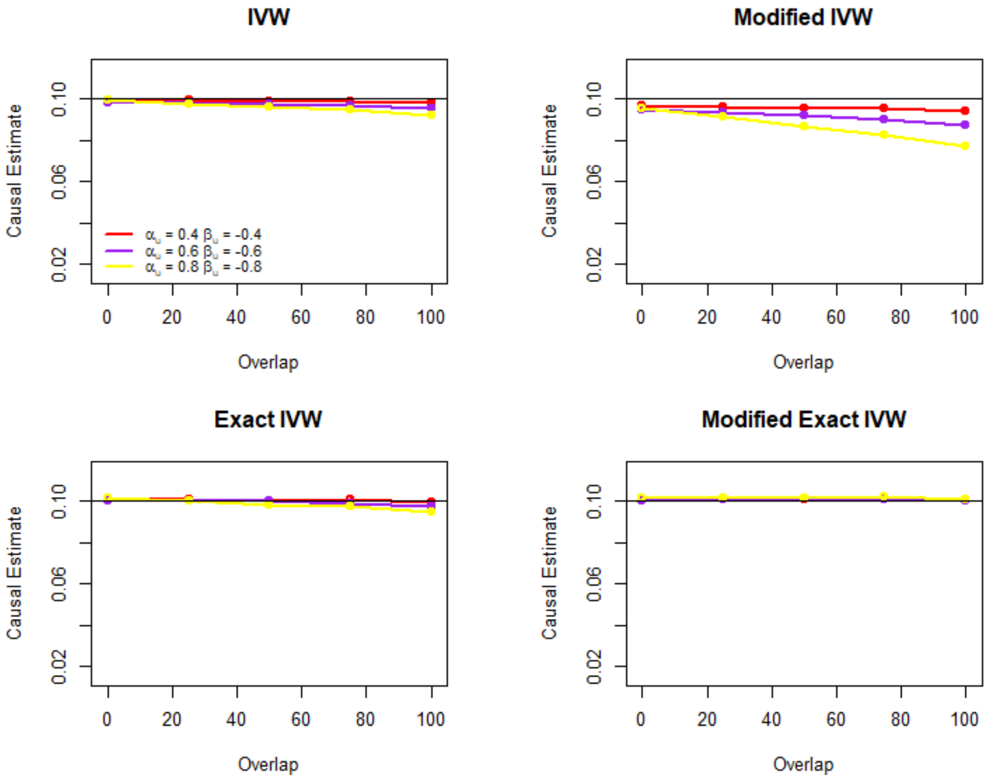


Figure B.12: Estimate with significance threshould of $5 \cdot 10^{-8}$ and positive causal effect ($\beta_c = 0.1$; horizontal line) with negative confounder. Abbreviations: IVW: “first order” inverse variance weight. Modified IVW: “second order” inverse variance weight that accounts for the covariance between association estimates. Exact: Original exact method without accounting for correlation. Modified Exact: exact method that accounts for the correlation in the weights. β_u and α_u represent the confounder effect on outcome and exposure, respectively.

Appendix C

Supplementary material of Chapter 4

C.1 Existing Methods

C.1.1 MR-PRESSO

MR-PRESSO (Verbanck *et al.*, 2018) method can be used as outlier detectation method that is relied on the summary statistics. It also requires that at least 50% of the variants are valid instruments and relies on the InSIDE assumption (Bowden *et al.*, 2015). The *MRPRESSO* software is available online `https://github.com/rondolab/MR-PRESSO`. MR-PRESSO is not considered in this thesis due to the cost of computation time. In "mr_presso" function, enough simulations are required to generate the null distribution of RSS. The author suggests at least 1000 replicates. However, it would not be enough if the number of variants is relatively large. For example, if the number of IVs is 120, the lower bound of empirical p-value would be

0.12 with 1000 replicates. This leads to no IV being called "invalid" at a significance level of 0.05. The computation time for one dataset including 120 potential IVs with 12000 replicates (with a lower bound of empirical p-value of 0.01) is 1.15 hours on anatolius. It is a multi-core virtual server with 40GB RAM and the operating system is CentOS Linux release 7.

C.1.2 MR-Lasso

In MR-Lasso (Rees *et al.*, 2019), the IVW regression is augmented so that an intercept term is allowed for each genetic variant. The intercept term accounts for the genetic variant-specific pleiotropy effect and it is penalized by L_1 loss function. $\hat{\beta}_{g_ix}$ is the estimate of the effect of variant i on the exposure, and $\hat{\beta}_{g_iy}$ is the estimate of the effect of variant i on the outcome. If the intercept term β_{0i} for genetic variant i shrinks to zero, then variant i is assumed to be a valid IV. Otherwise, the genetic variant is pleiotropic. The objective function is written as (Rees *et al.*, 2019):

$$\sum_i \text{var}(\hat{\beta}_{g_iy})^{-1}(\hat{\beta}_{g_iy} - \beta_{0i} - \beta_c\hat{\beta}_{g_ix})^2 + \lambda \sum_i |\beta_{0i}|$$

We implement MR-Lasso by simply using "mr_lasso" function in *MendelianRandomization* R package. The procedure for selecting the tuning parameter λ in "mr_lasso" is described in Rees *et al.* (2019).

C.1.3 sisVIVE

Kang *et al.* (2016) proposed a method for estimation of causal effects, called some invalid some valid IV estimator (sisVIVE). They showed that causal effects are identified and can be estimated when less than 50% of instruments are valid. It resembles the lasso (Tibshirani, 1996) procedures, but it only penalizes $\boldsymbol{\phi}$. The Lagrangian form is as follows

$$\underset{\boldsymbol{\phi},\beta_c}{\arg\min} \frac{1}{2}\|\boldsymbol{P_G}(\boldsymbol{Y} - \boldsymbol{G\phi} - \boldsymbol{X}\beta_c)\|_2^2 + \lambda\|\boldsymbol{\phi}\|_1$$

where $\boldsymbol{P_G}$ is the projection matrix of $\boldsymbol{G}$. The R package called *sisVIVE* is available. In *sisVIVE*, cross-validation is used for choosing λ. In our thesis, we use the same stopping rule based on Cochran Q statistics as in Lasso method. This stopping rule is considerably less computationally expensive than cross-validation. Additionally, the original *sisVIVE* code requires a lot of memory for a large dataset, which results in a longer computational time. As a result, we make a minor modification to *sisVIVE* so that it can be fitted for relatively large data sets. The computing time decreases noticeably as a result. We have informed the author of this point by email.

C.2 Threshold-based Method

We consider the exact linear regression model part used in Lasso method:

$$\boldsymbol{Y} = \boldsymbol{G\phi} + \boldsymbol{X}\beta_c + \boldsymbol{\rho}\boldsymbol{\hat{v}_2} + \boldsymbol{e} \tag{C.1}$$

Since $\boldsymbol{\hat{v}_2}$ is the exact linear function of $\boldsymbol{X}$ and $\boldsymbol{G}$, in which case equation (C.1) suffers from perfect collinearity. To avoid perfect collinearity, we randomly split $\boldsymbol{G}$

into half: $\boldsymbol{G_1}$ and $\boldsymbol{G_2}$.

$$\boldsymbol{Y} \sim \boldsymbol{G_1} + \boldsymbol{X} + \boldsymbol{\hat{v}_2}$$

$$\boldsymbol{Y} \sim \boldsymbol{G_2} + \boldsymbol{X} + \boldsymbol{\hat{v}_2}$$

Then we perform a hypothesis testing $H_0 : \phi_i = 0$ for each genetic variant i. If p-value>significance level (e.g. 5%), then that variant is called "valid".

C.3 Other Bootstrap Intervals

- Normal intervals with bootstrap estimate of bias

$$\left[\hat{\beta}_c - \text{bias}_{\text{boot}} - z_{1-\alpha/2}\ \hat{\sigma}_b, \hat{\beta}_c - \text{bias}_{\text{boot}} + z_{\alpha/2}\hat{\sigma}_b\right] \tag{C.1}$$

  where $\hat{\sigma}_b = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\beta}_r^* - \frac{1}{R}\sum_{b=1}^{R}\hat{\beta}_b^*)^2}$, $\text{bias}_{\text{boot}} = \frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_r^* - \hat{\beta}_c)$, and R is the number of bootstrap samples.

- Basic intervals

$$\left[2\hat{\beta}_c - \hat{\beta}^*_{(\alpha/2)}, 2\hat{\beta}_c - \hat{\beta}^*_{(1-\alpha/2)}\right] \tag{C.2}$$

  where $\hat{\beta}^*_{(\alpha/2)}$ and $\hat{\beta}^*_{(1-\alpha/2)}$ denote $\alpha/2$ and $1-\alpha/2$ percentiles of the bootstrap estimates $\hat{\beta}_c^*$, respectively.

- Studentized intervals

$$\left[\hat{\beta}_c - z^*_{1-\alpha/2}\ \hat{\sigma}, \hat{\beta}_c - z^*_{\alpha/2}\hat{\sigma}\right] \tag{C.3}$$

  where $z^*_{1-\alpha/2}$ and $z^*_{\alpha/2}$ denote $1-\alpha/2$ and $\alpha/2$ percentiles of the bootstrap z-statistic $z_r^* = \frac{\hat{\beta}_r^* - \hat{\beta}_c}{\hat{\sigma}^*}$, respectively.

C.4 Supplementary Tables

Table C.1: True positive rate (TPR) and false positive rate (FPR) over 1000 simulations. True positive rate is the probability that the invalid IVs are truly identified and false positive rate is the probability that valid IVs are wrongly classified as "invalid" ones. The entire data (100k) is used for both selection and estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	80 valid IVs		60 valid IVs		40 valid IVs	
Method	TPR	FPR	TPR	FPR	TPR	FPR
Scenario: balanced pleiotropy						
5% Sig.level	0.838	0.114	0.843	0.163	0.843	0.166
5% FDR	0.801	0.055	0.818	0.091	0.826	0.128
5% SID	0.717	0.006	0.723	0.012	0.721	0.018
Scenario: directional pleiotropy						
5% Sig.level	0.901	0.171	0.903	0.214	0.905	0.253
5% FDR	0.881	0.107	0.890	0.166	0.896	0.220
5% SID	0.824	0.020	0.828	0.036	0.828	0.055

Table C.2: Coverage and average width of 95% Wald-type intervals over 1000 simulations. The entire data (100k) is used for both selection and estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	80 valid IVs		60 valid IVs		40 valid IVs	
Method	Coverage	Width	Coverage	Width	Coverage	Width
Scenario: balanced pleiotropy;						
5% Sig.level	0.744	0.050	0.649	0.056	0.583	0.065
5% FDR	0.802	0.048	0.679	0.054	0.582	0.063
5% SID	0.802	0.045	0.669	0.050	0.549	0.055
Scenario: directional pleiotropy						
5% Sig.level	0.641	0.053	0.582	0.062	0.548	0.073
5% FDR	0.701	0.050	0.614	0.059	0.554	0.071
5% SID	0.794	0.047	0.680	0.053	0.503	0.061

Table C.3: Coverage and average width of bootstrap normal CIs over 1000 simulations for 5% Significance level method (Sig.level), 5% false discovery rate (FDR), 5% Šidák correction (SID). The entire data (100k) is used for both selection and estimation.

	Balanced pleiotropy						Directional pleiotropy					
	80 valid IVs		60 valid IVs		40 valid IVs		80 valid IVs		60 valid IVs		40 valid IVs	
Method	Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width
5% Sig.level	0.838	0.062	0.778	0.073	0.703	0.088	0.802	0.070	0.781	0.088	0.757	0.115
5% FDR	0.887	0.060	0.791	0.071	0.710	0.086	0.881	0.107	0.794	0.083	0.753	0.111
5% SID	0.893	0.057	0.813	0.067	0.697	0.082	0.881	0.059	0.820	0.072	0.720	0.094

Table C.4: True positive rate (TPR) and false positive rate (FPR) for threshold-based methods with sample-splitting over 1000 simulations. True positive rate is the probability that the invalid IVs are truly identified and false positive rate is the probability that valid IVs are wrongly classified as "invalid" ones. The sample size is 100,000; n_1 is the number of observation used in selection. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate; 5% SID, 5% Šidák correction.

	Balanced pleiotropy						Directional pleiotropy					
	80 valid IVs		60 valid IVs		40 valid IVs		80 valid IVs		60 valid IVs		40 valid IVs	
Method	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
						$n_1 = 10,000$						
5% Sig.level	0.542	0.058	0.548	0.061	0.548	0.065	0.696	0.064	0.699	0.072	0.702	0.078
5% FDR	0.403	0.009	0.440	0.016	0.460	0.023	0.606	0.016	0.634	0.030	0.653	0.040
5% SID	0.273	0.001	0.279	0.001	0.279	0.000	0.467	0.001	0.475	0.002	0.478	0.002
						$n_1 = 30,000$						
5% Sig.level	0.716	0.072	0.721	0.081	0.719	0.090	0.820	0.091	0.826	0.112	0.826	0.127
5%FDR	0.638	0.021	0.664	0.035	0.678	0.052	0.777	0.038	0.796	0.066	0.806	0.090
5%SID	0.513	0.001	0.518	0.002	0.522	0.003	0.683	0.003	0.686	0.006	0.690	0.010
						$n_1 = 50,000$						
5% Sig.level	0.776	0.085	0.782	0.097	0.778	0.113	0.861	0.117	0.864	0.145	0.863	0.169
5%FDR	0.720	0.031	0.741	0.051	0.751	0.074	0.829	0.057	0.843	0.097	0.850	0.134
5%SID	0.610	0.002	0.616	0.004	0.616	0.006	0.752	0.007	0.755	0.013	0.759	0.020

Table C.5: Average time for different methods over 5 simulations. Scenario: a total of 120 potential IVs with 80 valid; directional pleiotropy. The sample size is 100,000; n_1 is the number of observation used in selection. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction. Simulations are performed on a multi-core server. The hardware is 20 cores @ 2.67Ghz with 40GB RAM and the operating System is CentOS Linux release 7.

Method	entire data	$n_1 = 10k$	$n_1 = 30k$	$n_1 = 50k$
5% Sig.level	14.53	4.64	4.70	6.27
5% FDR	14.80	5.04	5.28	6.42
5% SID	16.06	5.37	5.47	6.91

Table C.6: Scenario: balanced pleiotropy. Coverage of causal effect ($\beta = 0.1$) and average width over 1000 simulations with different bootstrap CIs (percentile CIs, normal CIs) and Wald-type intervals. Causal effect is estimated by two stage least squares (TSLS) method. 1000 bootstrapped samples are generated (case resampling bootstrap). The sample size is 100,000; $n_1 = 10,000$ observations are used in selection and the remaining are used for estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	.	80 valid IVs		60 valid IVs		40 valid IVs	
Method		Coverage	CI width	Coverage	CI width	Coverage	CI width
Scenario:Balanced pleiotropy; $n_1 = 10,000$							
5% Sig.level	Wald CIs	0.611	0.047	0.461	0.050	0.362	0.053
	Percentile CIs	0.910	0.078	0.863	0.099	0.805	0.125
	Normal CIs	0.827	0.079	0.770	0.100	0.731	0.125
5% FDR	Wald CIs	0.507	0.045	0.373	0.047	0.330	0.050
	Percentile CIs	0.851	0.081	0.798	0.101	0.768	0.124
	Normal CIs	0.779	0.081	0.724	0.101	0.706	0.125
5% SID	Wald CIs	0.436	0.044	0.336	0.045	0.260	0.046
	Percentile CIs	0.780	0.084	0.707	0.103	0.664	0.123
	Normal CIs	0.722	0.084	0.659	0.103	0.635	0.124
MR-Lasso	Wald CIs	0.596	0.046	0.451	0.049	0.357	0.053
	Percentile CIs	0.915	0.081	0.887	0.102	0.837	0.129
	Normal CIs	0.840	0.091	0.887	0.102	0.772	0.130
Lasso	Wald CIs	0.524	0.045	0.425	0.048	0.345	0.051
	Percentile CIs	0.879	0.082	0.842	0.101	0.799	0.125
	Normal CIs	0.790	0.081	0.743	0.102	0.728	0.126
sisVIVE	Wald CIs	0.513	0.045	0.402	0.047	0.325	0.051
	Percentile CIs	0.869	0.082	0.836	0.102	0.790	0.127
	Normal CIs	0.776	0.083	0.735	0.103	0.724	0.127

Table C.7: Scenario: directional pleiotropy. Coverage of causal effect ($\beta = 0.1$) and average width over 1000 simulations with different bootstrap CIs (percentile CIs, normal CIs) and Wald-type intervals. Causal effect is estimated by two stage least squares (TSLS) method. 1000 bootstrapped samples are generated (case resampling bootstrap). The sample size is 100,000; $n_1 = 10,000$ observations are used in selection and the remaining are used for estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	.	80 valid IVs		60 valid IVs		40 valid IVs	
Method		Coverage	CI width	Coverage	CI width	Coverage	CI width
Scenario:Directional pleiotropy; $n_1 = 10,000$							
5% Sig.level	Wald CIs	0.624	0.049	0.473	0.053	0.357	0.059
	Percentile CIs	0.905	0.080	0.886	0.108	0.843	0.148
	Normal CIs	0.871	0.081	0.823	0.108	0.759	0.148
5% FDR	Wald CIs	0.533	0.047	0.431	0.051	0.303	0.056
	Percentile CIs	0.866	0.084	0.858	0.111	0.811	0.149
	Normal CIs	0.800	0.085	0.779	0.111	0.741	0.149
5% SID	Wald CIs	0.373	0.045	0.316	0.048	0.230	0.050
	Percentile CIs	0.787	0.094	0.766	0.123	0.702	0.161
	Normal CIs	0.730	0.095	0.694	0.124	0.658	0.161
MR-Lasso	Wald CIs	0.590	0.047	0.501	0.052	0.359	0.058
	Percentile CIs	0.918	0.082	0.919	0.107	0.859	0.145
	Normal CIs	0.831	0.082	0.827	0.107	0.777	0.146
Lasso	Wald CIs	0.522	0.046	0.428	0.050	0.319	0.055
	Percentile CIs	0.881	0.084	0.884	0.107	0.808	0.141
	Normal CIs	0.808	0.085	0.784	0.107	0.728	0.141
sisVIVE	Wald CIs	0.509	0.046	0.411	0.050	0.310	0.055
	Percentile CIs	0.873	0.086	0.880	0.109	0.800	0.144
	Normal CIs	0.797	0.086	0.784	0.109	0.711	0.144

Table C.8: Scenario: balanced pleiotropy. Coverage of causal effect ($\beta = 0.1$) and average width over 1000 simulations with different bootstrap CIs (percentile CIs, normal CIs) and Wald-type intervals. Causal effect is estimated by two stage least squares (TSLS) method. 1000 bootstrapped samples are generated (case resampling bootstrap). The sample size is 100,000; $n_1 = 30,000$ observations are used in selection and the remaining are used for estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	.	80 valid IVs		60 valid IVs		40 valid IVs	
Method		Coverage	CI width	Coverage	CI width	Coverage	CI width
Scenario:Balanced pleiotropy; $n_1 = 30,000$							
5% Sig.level	Wald CIs	0.842	0.056	0.711	0.061	0.576	0.067
	Percentile CIs	0.940	0.069	0.888	0.084	0.809	0.105
	Normal CIs	0.924	0.070	0.858	0.084	0.777	0.105
5% FDR	Wald CIs	0.798	0.054	0.661	0.058	0.530	0.065
	Percentile CIs	0.923	0.067	0.854	0.082	0.790	0.103
	Normal CIs	0.887	0.067	0.825	0.082	0.762	0.103
5% SID	Wald CIs	0.674	0.052	0.505	0.054	0.415	0.058
	Percentile CIs	0.872	0.069	0.793	0.084	0.707	0.104
	Normal CIs	0.817	0.069	0.734	0.084	0.661	0.104
MR-Lasso	Wald CIs	0.828	0.054	0.732	0.059	0.622	0.066
	Percentile CIs	0.957	0.067	0.931	0.081	0.868	0.100
	Normal CIs	0.913	0.068	0.884	0.081	0.825	0.101
Lasso	Wald CIs	0.776	0.053	0.689	0.057	0.573	0.064
	Percentile CIs	0.940	0.067	0.909	0.079	0.827	0.096
	Normal CIs	0.867	0.067	0.841	0.079	0.775	0.097
sisVIVE	Wald CIs	0.765	0.053	0.673	0.057	0.568	0.063
	Percentile CIs	0.941	0.067	0.909	0.080	0.824	0.098
	Normal CIs	0.867	0.067	0.842	0.080	0.768	0.098

Table C.9: Scenario: directional pleiotropy. Coverage of causal effect ($\beta = 0.1$) and average width over 1000 simulations with different bootstrap CIs (percentile CIs, normal CIs) and Wald-type intervals. Causal effect is estimated by two stage least squares (TSLS) method. 1000 bootstrapped samples are generated (case resampling bootstrap). The sample size is 100,000; $n_1 = 30,000$ observations are used in selection and the remaining are used for estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	.	80 valid IVs		60 valid IVs		40 valid IVs	
Method		Coverage	CI width	Coverage	CI width	Coverage	CI width
Scenario:Directional pleiotropy; $n_1 = 30,000$							
	Wald CIs	0.851	0.058	0.755	0.065	0.588	0.074
5% Sig.level	Percentile CIs	0.925	0.071	0.893	0.089	0.826	0.120
	Normal CIs	0.930	0.071	0.885	0.090	0.812	0.120
	Wald CIs	0.809	0.056	0.722	0.063	0.558	0.072
5% FDR	Percentile CIs	0.922	0.069	0.881	0.088	0.739	0.119
	Normal CIs	0.896	0.069	0.854	0.088	0.694	0.119
	Wald CIs	0.695	0.054	0.547	0.058	0.392	0.064
5% SID	Percentile CIs	0.870	0.070	0.822	0.090	0.816	0.118
	Normal CIs	0.830	0.070	0.773	0.090	0.802	0.118
	Wald CIs	0.858	0.055	0.790	0.061	0.675	0.071
MR-Lasso	Percentile CIs	0.962	0.067	0.949	0.081	0.893	0.105
	Normal CIs	0.930	0.067	0.902	0.081	0.857	0.105
	Wald CIs	0.803	0.054	0.738	0.060	0.610	0.069
Lasso	Percentile CIs	0.948	0.066	0.931	0.079	0.859	0.100
	Normal CIs	0.882	0.066	0.865	0.079	0.801	0.100
	Wald CIs	0.790	0.054	0.717	0.059	0.594	0.068
sisVIVE	Percentile CIs	0.944	0.066	0.929	0.079	0.863	0.100
	Normal CIs	0.878	0.066	0.858	0.079	0.798	0.101

Table C.10: Scenario: balanced pleiotropy. Coverage of causal effect ($\beta = 0.1$) and average width over 1000 simulations with different bootstrap CIs (percentile CIs, normal CIs) and Wald-type intervals. Causal effect is estimated by two stage least squares (TSLS) method. 1000 bootstrapped samples are generated (case resampling bootstrap). The sample size is 100,000; $n_1 = 50,000$ observations are used in selection and the remaining are used for estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	.	80 valid IVs		60 valid IVs		40 valid IVs	
Method		Coverage	CI width	Coverage	CI width	Coverage	CI width
Scenario:Balanced pleiotropy; $n_1 = 50,000$							
5% Sig.level	Wald CIs	0.904	0.067	0.828	0.075	0.725	0.084
	Percentile CIs	0.963	0.077	0.930	0.091	0.851	0.112
	Normal CIs	0.943	0.078	0.910	0.091	0.849	0.112
5% FDR	Wald CIs	0.880	0.064	0.794	0.071	0.699	0.081
	Percentile CIs	0.953	0.074	0.917	0.088	0.884	0.108
	Normal CIs	0.923	0.074	0.890	0.088	0.829	0.108
5% SID	Wald CIs	0.818	0.062	0.689	0.067	0.561	0.072
	Percentile CIs	0.938	0.072	0.870	0.085	0.770	0.102
	Normal CIs	0.877	0.072	0.810	0.085	0.712	0.103
MR-Lasso	Wald CIs	0.895	0.064	0.861	0.071	0.784	0.081
	Percentile CIs	0.990	0.074	0.971	0.086	0.921	0.104
	Normal CIs	0.947	0.075	0.919	0.086	0.883	0.104
Lasso	Wald CIs	0.867	0.063	0.821	0.069	0.723	0.078
	Percentile CIs	0.984	0.072	0.958	0.083	0.896	0.099
	Normal CIs	0.914	0.073	0.894	0.083	0.851	0.099
sisVIVE	Wald CIs	0.861	0.063	0.815	0.069	0.714	0.077
	Percentile CIs	0.984	0.072	0.958	0.083	0.899	0.099
	Normal CIs	0.911	0.072	0.889	0.083	0.851	0.099

Table C.11: Scenario: directional pleiotropy. Coverage of causal effect ($\beta = 0.1$) and average width over 1000 simulations with different bootstrap CIs (percentile CIs, normal CIs) and Wald-type intervals. Causal effect is estimated by two stage least squares (TSLS) method. 1000 bootstrapped samples are generated (case resampling bootstrap). The sample size is 100,000; $n_1 = 50,000$ observations are used in selection and the remaining are used for estimation. Abbreviation: 5% Sig.level, 5% significance level; 5% FDR, 5% false discovery rate. 5% SID, 5% Šidák correction.

	.	80 valid IVs		60 valid IVs		40 valid IVs	
Method		Coverage	CI width	Coverage	CI width	Coverage	CI width
Scenario:Directional pleiotropy; $n_1 = 50,000$							
	Wald CIs	0.926	0.071	0.858	0.080	0.745	0.093
5% Sig.level	Percentile CIs	0.962	0.081	0.921	0.099	0.863	0.129
	Normal CIs	0.951	0.082	0.922	0.099	0.878	0.129
	Wald CIs	0.897	0.067	0.842	0.077	0.724	0.090
5% FDR	Percentile CIs	0.963	0.077	0.913	0.095	0.850	0.125
	Normal CIs	0.942	0.077	0.926	0.095	0.864	0.126
	Wald CIs	0.835	0.064	0.708	0.071	0.568	0.080
5% SID	Percentile CIs	0.937	0.073	0.880	0.089	0.796	0.117
	Normal CIs	0.883	0.074	0.843	0.090	0.774	0.117
	Wald CIs	0.920	0.066	0.883	0.074	0.844	0.086
MR-Lasso	Percentile CIs	0.986	0.075	0.986	0.097	0.949	0.107
	Normal CIs	0.953	0.075	0.947	0.087	0.92	0.107
	Wald CIs	0.891	0.065	0.864	0.072	0.777	0.083
Lasso	Percentile CIs	0.981	0.072	0.977	0.083	0.935	0.102
	Normal CIs	0.929	0.072	0.937	0.083	0.875	0.102
	Wald CIs	0.888	0.064	0.862	0.072	0.764	0.083
sisVIVE	Percentile CIs	0.981	0.072	0.977	0.083	0.936	0.102
	Normal CIs	0.926	0.072	0.931	0.083	0.871	0.102

Table C.12: Coverages and average widths of different 95% bootstrap intervals for new Lasso method over 1000 simulations. The entire data ($100k$) is used for both selection and estimation. Abbreviation: Normal, Normal intervals; Normal (Bias correction), Normal intervals with bootstrap estimate of bias; Basic, Basic intervals; Percentile, Percentile intervals; Studentized, Studentized intervals

	80 valid IVs		60 valid IVs		40 valid IVs	
Method	Coverage	Width	Coverage	Width	Coverage	Width
Scenario: balanced pleiotropy;						
Normal	0.971	0.061	0.928	0.070	0.878	0.083
Normal (Bias correction)	0.946	0.061	0.902	0.070	0.871	0.083
Basic	0.935	0.061	0.899	0.070	0.859	0.086
Percentile	0.948	0.061	0.896	0.070	0.814	0.083
Studentized	1	0.077	1	0.095	1	0.127
Scenario: directional pleiotropy						
Normal	0.972	0.061	0.964	0.071	0.912	0.086
Normal (Bias correction)	0.929	0.062	0.916	0.071	0.891	0.086
Basic	0.958	0.062	0.916	0.071	0.886	0.086
Percentile	0.927	0.062	0.905	0.071	0.818	0.086
Studentized	1	0.079	1	0.091	1	0.123

C.5 Supplementary Figures

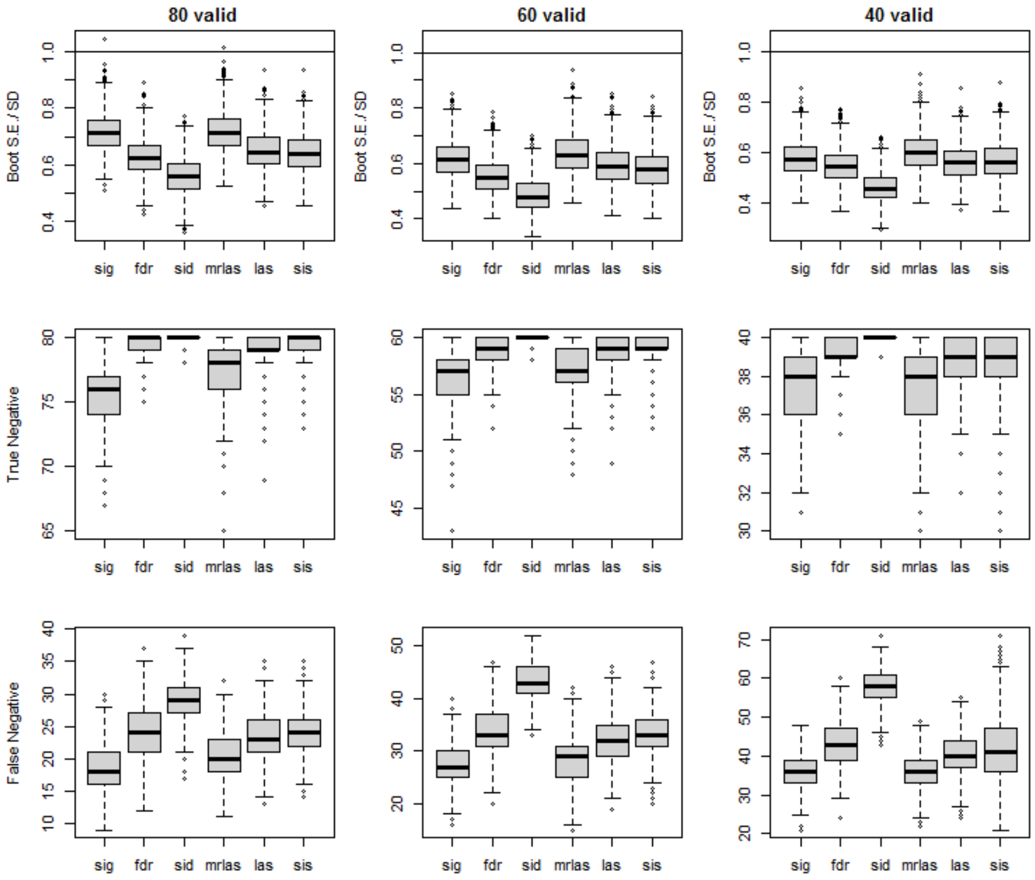


Figure C.1: Scenario: balanced pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations.The sample size is $100k$. $10k$ observations are used for selection and the remaining for estimation. Abbreviation: sig, 5% significance level; fdr, 5% false discovery rate; sid, 5% Šidák correction; mrlas, MR-Lasso; las:Lasso; sis, sisVIVE.

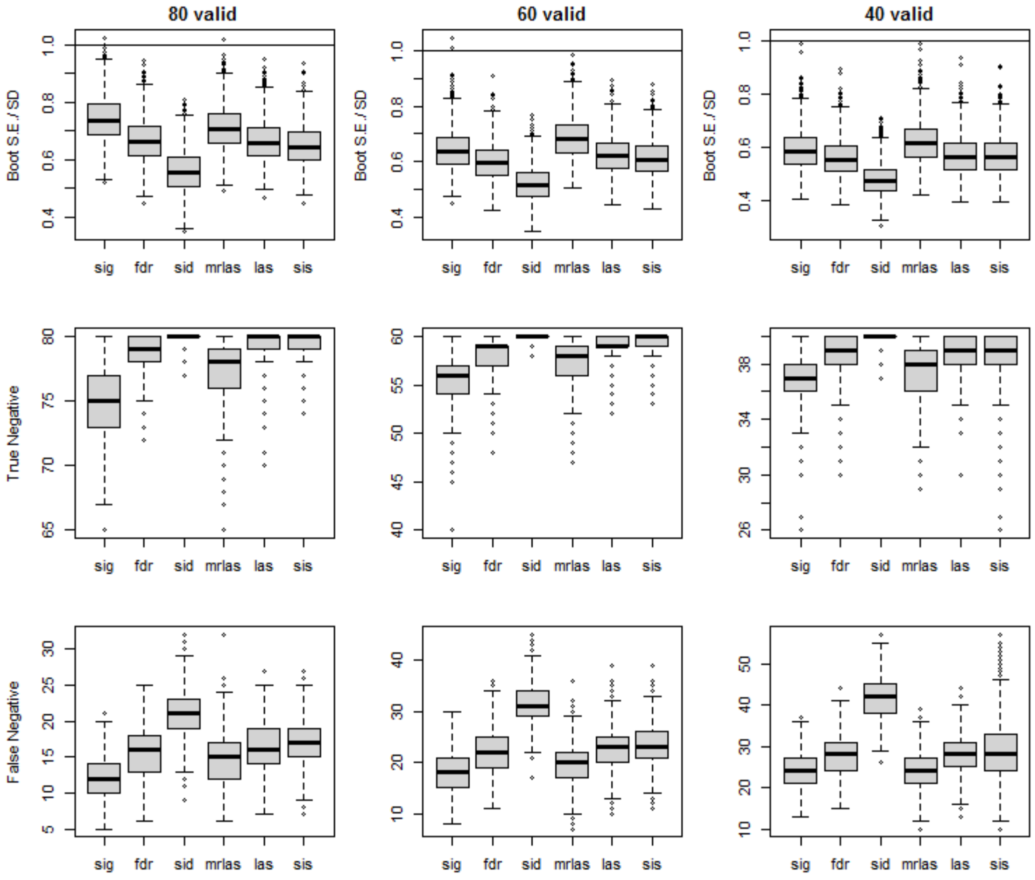


Figure C.2: Scenario: directional pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations.The sample size is 100k. 10k observations are used for selection and the remaining for estimation. Abbreviation: sig, 5% significance level; fdr, 5% false discovery rate; sid, 5% Šidák correction; mrlas, MR-Lasso; las:Lasso; sis, sisVIVE.

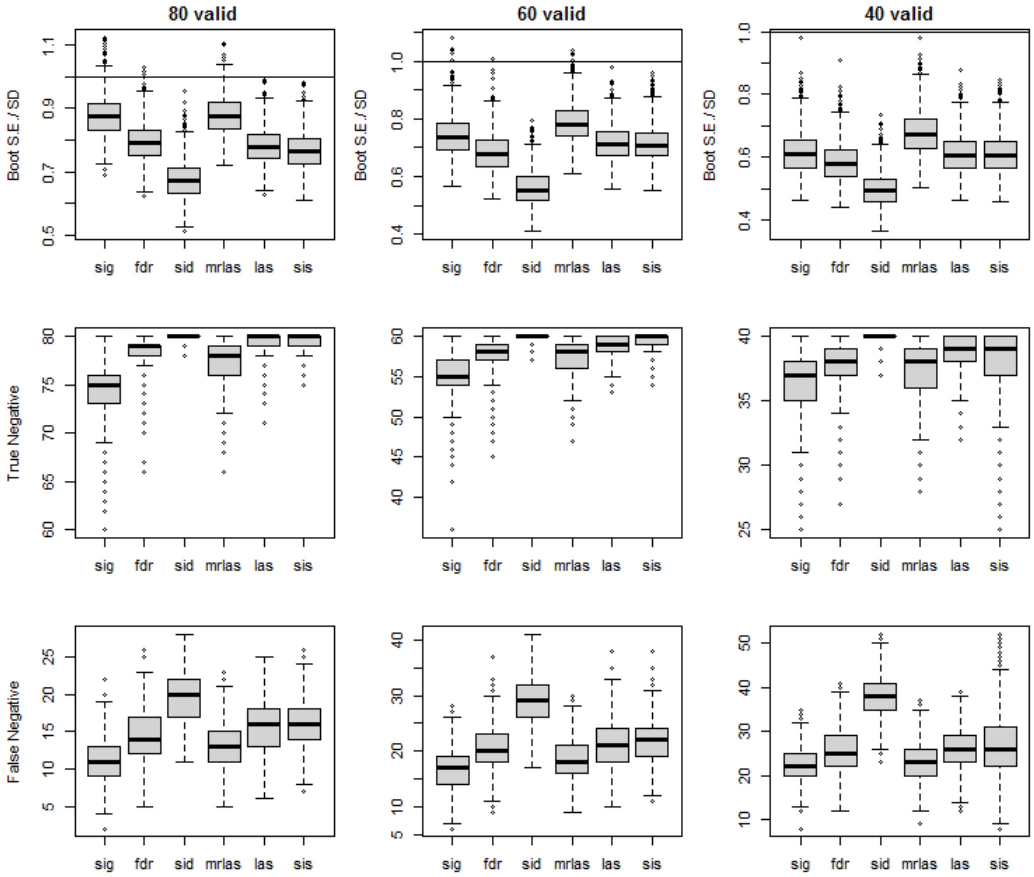


Figure C.3: Scenario: balanced pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations. The sample size is $100k$. $30k$ observations are used for selection and the remaning for estimation. Abbreviation: sig, 5% significance level; fdr, 5% false discovery rate; sid, 5% Šidák correction; mrlas, MR-Lasso; las:Lasso; sis, sisVIVE.

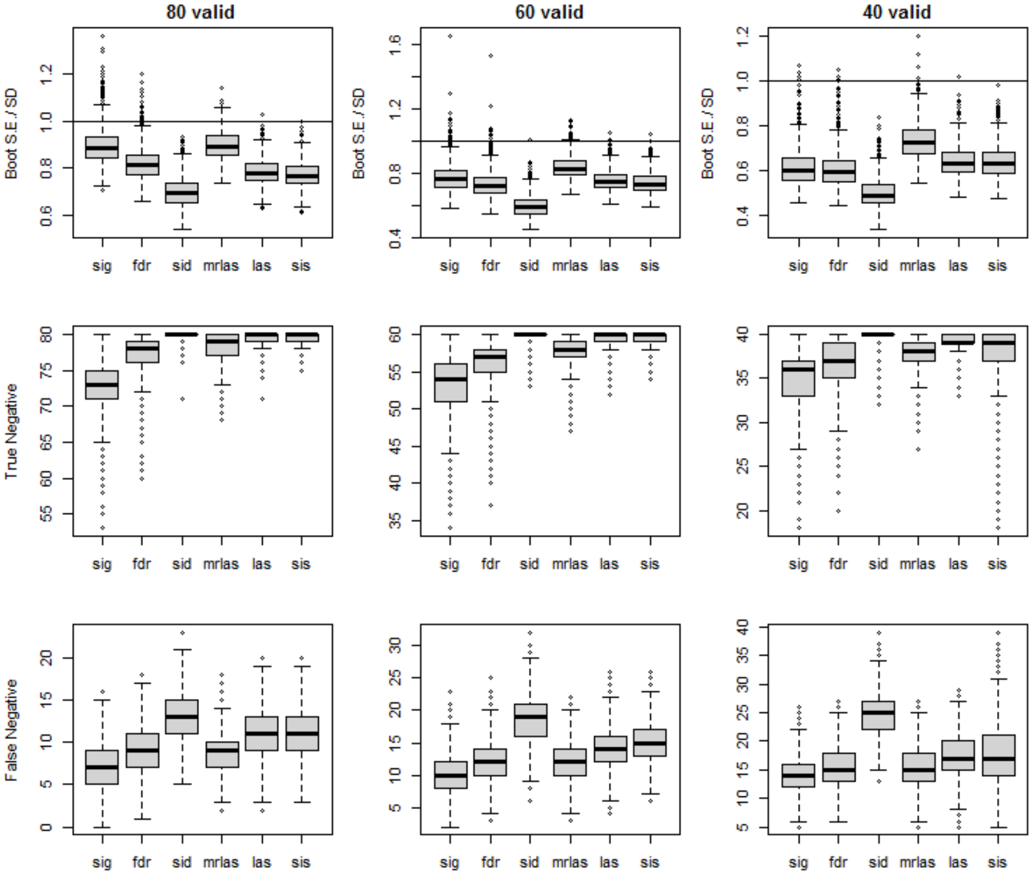


Figure C.4: Scenario: directional pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations. The sample size is 100k. 30k observations are used for selection and the remaining for estimation. Abbreviation: sig, 5% significance level; fdr, 5% false discovery rate; sid, 5% Šidák correction; mrlas, MR-Lasso; las:Lasso; sis, sisVIVE.

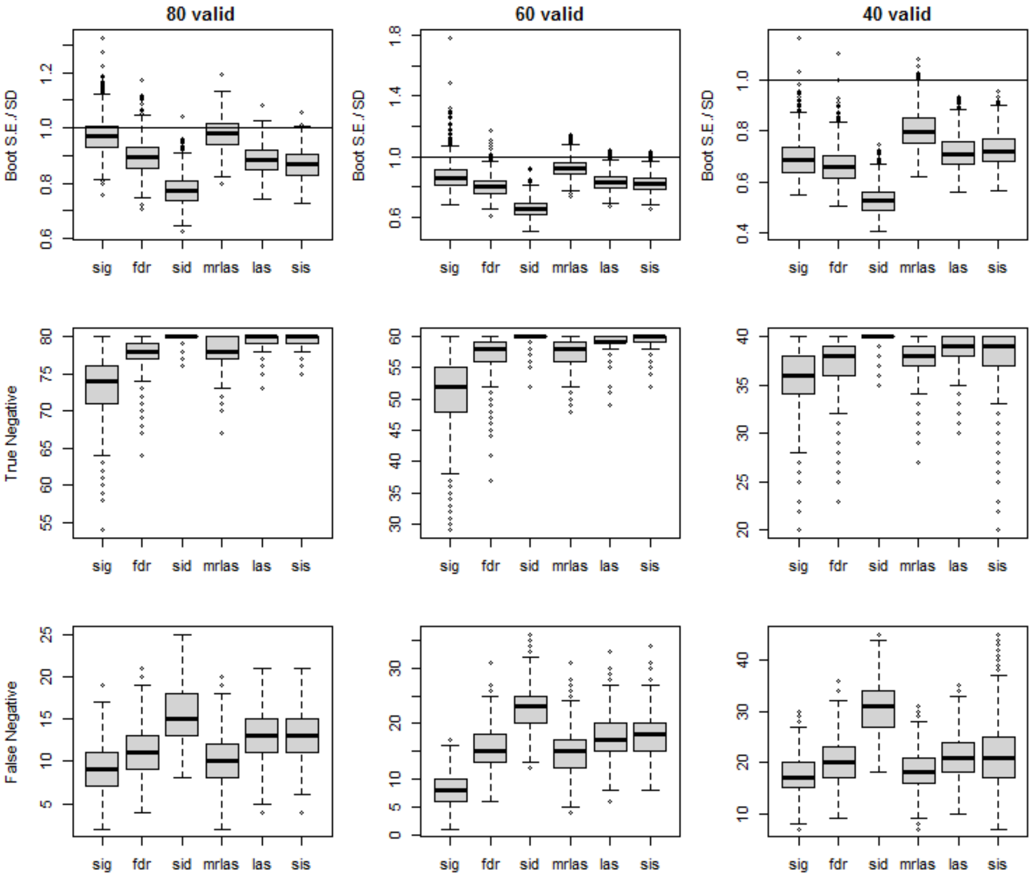


Figure C.5: Scenario: balanced pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations. The sample size is $100k$. $50k$ observations are used for selection and the remaining for estimation. Abbreviation: sig, 5% significance level; fdr, 5% false discovery rate; sid, 5% Šidák correction; mrlas, MR-Lasso; las:Lasso; sis, sisVIVE.

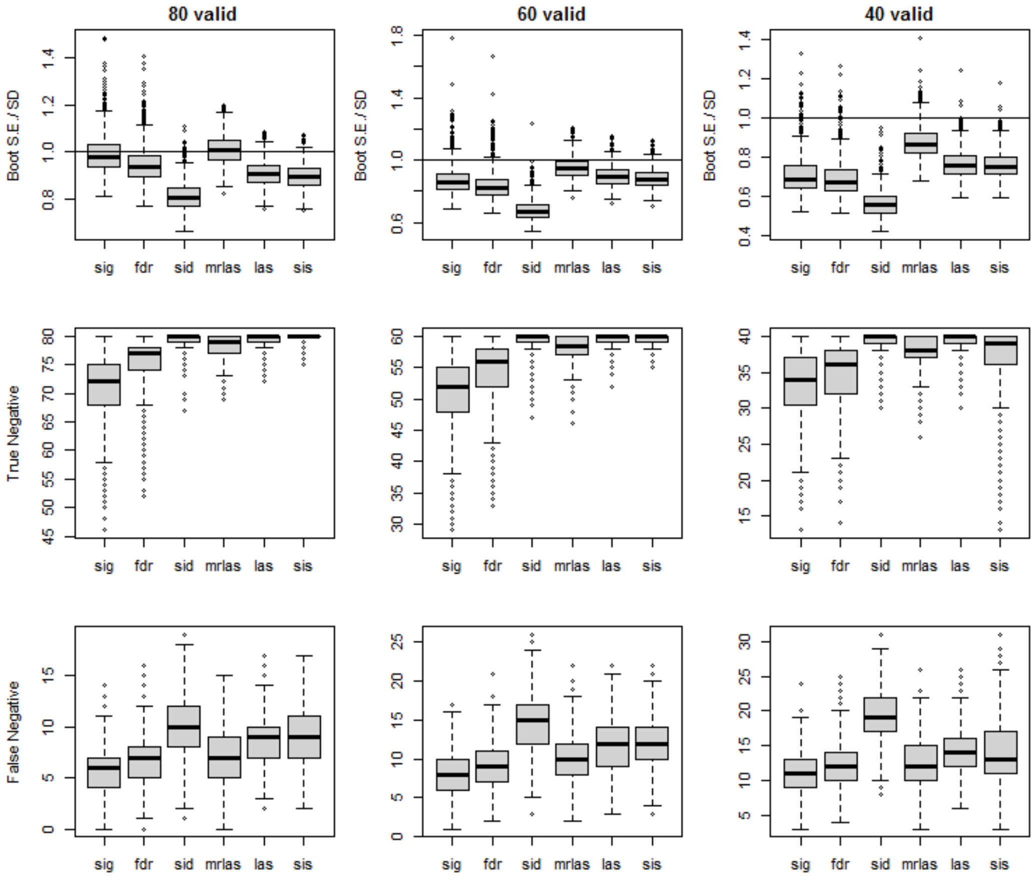


Figure C.6: Scenario: directional pleiotropy. Boxplot of bootstrap SE/SD, true positives, false negatives over 1000 simulations. The sample size is $100k$. $50k$ observations are used for selection and the remaining for estimation. Abbreviation: sig, 5% significance level; fdr, 5% false discovery rate; sid, 5% Šidák correction; mrlas, MR-Lasso; las:Lasso; sis, sisVIVE.

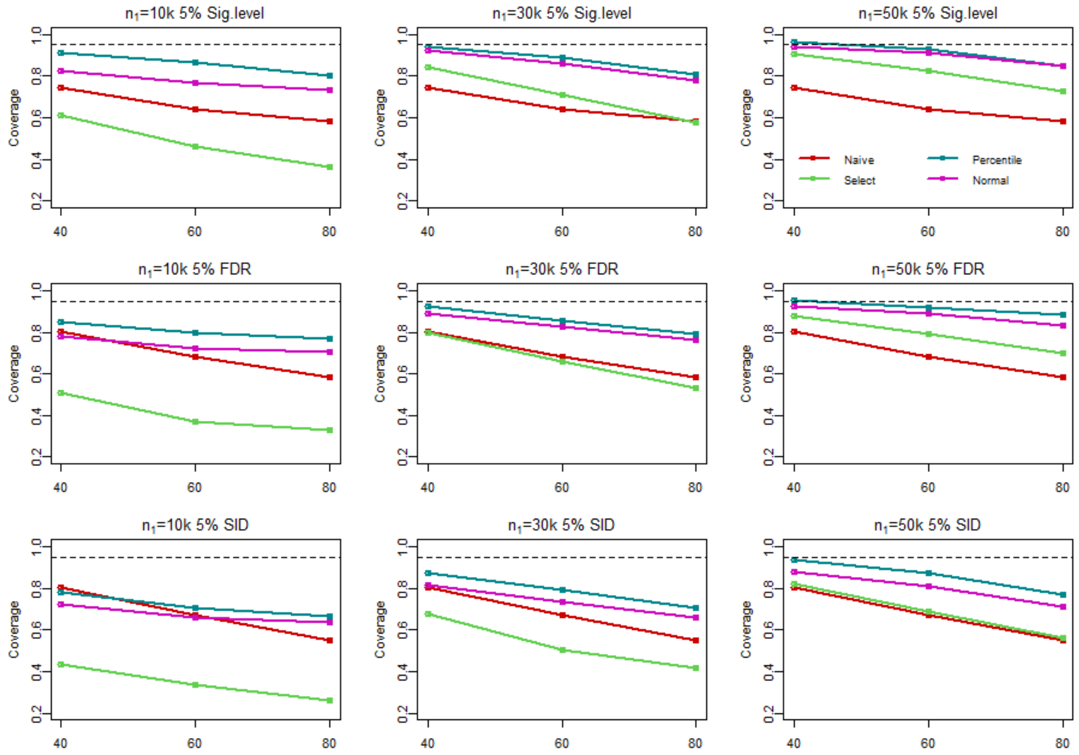


Figure C.7: Scenario: balanced pleiotropy. Comparison among bootstrap CIs and Wald CIs accounting for selection with different splits of data for selection and Wald CIs without accounting for selection. Abbreviations: Naive, Wald CIs using all data for selection and estimation; Select, Wald CIs; Percentile, Percentile CIs; Normal, Normal CIs. For Select, Percentile and Normal CIs, part of data ($n_1 = 10k$, $30k$ or $50k$) are used for selecting valid IVs and the remaining for estimation. The sample size is $100k$.

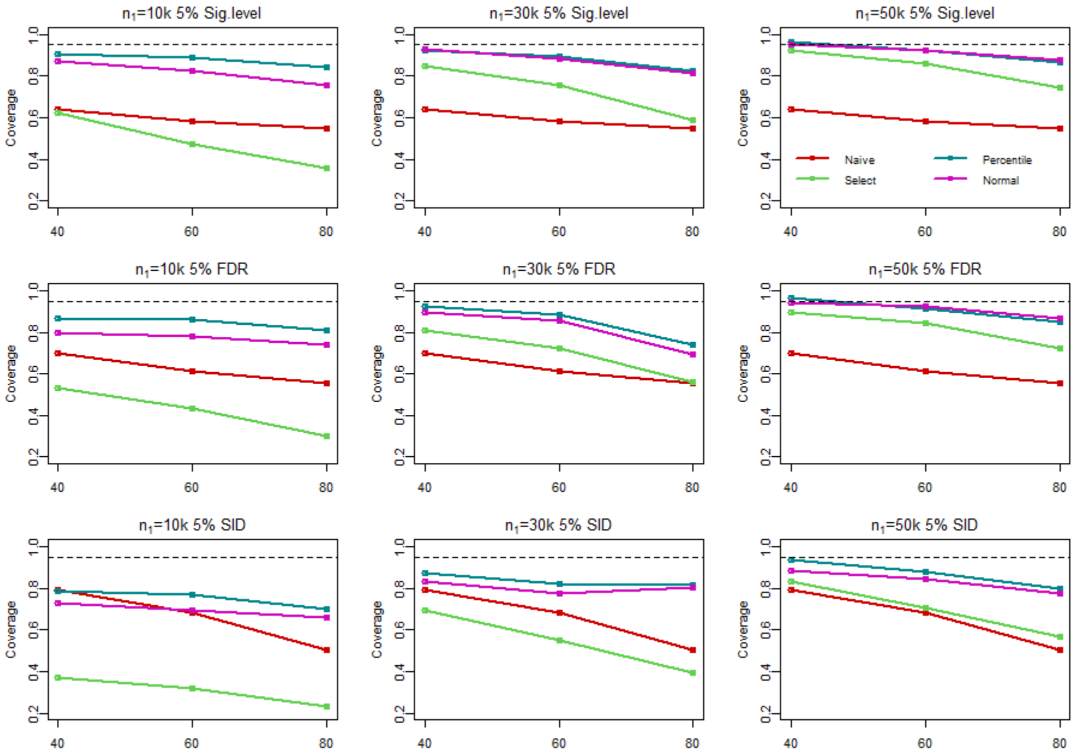


Figure C.8: Scenario: directional pleiotropy. Comparison among bootstrap CIs and Wald CIs accounting for selection with different splits of data for selection and Wald CIs without accounting for selection. Abbreviations: Naive, Wald CIs using all data for selection and estimation; Select, Wald CIs; Percentile, Percentile CIs; Normal, Normal CIs. For Select, Percentile and Normal CIs, part of data ($n_1 = 10k$, $30k$ or $50k$) are used for selecting valid IVs and the remaining for estimation. The sample size is $100k$.

Bibliography

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56.

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, **20**(1), 46–63.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, **91**(434), 444–455.

Annamalai, A., Kosir, U., and Tek, C. (2017). Prevalence of obesity and diabetes in patients with schizophrenia. *World Journal of Diabetes*, **8**(8), 390–396.

Bao, Y., Clarke, P. S., Smart, M., and Kumari, M. (2019). Assessing the robustness of sisVIVE in a Mendelian randomization study to estimate the causal effect of body mass index on income using multiple SNPs from understanding society. *Statistics in Medicine*, **38**(9), 1529–1542.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica: Journal of the Econometric Society*, **25**(1), 77–83.

Bi, N., Kang, H., and Taylor, J. (2019). Inference after selecting plausibly valid instruments with application to Mendelian randomization. ArXiv:2107.12345.

Bigdeli, T. B., Lee, D., Webb, B. T., Riley, B. P., Vladimirov, V. I., Fanous, A. H., Kendler, K. S., and Bacanu, S.-A. (2016). A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. *Bioinformatics*, **32**(17), 2598–2603.

Boccia, S., Hashibe, M., Gallì, P., De Feo, E., Asakage, T., Hashimoto, T., Hiraki, A., Katoh, T., Nomura, T., Yokoyama, A., *et al.* (2009). Aldehyde dehydrogenase 2 and head and neck cancer: a meta-analysis implementing a Mendelian randomization approach. *Cancer Epidemiology and Prevention Biomarkers*, **18**(1), 248–254.

Bonilla, C., Gilbert, R., Kemp, J. P., Timpson, N. J., Evans, D. M., Donovan, J. L., Hamdy, F. C., Neal, D. E., Fraser, W. D., and Davey Smith, G. (2013). Using genetic proxies for lifecourse sun exposure to assess the causal relationship of sun exposure with circulating vitamin d and prostate cancer risk. *Cancer Epidemiology, Biomarkers & Prevention*, **22**(4), 597–606.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). 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.

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, **44**(2), 512–525.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016a). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. *International Journal of Epidemiology*, **45**(6), 1961–1974.

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

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

Bowden, J., Hemani, G., and Davey Smith, G. (2018). Invited commentary: detecting individual and global horizontal pleiotropy in Mendelian randomizatio–a job for the humble heterogeneity statistic? *American Journal of Epidemiology*, **187**(12), 2681–2685.

Bowden, J., Del Greco M, F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J., and Davey Smith, G. (2019). Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *International Journal of Epidemiology*, **48**(3), 728–742.

Brion, M.-J. A., Shakhbazov, K., and Visscher, P. M. (2013). Calculating statistical power in Mendelian randomization studies. *International Journal of Epidemiology*, **42**(5), 1497–1501.

Brumpton, B., Sanderson, E., Heilbron, K., Hartwig, F. P., Harrison, S., Vie, G. Å., Cho, Y., Howe, L. D., Hughes, A., Boomsma, D. I., *et al.* (2020). Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications*, **11**(1), 3519.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., *et al.* (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, **47**(11), 1236–1241.

Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine*, **30**(11), 1312–1323.

Burgess, S. and Thompson, S. G. (2013). Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology*, **42**(4), 1134–1144.

Burgess, S., Thompson, S. G., and CRP CHD Genetics Collaboration (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, **40**(3), 755–764.

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

Burgess, S., Davies, N. M., and Thompson, S. G. (2016). Bias due to participant overlap in two-sample Mendelian randomization. *Genetic Epidemiology*, **40**(7), 597–608.

Burgess, S., Foley, C. N., Allara, E., Staley, J. R., and Howson, J. M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nature Communications*, **11**(1), 376.

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., Hartwig, F. P., Holmes, M. V., Minelli, C., Relton, C. L., *et al.* (2023). Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome Open Research*, **4**(186).

Cholesterol Treatment Trialists' (CTT) Collaborators (2005). Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *The Lancet*, **366**(9493), 1267–1278.

Cholesterol Treatment Trialists' (CTT) Collaborators (2012). The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *The Lancet*, (9841), 581–591.

Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., *et al.* (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics*, **45**(11), 1345–1352.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association*, **104**(487), 1015–1028.

Elsworth, B. L., Lyon, M. S., Alexander, T., Liu, Y., Matthews, P., Hallett, J., Bates, P., Palmer, T., Haberland, V., Davey Smith, G., *et al.* (2020). The MRC IEU OpenGWAS data infrastructure. BioRxiv:2008.244293.

Faye, L. L., Sun, L., Dimitromanolakis, A., and Bull, S. B. (2011). A flexible genome-wide bootstrap method that accounts for ranking and threshold-selection bias in GWAS interpretation and replication study design. *Statistics in Medicine*, **30**(15), 1898–1912.

Ferguson, J. P., Cho, J. H., Yang, C., and Zhao, H. (2013). Empirical Bayes correction for the Winner's Curse in genetic association studies. *Genetic Epidemiology*, **37**(1), 60–68.

Forde, A., Hemani, G., and Ferguson, J. (2023). Review and further developments in statistical corrections for Winner's curse in genetic association studies. *PLoS Genetics*, **19**(9), e1010546.

Fuller, W. (1977). Some properties of a modification of the limited information estimator. *Econometrica: Journal of the Econometric Society*, **45**(4), 939–953.

Gao, C., Patel, C. J., Michailidou, K., Peters, U., Gong, J., Schildkraut, J., Schumacher, F. R., Zheng, W., Boffetta, P., and Stucker, I. (2016). Mendelian randomization study of adiposity-related traits and risk of breast, ovarian, prostate, lung and colorectal cancer. *International Journal of Epidemiology*, **45**(3), 896–908.

Garner, C. (2007). Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology*, **31**(4), 288–295.

Ghosh, A., Zou, F., and Wright, F. A. (2008). Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *The American Journal of Human Genetics*, **82**(5), 1064–1074.

Gill, D., Georgakis, M. K., Walker, V. M., Schmidt, A. F., Gkatzionis, A., Freitag, D. F., Finan, C., Hingorani, A. D., Howson, J. M., Burgess, S., *et al.* (2021). Mendelian randomization for studying the effects of perturbing drug targets. *Wellcome Open Research*, **6**(16).

Gkatzionis, A. and Burgess, S. (2019). Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*, **48**(3), 691–701.

Göring, H. H., Terwilliger, J. D., and Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *The American Journal of Human Genetics*, **69**(6), 1357–1369.

Gray, R. and Wheatley, K. (1991). How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplantation*, **7**(Suppl 3), 9–12.

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., Ip, H. F., Marioni, R. E., McIntosh, A. M., Deary, I. J., *et al.* (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*, **3**(5), 513–525.

Hahn, J., Hausman, J., and Kuersteiner, G. (2004). Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal*, **7**(1), 272–306.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, **50**(4), 1029–1054.

Hartwig, F. P., Bowden, J., Loret de Mola, C., Tovo-Rodrigues, L., Davey Smith, G., and Horta, B. L. (2016). Body mass index and psychiatric disorders: a Mendelian randomization study. *Scientific Reports*, **6**(1), 1–11.

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

Hartwig, F. P., Davies, N. M., and Davey Smith, G. (2018). Bias in Mendelian randomization due to assortative mating. *Genetic Epidemiology*, **42**(7), 608–620.

He, C., Zhang, M., Li, J., Wang, Y., Chen, L., Qi, B., Wen, J., Yang, J., Lin, S., Liu, D., *et al.* (2022). Novel insights into the consequences of obesity: a phenotype-wide Mendelian randomization study. *European Journal of Human Genetics*, **30**(5), 540–546.

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., *et al.* (2018). The MR-Base platform supports systematic causal inference across the human phenome. *Elife*, **7**, e34408.

Higgins, J. P., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *The BMJ*, **327**(7414), 557–560.

Holmes, M. V., Asselbergs, F. W., Palmer, T. M., Drenos, F., Lanktree, M. B., Nelson, C. P., Dale, C. E., Padmanabhan, S., Finan, C., Swerdlow, D. I., *et al.* (2015). Mendelian randomization of blood lipids for coronary heart disease. *European Heart Journal*, **36**(9), 539–550.

Hubert, H. B., Feinleib, M., McNamara, P. M., and Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. *Circulation*, **67**(5), 968–977.

Jarvis, D., Mitchell, J. S., Law, P. J., Palin, K., Tuupanen, S., Gylfe, A., Hänninen, U. A., Cajuso, T., Tanskanen, T., and Kondelin, J. (2016). Mendelian randomisation analysis strongly implicates adiposity with risk of developing colorectal cancer. *British Journal of Cancer*, **115**(2), 266–272.

Jeffries, N. O. (2007). Multiple comparisons distortions of parameter estimates. *Biostatistics*, **8**(2), 500–504.

Jiang, T., Gill, D., Butterworth, A. S., and Burgess, S. (2023). An empirical investigation into the impact of winner's curse on estimates from Mendelian randomization. *International Journal of Epidemiology*, **52**(4), 1209–1219.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, **111**(513), 132–144.

Katan, M. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *The Lancet*, **327**(8479), 507–508.

Kho, P.-F., Amant, F., Annibali, D., Ashton, K., Attia, J., Auer, P. L., Beckmann, M. W., Black, A., Brinton, L., and Buchanan, D. D. (2021). Mendelian randomization analyses suggest a role for cholesterol in the development of endometrial cancer. *International Journal of Cancer*, **148**(2), 307–319.

Koller, M. and Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis*, **55**(8), 2504–2515.

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., *et al.* (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, **45**(12), 1452–1458.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, **27**(8), 1133–1163.

LeBlanc, M., Zuber, V., Thompson, W. K., Andreassen, O. A., Schizophrenia and Bipolar Disorder Working Groups of the Psychiatric Genomics Consortium, Frigessi, A., and Andreassen, B. K. (2018). A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. *BMC Genomics*, **19**(1), 1–15.

Lee, J. J., McGue, M., Iacono, W. G., and Chow, C. C. (2018). The accuracy of LD score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genetic Epidemiology*, **42**(8), 783–795.

Lewis, S. J. and Davey Smith, G. (2005). Alcohol, ALDH2, and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomization approach. *Cancer Epidemiology Biomarkers & Prevention*, **14**(8), 1967–1971.

Lin, D.-Y. and Sullivan, P. F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. *The American Journal of Human Genetics*, **85**(6), 862–872.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., *et al.* (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, **518**(7538), 197–206.

Markozannes, G., Kanellopoulou, A., Dimopoulou, O., Kosmidis, D., Zhang, X., Wang, L., Theodoratou, E., Gill, D., Burgess, S., and Tsilidis, K. K. (2022). Systematic review of Mendelian randomization studies on risk of cancer. *BMC Medicine*, **20**(1), 41.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., and Klungel, O. H. (2006). Instrumental variables: application and limitations. *Epidemiology*, **17**(3), 260–267.

Mounier, N. and Kutalik, Z. (2023). Bias correction for inverse variance weighting Mendelian randomization. *Genetic Epidemiology*, **47**(4), 314–331.

Nimptsch, K., Song, M., Aleksandrova, K., Katsoulis, M., Freisling, H., Jenab, M., Gunter, M. J., Tsilidis, K. K., Weiderpass, E., and Bueno-De-Mesquita, H. B. (2017). Genetic variation in the ADIPOQ gene, adiponectin concentrations and risk of colorectal cancer: a Mendelian randomization analysis using data from three large cohort studies. *European Journal of Epidemiology*, **32**(5), 419–430.

Niu, W., Pang, Q., Lin, T., Wang, Z., Zhang, J., Tai, M., Zhang, L., Zhang, L., Gu, M., and Liu, C. (2016). A causal role of genetically elevated circulating interleukin-10 in the development of digestive cancers: evidence from Mendelian randomization analysis based on 29,307 subjects. *Medicine*, **95**(7), e2799.

Ong, J.-S., Cuellar-Partida, G., Lu, Y., Australian Ovarian Cancer Study and Fasching, P. A., Hein, A., Burghaus, S., Beckmann, M. W., *et al.* (2016). Association of vitamin D levels and risk of ovarian cancer: a Mendelian randomization study. *International Journal of Epidemiology*, **45**(5), 1619–1630.

Orho-Melander, M., Hindy, G., Borgquist, S., Schulz, C.-A., Manjer, J., Melander, O., and Stocks, T. (2018). Blood lipid genetic scores, the HMGCR gene and cancer risk: a Mendelian randomization study. *International Journal of Epidemiology*, **47**(2), 495–505.

Ozaki, K., Ohnishi, Y., Iida, A., Sekine, A., Yamada, R., Tsunoda, T., Sato, H., Sato, H., Hori, M., Nakamura, Y., *et al.* (2002). Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, **32**(4), 650–654.

Painter, J. N., O'Mara, T. A., Marquart, L., Webb, P. M., Attia, J., Medland, S. E., Cheng, T., Dennis, J., Holliday, E. G., and McEvoy, M. (2016). Genetic risk score Mendelian randomization shows that obesity measured as body mass index, but not waist: hip ratio, is causal for endometrial cancer. *Cancer Epidemiology, Biomarkers & Prevention*, **25**(11), 1503–1510.

Pantelis, C., Papadimitriou, G. N., Papiol, S., Parkhomenko, E., Pato, M. T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D. O., Pietiläinen, O., *et al.* (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, **511**(7510), 421–427.

Papadimitriou, N., Dimou, N., Gill, D., Tzoulaki, I., Murphy, N., Riboli, E., Lewis, S. J., Martin, R. M., Gunter, M. J., and Tsilidis, K. K. (2021). Genetically predicted circulating concentrations of micronutrients and risk of breast cancer: a Mendelian randomization study. *International Journal of Cancer*, **148**(3), 646–653.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**(4), 669–688.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pickrell, J. K., Berisa, T., Liu, J. Z., Ségurel, L., Tung, J. Y., and Hinds, D. A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, **48**(7), 709–717.

Pierce, B. L. and Burgess, S. (2013). Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, **178**(7), 1177–1184.

Pierce, B. L. and VanderWeele, T. J. (2012). The effect of non-differential measurement error on bias, precision and power in Mendelian randomization studies. *International Journal of Epidemiology*, **41**(5), 1383–1393.

Pierce, B. L., Kraft, P., and Zhang, C. (2018). Mendelian randomization studies of cancer risk: a literature review. *Current Epidemiology Reports*, **5**(2), 184–196.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**(7), 459–463.

Qi, G. and Chatterjee, N. (2019). Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nature Communications*, **10**(1), 1941.

Rees, J. M., Wood, A. M., Dudbridge, F., and Burgess, S. (2019). Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PloS One*, **14**(9), e0222362.

Reiersøl, O. (1945). *Confluence Analysis by Means of Instrumental Sets of Variables.* Ph.D. thesis, Stockholms Högskola.

Richmond, R. C., Davey Smith, G., Ness, A. R., den Hoed, M., McMahon, G., and Timpson, N. J. (2014). Assessing causality in the association between child adiposity and physical activity levels: a Mendelian randomization analysis. *PLoS Medicine*, **11**(3), e1001618.

Sadreev, I. I., Elsworth, B. L., Mitchell, R. E., Paternoster, L., Sanderson, E., Davies, N. M., Millard, L. A., Davey Smith, G., Haycock, P. C., Bowden, J., *et al.* (2021). Navigating sample overlap, winner's curse and weak instrument bias in Mendelian randomization studies using the UK Biobank. MedRxiv:2021.21259622.

Schwartz, G. G., Olsson, A. G., Abt, M., Ballantyne, C. M., Barter, P. J., Brumm, J., Chaitman, B. R., Holme, I. M., Kallend, D., Leiter, L. A., *et al.* (2012). Effects of dalcetrapib in patients with a recent acute coronary syndrome. *New England Journal of Medicine*, **367**(22), 2089–2099.

Sheehan, N. A., Didelez, V., Burton, P. R., and Tobin, M. D. (2008). Mendelian randomisation and causal inference in observational epidemiology. *PLoS Medicine*, **5**(8), e177.

Shepardson, N. E., Shankar, G. M., and Selkoe, D. J. (2011a). Cholesterol level and statin use in Alzheimer disease: I. Review of epidemiological and preclinical studies. *Archives of Neurology*, **68**(10), 1239–1244.

Shepardson, N. E., Shankar, G. M., and Selkoe, D. J. (2011b). Cholesterol level and statin use in Alzheimer disease: II. Review of human trials and recommendations. *Archives of Neurology*, **68**(11), 1385–1392.

Shi, J., Wu, L., Zheng, W., Wen, W., Wang, S., Shu, X., Long, J., Shen, C.-Y., Wu, P.-E., and Saloustros, E. (2018). Genetic evidence for the association between schizophrenia and breast cancer. *Journal of Psychiatry and Brain Science*, **3**(4), 7.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*, **89**(5), 607–618.

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

Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, **65**(3), 557–586.

Staley, J. R., Blackshaw, J., Kamat, M. A., Ellis, S., Surendran, P., Sun, B. B., Paul, D. S., Freitag, D., Burgess, S., Danesh, J., *et al.* (2016). PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics*, **32**(20), 3207–3209.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, **20**(4), 518–529.

Sun, L. and Bull, S. B. (2005). Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology*, **28**(4), 352–367.

Sun, L., Dimitromanolakis, A., Faye, L. L., Paterson, A. D., Waggott, D., and Bull, S. B. (2011). BR-squared: a practical solution to the winner's curse in genome-wide scans. *Human Genetics*, **129**(5), 545–552.

Taylor, A. E., Martin, R. M., Geybels, M. S., Stanford, J. L., Shui, I., Eeles, R., Easton, D., Kote-Jarai, Z., Al Olama, A. A., Benlloch, S., *et al.* (2017). Investigating the possible causal role of coffee consumption with prostate cancer risk and progression using Mendelian randomization analysis. *International Journal of Cancer*, **140**(2), 322–328.

The Emerging Risk Factors Collaboration (2009). Major lipids, apolipoproteins, and risk of vascular disease. *Journal of the American Medical Association*, **302**(18), 1993–2000.

The FIELD Study Investigators (2005). Effects of long-term fenofibrate therapy on cardiovascular events in 9795 people with type 2 diabetes mellitus (the FIELD study): randomised controlled trial. *The Lancet*, **366**(9500), 1849–1861.

The Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium (2012). The interleukin-6 receptor as a target for prevention of coronary heart disease: a Mendelian randomisation analysis. *The Lancet*, **379**(9822), 1214–1224.

Thrift, A. P., Gong, J., Peters, U., Chang-Claude, J., Rudolph, A., Slattery, M. L., Chan, A. T., Locke, A. E., Kahali, B., and Justice, A. E. (2015). Mendelian randomization study of body mass index and colorectal cancer risk. *Cancer Epidemiology, Biomarkers & Prevention*, **24**(7), 1024–1031.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Tjønneland, A., Grønbæk, M., Stripp, C., and Overvad, K. (1999). Wine intake and diet in a random sample of 48763 Danish men and women. *The American Journal of Clinical Nutrition*, **69**(1), 49–54.

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*, **50**(5), 693–698.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, **54**(3), 426–482.

Wang, J., Wang, H., Chen, Y., Hao, P., and Zhang, Y. (2011). Alcohol ingestion and colorectal neoplasia: a meta-analysis based on a Mendelian randomization approach. *Colorectal Disease*, **13**(5), e71–e78.

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., *et al.* (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, **45**(11), 1274.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, **114**(527), 1339–1350.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wright, P. (1928). *The Tariff on Animal and Vegetable Oils*. Macmillan Company.

Xiao, R. and Boehnke, M. (2009). Quantifying and correcting for the winner's curse in genetic association studies. *Genetic Epidemiology*, **33**(5), 453–462.

Xu, L., Craiu, R. V., and Sun, L. (2011). Bayesian methods to overcome the winner's curse in genetic studies. *The Annals of Applied Statistics*, **5**(1), 201–231.

Yang, Z., Schooling, C. M., and Kwok, M. K. (2021). Credible Mendelian randomization studies in the presence of selection bias using control exposures. *Frontiers in Genetics*, **12**, 729326.

Yuan, S., Carter, P., Vithayathil, M., Kar, S., Giovannucci, E., Mason, A. M., Burgess, S., and Larsson, S. C. (2020a). Iron status and cancer risk in UK Biobank: a two-sample Mendelian randomization study. *Nutrients*, **12**(2), 526.

Yuan, S., Kar, S., Carter, P., Vithayathil, M., Mason, A. M., Burgess, S., and Larsson, S. C. (2020b). Is type 2 diabetes causally associated with cancer risk? Evidence from a two-sample Mendelian randomization study. *Diabetes*, **69**(7), 1588–1596.

Zhao, Q., Chen, Y., Wang, J., and Small, D. S. (2019). Powerful three-sample genome-wide design and robust statistical inference in summary-data mendelian randomization. *International Journal of Epidemiology*, **48**(5), 1478–1492.

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

Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, **9**(4), 621–634.

Zhou, Y.-H. and Wright, F. A. (2016). The projack: a resampling approach to correct for ranking bias in high-throughput studies. *Biostatistics*, **17**(1), 54–64.

Zöllner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, **80**(4), 605–615.

Zou, L., Guo, H., and Berzuini, C. (2020). Overlapping-sample Mendelian randomisation with multiple exposures: a Bayesian approach. *BMC Medical Research Methodology*, **20**(1), 1–15.