Statistical models for Mendelian randomization analysis using summary-level data

Okezie Uche-Ikonne

Dissertation for PhD

Department of Mathematics and Statistics

Lancaster University

Lancaster, UK

February 4, 2022

Abstract

Mendelian randomization (MR) is a method that uses genetic variants as instrument variables to investigate causality in epidemiology. The application of MR has increased over the years due to genotype-exposure and genotype-disease estimates being published in large genome-wide association studies (GWAS). This research investigates statistical models using GWAS estimates.

To increase the application of Bayesian models in MR, an R package `mrbayes`, which implements univariate and multivariate Bayesian estimation for commonly used two-sample MR estimators, specifically; the inverse variance weighted (IVW), MR-Egger, and radial MR-Egger models. The thesis investigated the use of multivariate Bayesian models with hierarchical priors (BayesLasso, Horseshoe, and Horseshoe+) that account for high-throughput data. Simulations showed these models produced consistent estimates in the presence of pleiotropy and invalid instruments. This thesis also investigated weighted and conditional quantile estimators. Quantile models were shown to produce less bias estimates in simulations.

This thesis has described and reviewed the MR approach and then developed and assessed Bayesian methods for genotype summary level data for application in MR analyses. This research shows how prior distributions can be used to make MR models more robust to the standard IV assumptions.

Contents

List of Tables

List of Figures

Acknowledgements

The completion of this thesis is the result of a collaborative effort involving a number of different contributors. First and foremost, I would like to express my gratitude to the Tertiary Education Trust Fund (TETFUND), which sponsored my PhD programme through Abia State University. I am thankful for the opportunity that has been provided to me to learn and make a positive contribution to society.

I would like to express my gratitude to my supervisory team, Tom Palmer, Frank Dondelinger, and Brian Francis, for making it easy to adjust to the programme and for facilitating the transition of supervision when they were leaving Lancaster University. I am especially grateful to Tom for accepting me as a student and a friend; the scientific and social discussions that have taken place have been extremely beneficial. Tom, I am extremely grateful to you for your mentorship and genuine enthusiasm you showed for me to complete my PhD programme despite numerous obstacles. I look forward to future collaborations with you and look forward to learning more from you.

It has been a privilege to meet and work with so many wonderful people in the Lancaster community, including colleagues from B18 and other Nigerian researchers. Every meeting with each of you helped me to grow as a person, and for that I am grateful to each and everyone of you. I would like to express my gratitude to my friends and colleagues in Nigeria for their indirect support and motivation. I would like to express my gratitude to Ure for being such an excellent companion, and I look forward to sharing many more memories with you in the future.

Finally, and perhaps most importantly, I am grateful to my father, mother, and Chike for many reasons. For inspiring me to begin and complete this programme, for providing moral support when I encountered obstacles, and for helping providing me with the motivation to see it through to the end. To express my gratitude, words fail me. As actor Michael J Fox once said, *"Family is not an important thing; it is everything."* My family has been everything to me, and for that I am grateful beyond measure for such blessings.

Declaration

I declare that this thesis was written entirely by me and has not been submitted to any other institution for the award of any other degree.

Chapter 3 has been published as Uche-Ikonne O, Dondelinger F, Palmer T. Software Application Profile: Bayesian estimation of inverse variance weighted and MR-Egger models for two-sample Mendelian randomization studies—mrbayes, International Journal of Epidemiology, 2020, **50**(1), 43–49. The pdf of the paper is included in the Addendum.

My coauthors served in a supervisory capacity, I conceptualised the main research ideas, formulated the statistical models, and wrote the statistical software code.

Okezie Uche-Ikonne

Chapter 1

Introduction to Mendelian randomization

1.1 Background

Epidemiology is the study of the relationships between exposures and disease outcomes. Although work related to epidemiology has identified possible exposures, there are limitations that make it difficult to distinguish between correlation and causation. These limitations may be due to confounding variables or reverse causation. The randomized controlled trial study design (RCT) is the gold standard for drug and clinical studies. However, RCTs also have certain limitations especially relating to time, expenses, and ethics. Due to merits of lower costs and less time, an epidemiological study can inform drug trials targeting a particular disease. Examples of such cases are: an observational study concluded an inverse association between Vitamin-C and coronary heart disease (Khaw et al., 2001), however an RCT showed a null estimate (Baigent et al., 2005). Similarly, observational studies recommended that hormone-replacement therapy would reduce cardiovascular mortality and breast cancer,

however, an RCT did not support this, but rather showed increased mortality (Banks et al., 2004). A high profile clinical study found that selenium supplements had no effect on prostate cancer despite substantial evidence suggesting otherwise from epidemiological studies (Lippman et al., 2009).

Researchers have proposed that genetic studies can aid in establishing a valid investigation towards identifying cause and effect relationships in healthcare. The method proposed is Mendelian randomization (MR) which uses genetic variants as instruments to infer causal effects (Davey Smith & Ebrahim, 2003). The MR approach was conceptualized from an early paper by Katan (1986), which was the first occurence of using genetic variants as instruments (Thomas et al., 2007). Katan's idea was motivated by observational studies that indicated an association between low cholestrol level and increased cancer rates (Elwood, 2017); which implied that the association may be causal with a decrease in cholesterol triggering an increase or decrease in the risk of cancer. However, the study was limited by confounding factors (e.g., dietary factors), Katan proposed to compare cancer risks in people with different polymorphisms of the apolipoprotein E gene (APOE). The study mitigated unmeasured confounders as individuals with E2 allele have lower levels of cholesterol. The term "Mendelian randomization" was used by Gray & Wheatley (1991), however the application was different from epidemiological applications (Wheatley & Gray, 2004).

The underlying principles of the MR approach come from Mendel's laws which were published by Lock et al. (1916). Mendel's first law states that two alleles separate in equal numbers of the germ cells within the reproductive stage. The more important law within MR studies is Mendel's second law which is the law of independent assortment. This indicates the independence of the genetic variants except those in linkage disequilibrium, the term is explained in section 1.3.2.

1.1.1 Epidemiological terminology

This section introduces some concepts relating to epidemiology. We assume a study to estimate the causal effect of an exposure X on a disease outcome Y. In epidemiological studies it is important to consider different types of variables; including confounders, colliders, and mediators. Relationships between variables in statistics can be graphically represented in causal path diagrams also known as direct acyclic graphs (DAGs). Causal diagrams were introduced in robotics and formalised by Pearl (2009), and they were formally introduced to epidemiology by Greenland et al. (1999). The graphs are connected by arrows also known as an arc, the points in the graph representing the variables are nodes. A causal path in the graph is a directed path from one node to another; they can be traced through a sequence of single arrows entering and leaving a node. A backdoor pathway between an exposure and outcome is a pathway which begins with an arrow pointing towards the exposure and ends with an arrow pointing into the outcome, as per the Backdoor Criterion due to Pearl (2009). Using this terminology confounding is the existence of an open backdoor pathway between an exposure and an outcome. And hence a confounder is any variable on an open backdoor pathway that when adjusted for would block the backdoor pathway. Our aim in observational studies is therefore to adjust for a sufficient set of confounders in order to block all open backdoor pathways between the exposure and outcome. A collider is a variable that is caused by the exposure and outcome independently, controlling for a collider would bias the estimate (Rothman et al., 2008). Hence, in observational studies our aim is not to adjust for colliders. A mediator is a variable that lies on the causal pathway between the exposure and outcome. Adjusting for a mediator partitions a causal effect into a direct and indirect effect. Hence, in observational studies we take care to adjust for mediators only if we wish to estimate specific direct and indirect effects. Figure 1.1 gives a graphical representation where Z is used as a confounder, collider and mediator.

Figure 1.1: Scenarios showing the variable Z as a confounder (left), mediator (middle) and collider (right) in a causal diagram.

1.1.2 Genetics and GWAS studies

This section introduces some genetic terms relating to MR study. The human genome comprises of 23 pairs of chromosomes consisting of 22 autosomal pairs and a pair of sex chromosomes. There are approximately 3×10^9 base pairs of DNA in the human genome. A deoxyribonucliec acid (DNA) sequence is made up of four nucleotide bases: A, C, G, and T. A SNP is defined as a variation in which more than 1% of a population which does not carry the same nucleotide at a specific position in the DNA sequence. Although a specific SNP may not cause a disorder, some SNPs are linked to specific diseases, e.g, Mendelian traits such as Huntingdon's chorea. These associations enable scientists to search for SNPs in order to assess an individual's genetic general tendency to develop a disease. Furthermore, if certain SNPs are known to be associated with a trait, scientists may examine DNA stretches near these SNPs in an attempt to identify the gene or genes responsible for the trait.

Individuals with two copies of a similar allele are considered homozygous in a genetic locus, whereas people with two divergent alleles are called heterozygous. Given a

common allele is given as a the risk allele as A, the three related genotypes are homozygous (aa), heterozygotes (Aa) and the rare homozygotes (AA). The frequency of these related genotypes is described by Hardy-Weinberg equilibrium (Hardy et al., 1908). Mendelian traits are defined dominant when a single copy of the mutant (risk) allele is sufficient to cause the disease, traits are defined recessive when only those individuals with two copies of the risk allele have the disease. It is also possible to inherit traits as co-dominant, which defines a relationship in which the phenotypes induced by each allele manifest when both alleles are present.

Research by Pauling et al. (1949) observed mutations from specific genes in sickle-cell patients which led to change of haemoglobin in red blood cells leading to the connection between genetics and diseases.

Genome Wide Association Studies (GWAS) use SNP arrays to collect data to find out the specific variants associated with common complex traits. Analyses are conducted to check how likely a variant is related to a trait, a p-value of 5×10^{-8} indicates significance, adjusted for multiple testing, that a variant is associated with a trait. There is an archive of GWAS studies that offer summary-level datasets (Akiyama et al., 2017; Sudlow et al., 2015).

For the data from GWAS to be utilized in an MR study, the genotypes of the controls of the study ought to be in Hardy-Weinberg equilibrium (Weinberg, 1908). Hardy-Weinberg equilibrium shows in the controls that the data is a representative sample of the population in order to reliably conclude gene-disease associations in the genetic association studies (Salanti et al., 2005). With the help of GWAS, where lots of genetic variants are tested with exposures, this has led to the study of polygenic and multifactorial disorders where researchers extend the investigation of a disease from single gene mutation towards multiple gene variation and environmental factors.

1.1.3 Concept of causality

A major objective of health studies are to differentiate between causation and association because an exposure might be associated to a disease outcome, but an intervention that affects exposure will not affect the disease unless the association is causal (Hernán, 2004; Sheehan & Didelez, 2020). The association can be notated in a conditional probability $P(Y = y|X = x)$ which shows the distribution of the outcome Y is described by the observation of an exposure $X = x$.

We illustrate an example that reflects association is not causation using the notation of Pearl (2009). The notation $(\text{do}(X = x))$ represents an intervention setting X to x. A hypothetical example could be a binary exposure variable indicating having either stained teeth or not and a disease outcome for coronary heart diseases (CHD). This is because stained teeth is informative of smoking which is causal to coronary heart disease. $P(Y = y|X = x)$ describes how coronary heart diseases can be predicted from teeth inspection. However, an intervention on the stained teeth $(P(Y = y|\text{do}(X = x))$ from a dentist will not have an effect on heart failure.

There are many different causal effect estimands we can target. Three common ones are the average causal effect (ACE), causal risk ratio (CRR), and causal odds ratio (COR). The ACE describes the average change of the outcome from comparing different settings of the exposure, often a 1 unit difference. CRR and COR are causal estimands for different values of the exposure, again often in terms of a 1 unit difference in the exposure, on the risk ratio and odds ratio scales respectively.

1.2 Instrumental variables

Instrumental variable (IV) analysis is a method proposed and primarily developed within the field of econometrics which was later introduced to epidemiology by

Greenland (2000). Within health research, two of the types of instrumental variables that have been used are: those that are controlled and randomized by the researcher (common within RCT) and those that is randomised by nature, applications to MR fall within the second category. Instrumental variable estimators target causal effect estimands, and hence if the instrumental variable assumptions are met in an observational study we can draw causal conclusions. Using the example of linear regression, the goal of an instrumental variable analysis is to resolve the bias in ordinary least squares parameter estimates caused by the inclusion of covariates associated with the error term; this bias is known in econometrics as endogeneity and as confounding in epidemiology. Figure 1.2 represents the causal DAG for an MR analysis which encodes the core instrumental variable conditions described in a later section.

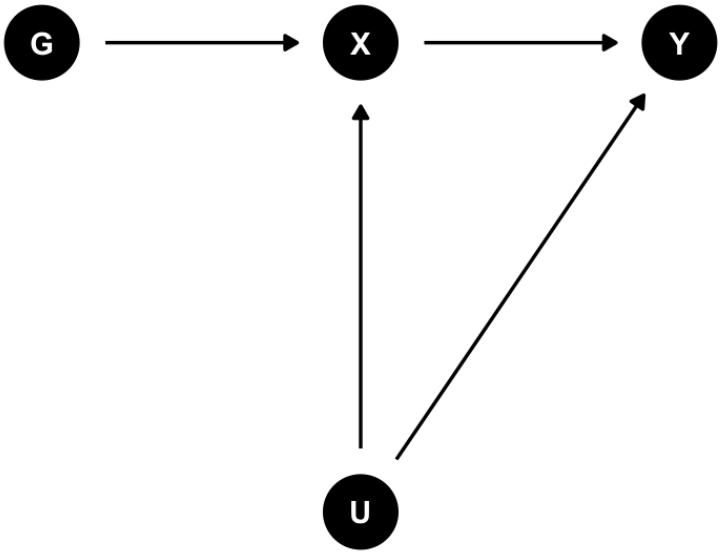

Figure 1.2: Causal DAG representing an MR analysis

1.2.1 Instrumental variable assumptions

The assumptions/core conditions for instrumental variable analysis are; from figure 1.2, parameters (G, U) represent the genotype and unmeasured confounder respectively, X represents phenotype/exposure and Y represents the outcome. G is an instrumental variable for the causal effect of X on Y if;

- It is associated with the phenotype of interest; G is directly associated with X.
- It is independent of the confounders between the exposure and outcome; G is marginally independent of U
- It is independent of the outcome given the phenotype and confounders; G is conditionally independent of Y in the presence of X and U.

In order to estimate the causal effect of X on Y a structural assumption is additionally required (Sheehan & Didelez, 2020).

- The distributions of the parameters representing the genotype, confounders and the conditional distribution of the outcome given the exposure and confounders remain the same regardless of how the instruments affect the exposure (naturally or by intervention)

These assumptions are not fully testable using observational study data. The association between the instrument/s and exposure is of course testable. However, for the second condition, this is not fully testable because of course all possible confounders will not have been measured in any given study.

1.3 Limitations to Mendelian randomization

Some of the limitations to MR analyses are outlined with examples in the following sections.

1.3.1 Population stratification

Population stratification is a limitation to a MR study, this occurs when the population investigated has well defined sub-groups. If the frequency of the genetic variant and the distribution of exposure varies in distinct sub-populations, a misleading link between the variant and the exposure will be produced due to sub-population differences, not the genetic variant's influence. Violations can also occur if there is a continuous variation in the population's structure rather than the unique sub-populations (Burgess & Thompson, 2015). Population stratification can have a confounding effect because different sub-groups may have different risk factors within a sample. Population stratification can be controlled by stratified analysis (Cardon & Palmer, 2003). However, stratification becomes difficult to control when there is admixture of samples. It is preferable to apply MR analysis within populations of homogenous origins (Davey Smith, 2006). Similar to Didelez & Sheehan (2007), figure 1.3 explains the effects of population stratification.

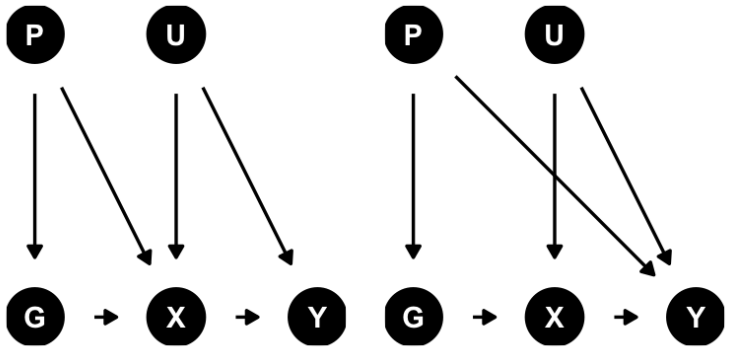

Figure 1.3: DAGs representing population stratification (P). The figure on the left side shows P as a confounder between the G and X. The figure on the right shows P as a confounder between G and Y

1.3.2 Linkage disequilibrium

The term "linkage disequilibrium" refers to a state that differs from the hypothetical scenario in which all loci are completely independent.

Linkage disequilibrium can have both favourable and unfavourable consequences in an MR analysis. It can be favourable if data is not available on a genotype in question but instead on a genotype in high linkage disequilibrium with it, which can then be used as the IV in the analysis. However, linkage disequilibrium can have the unfavourable consequence of accidentally including genetic variants as IVs that are correlated with additional exposures which also affect the outcome, which violates the IV assumptions.

Linkage disequilibrium can be mitigated by systematic testing of the interaction of known confounders with the measured genetic variant (Burgess & Thompson, 2015). The differences in patterns of linkage disequilibrium between populations may partly account for differing estimates for gene-disease association studies (Little & Khoury, 2003). The effects of the conditions from linkage disequilibrium are described in DAGs by Didelez & Sheehan (2007) denoted in figure 1.4. From figure 1.4, the variable

$G2$ describes the genetic variant in linkage disequilibrium, $G1$ describes the genetic variant used for instrument. The variables U, X and Y represent the confounders, exposure of interest and the disease outcome respectively. In the presence of linkage disequilibrium, ($G1$ and $G2$), if $G2$ is independently associated with confounders then it violates IV assumption 2, the assumption of no pleiotropy is additionally violated if the variant $G2$ is independently associated with the outcome of interest (Y).

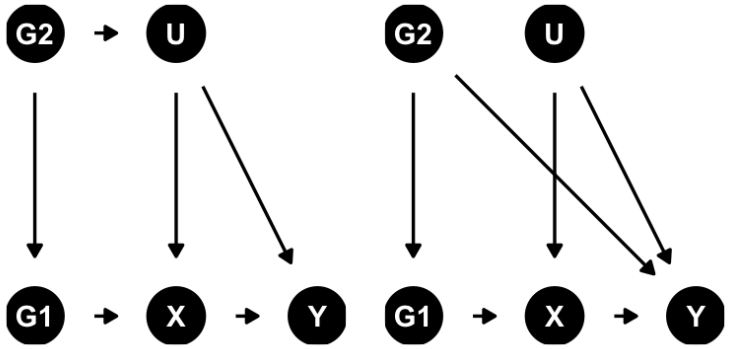

Figure 1.4: Examples of how linkage disequilibrium can violate the MR assumptions.

1.3.3 Canalization

Canalization is the ability of a population to produce the same type of phenotype regardless of the variability from its genotype (Debat & David, 2001). This occurs when an individual adapts to a genetic change in such a way the effect is reduced or absent (Burgess & Thompson, 2015). Canalization is evident in studies where genes are rendered inactive in an organism; the study subjects can develop a mechanism that compensates for the inactive gene through other biological pathways which may have downstream effects. This can become an issue in MR studies if the different levels of genetic variants differ not only from the exposure of interest but through canalization. The goal of MR, on the other hand, is to analyse the causal influence of the (non-genetic) exposure, not just to characterise the effects of genetic change. MR estimates may be unrepresentative of clinical interventions on the exposure undertaken

in a mature cohort if there is significant canalization (Burgess & Thompson, 2015).

1.3.4 Pleiotropy

Pleiotropy is defined as a genetic variant related to multiple exposures, if such variants are used as instruments it violates the second and third IV assumptions. Pleiotropy is divided into two categories; Horizontal pleiotropy means if the genetic variant is associated with an exposure that is not on the causal pathway to the exposure of interest. Vertical pleiotropy means the genetic variant is associated with another variable which is on the same causal pathway as the exposure of interest (mediation), this pleiotropy is usually inconsequential in MR studies. An example relates to the FTO gene which is related to food satisfaction (Wardle et al., 2008); a study on the causal effect of body mass index on a disease outcome, a variant from the FTO gene can be used as an instrumental variable due the relationship of food satisfaction to body mass index are on the same pathway, which is a good application of vertical pleiotropy. However, if the variant in the FTO gene affects another exposure which is not on the same causal pathway as BMI it may lead to misleading conclusions (horizontal pleiotropy) (Burgess & Thompson, 2015). Figure 1.5 graphically describes horizontal and vertical pleiotropy. Pleiotropy is a common issue within MR due to numerous metabolite-related instruments, especially horizontal pleiotropy (left of figure 1.5) which can severely bias MR studies. This has led to the development of different study designs and methods to mitigate pleiotropy which are later discussed in section 2.5.

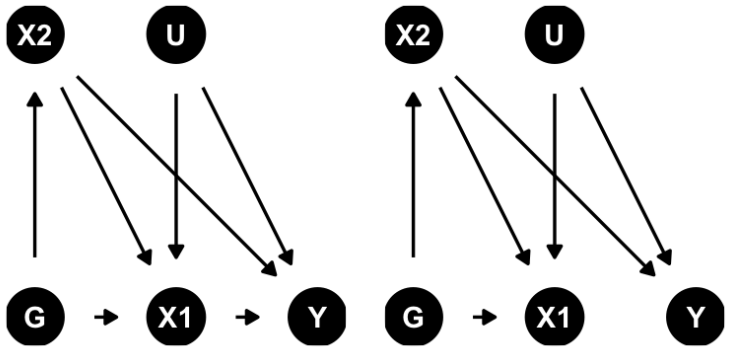

Figure 1.5: DAGs representing pleiotropy. $X1$ represents the exposure of interest; $X2$ represents another exposure; G represents the instruments; U represent confounders; Y represents the disease outcome. The left panel represents vertical pleiotropy and the right panel represents horizontal pleiotropy.

1.3.5 Genetic heterogeneity

An MR study requires the researcher to have substantial knowledge of the genetic variants and their functions (Vineis, 2004). Brennan (2004) noted that a problem in MR is having a vast knowledge of the alleles used as instruments and disease pathways. Biological knowledge is highly relevant as it can help to identify the pathways of the gene-disease association to confirm if it is only through the phenotype of interest. A strong gene-phenotype association would reduce the likelihood of the genotype being a 'weak' instrument, which would bias the estimator for instrumental variable (IV) analysis (Staiger & Stock, 1994). Also to conduct an MR study requires consideration of biological processes, an example is the segregation distortion of genes within the locus. The process occurs when the distribution of alleles in a particular locus differs from the surviving offsprings as a result of selective survival between conception and birth. The phenomenom affects MR analysis as it can induce a correlation between the genotype and confounders which was unlikely at population level (Davey Smith & Ebrahim, 2008). The existence of selective survival due to the genetic variant

can be a reference to the possibility that the differing genotype outcomes may bias genetic association studies. If the effect of a variant on a phenotype is dependent on the parent allele is known as parent-of-origin effect. Genetic heterogeneity occurs in MR studies due to multiple genes causally associated with a phenotype. However, genetic heterogeneity has no effect on the core assumptions if the genes do not affect confounder or disease risk through another phenotype.

1.4 Motivation of study

The rise in GWAS studies has contributed to the increase of summary-level datasets for MR studies. The development of robust statistical models is essential, with newly developed models suitable for use in summary-level data settings. This research is aimed at expanding statistical methods that range from software implementation, theoretical methods, and data application within summary level data analysis.

1.5 Outline of thesis

The thesis is outlined as: Chapter 2 is a review of the literature on the historical context of MR as an analytical method in epidemiology, the study designs and statistical models used for estimating causal effects and robust statistical methods to mitigate effects of weak instruments in MR. In Chapter 3 I introduce an convenient application of univariate and multivariate Bayesian analysis for IVW and MR-Egger models including the radial MR-Egger model. Chapter 4 extends the multivariate MR model by investigating Bayesian hierarchical priors in summary-level data with sparsity. Chapter 5 investigates the weighted and conditional percentile models in summary-level data and introduces sensitivity analysis models. Chapter 6 discusses the overall findings of the thesis and possible areas for further research.

Chapter 2

Literature review

This chapter explains the various criteria to take into account when performing MR analysis, including statistical methods for causal inference. There is also review of the recent development of MR in new study designs and methods of sensitivity analysis.

2.1 Sample size

A requirement for studies in MR is large sample sizes, which has led to suggestions and application of meta-analysis in MR studies (Lawlor et al., 2008). The reasons for large sample size are because genetic effects explain a small proportion of the variance in exposures (Frayling et al., 2007). Earlier studies into genetic associations were not replicated due to lack of statistical power and publication bias (Little & Khoury, 2003), until a study by Hirschhorn et al. (2002) led to the recognition of reproducible studies relating to gene disease associations. A study by Danesh et al. (2008) showed they had 80% power to detect an odds ratio of 1.2 with a minor allele frequency of 5% for their sample size of 37,000 cases and 120,000 controls. That recognition brought about the availability of large GWAS consortia. Examples are biobanks in the UK (Sudlow

et al., 2015) and China (Chen et al., 2011), allowing researchers to apply MR studies on large numbers of participants. GWAS consortia can also provide an opportunity for GWAS studies to be harmonized and made more easily available to researchers through web-based platforms (Hemani et al., 2016).

2.2 Selection of genetic variants

The selection of genetic variants as instruments is an important decision in an MR study. This section explains how genetic variants for MR research are chosen.

The choice of genetic variants is either from a single gene region or from multiple regions of the genome (a polygenic analysis). Selecting variants from a single gene region offers the advantage of specificity, which means that if a gene region has a specific biological link to the exposure, the MR study will be more accurate in determining the causal role of that exposure. However, several robust statistical analysis methods are not possible, as they assume independence of variants. Using polygenic variants has a major advantage to explain additional variability in the exposure which will improve the statistical power of an analysis (Brion et al., 2013).

There are two strategies for identifying variants in a polygenic analysis which include a biologically driven method and a statistically driven approach. The overall decision about which variants to include could include features from both methods (Burgess et al., 2019). A biological approach to genetic variant selection would include variants from regions with a biological relationship to the exposure of interest, however biological knowledge is not perfect. A common statistical method for choosing genetic variants is to include all variants that are linked with the exposure of interest at a certain level of statistical significance (usually, a genome-wide significance threshold, such as p-value $< 5 \times 10^{-8}$) in the analysis. The p-value selection is based on the data in which genetic associations with exposure are estimated, but it can lead to genetic

associations being overestimated. Furthermore, weak instrument bias is increased when genetic variants are chosen based on their associations with the exposure in the data under investigation. The bias in the estimates is in the direction of the observational association in a one-sample setting, and in the direction of the null in a two-sample setting (Burgess et al., 2011). Bias can be avoided by selecting genetic variants from a different dataset (Zhao et al., 2019). Cis-variants are the most credible instruments for MR studies because they have biological relevance to the exposure (e.g., molecular phenotypes such as gene expression and DNA methylation) (Burgess et al., 2019). However, with multifactorial exposures such as body mass index or blood pressure, it is not possible to find a cis-variant so polygenic analysis is necessary. There is no simple technique for deciding which genetic variants to include in a study. Burgess et al. (2019) suggests a balance between adding fewer variants which potentially have insufficient power and including more variants which potentially include pleiotropic variants.

2.3 Study designs

The progression of MR analysis has led to various study designs to improve statistical validity, some of these designs are applied independently or combined. They are discussed in this section.

2.3.1 One-sample MR

The study design assesses exposure and outcome variables from a single sample, allowing for the investigation of the causal effect in a single population sample. A one-sample study design is used for the majority of individual-level data collection. The study design has the benefit of being simple to collect data and having fewer influences from the demographic groups. However, chance variation limits the study

design; that is if the instrument-exposure relationship is weak (Burgess et al., 2019).

2.3.2 Two-sample MR

Results from GWAS studies have given researchers wider access to large samples of summary-level data leading to MR studies conducted from different samples. Two-sample MR analysis is conducted as the instrument-exposure and instrument-outcome associations are obtained from different samples which do not overlap each other (Burgess et al., 2015c). From figure 2.1 the DAG shows the exposure variable from sample 1 (X_1) and the outcome variable from sample 2 (Y_2). Two-sample MR is appealing due to the difficulty of acquiring sample measures for exposure and outcome from the same sample set. The quality of the datasets are dependent on individual studies, but due to their use of independent replication samples, GWAS results are robust to replication. Two-sample MR methods are available in several highly used R packages, for example TwoSampleMR applies the methods to data from the MR-Base database of GWAS results (Hemani et al., 2016). Hartwig et al. (2016) discussed several ways to simply improve the quality of two-sample MR studies, including ensuring that the genotypes are harmonized with respect to the coding of their alleles in the two-samples.

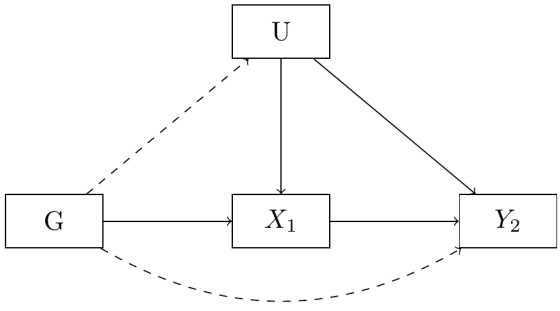

Figure 2.1: DAG representing Two-sample MR. X_1 represents exposure from sample 1 and Y_2 represents outcome from sample 2.

2.3.3 Multivariable MR

The aetiology of certain outcomes is normally characterized by multiple exposures. This has led to the development of designs appropriate for the application of MR models including multiple exposures. The Multivariable MR (MVMR) study design incorporates multiple genetic variants that are associated with multiple exposures (Burgess et al., 2015b). From figure 2.2, the variables X and Z represent the exposures with their respective instruments (G_X and G_Z) causal to the disease outcome Y, with confounder U. MVMR study simultaneously estimate the 'direct' ($X \rightarrow Y$ and $Z \rightarrow Y$) effect of each exposure on the outcome. The design of the study also investigates the 'indirect' effects ($X \rightarrow Z \rightarrow Y$) of the exposure by extending the paradigm to the causal networks. And indeed the direct and indirect effects can be combined into a total effect if this is the aim of the investigator.

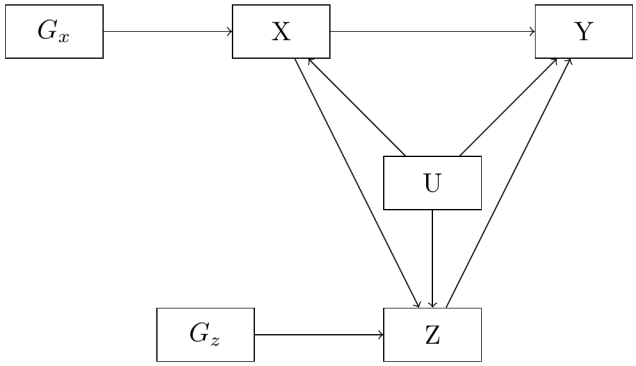

Figure 2.2: DAG representing a multivariate MR study design, where G_x and G_z represent the instruments for X and Z, X represents the exposure, Z a second exposure which is also mediator between X and the outcome, and Y the outcome.

2.3.4 Two-step MR

Increased interest in the role of epigenetics in estimating causal effects of environmental exposures on outcomes has led to DNA methylation mediators being incorporated into the MR framework. Two-step MR study design is a framework to understand

the role of these genetic markers (Relton & Davey Smith, 2012). Through integrative genomics. The study involves two steps; the first is the use of instruments to estimate the causal effect of the exposure on the mediator, while the second involves the using instruments to estimate the causal effect of the mediator on the outcome summarized in figure 2.3. The steps are then combined to investigate evidence of a causal effect of the exposures through the proposed mediator on the outcome.

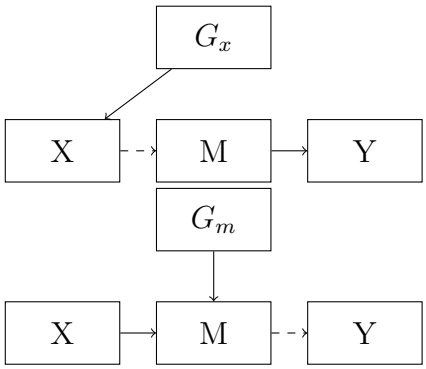

Figure 2.3: DAG representing the two-step study design. The dashed arrows represent the causal effect estimated in each step. G_x is the instrument for the exposure, X the exposure, M the mediator, Y the outcome, and G_m the instrument for the mediator.

2.3.5 Bi-directional MR

Bi-directional MR is a study design depicted in figure 2.4 when the investigator is not sure whether the exposure (X_A) is the cause of the outcome (X_B) (first direction) or the outcome (X_B) of the exposure (X_A) (second direction). The instruments generated for both the exposure (G_A) and outcome (G_B) are independent of each other. The study design is relevant as it dissects the direction of causality to establish higher confidence in MR studies and to prevent reverse causation. Some examples are, the direction of causality between serum uric acid (SUA) and adiposity (Lyngdoh et al., 2012); and the association of adiposity and inflammation (Welsh et al., 2010).

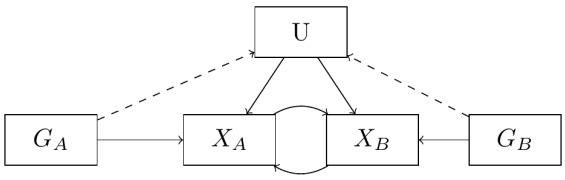

Figure 2.4: DAG for a bi-directional MR design.

2.3.6 Relationship between the study designs

The research designs are used for specific cases of MR analysis, but they share some features. In this section, I discuss some of the characteristics and how they relate to each other.

The use of two-sample MR has risen in MR analysis through time due to the multiple advantages highlighted previously. For most of the study designs, the exposure and disease outcome originate from independent samples, thus two-sample MR integrates with most of the study designs mentioned earlier. The two-step and multivariate MR study designs assess both direct and indirect effects; it is reasonable to suggest that two-step MR is a multivariate MR study design when researching with two exposures and the exposure associations are derived from separate instruments. MVMR study designs, on the other hand, can additionally estimate the total effect, incorporate a large number of exposures, and generate instrument-exposure estimates using similar instruments or independent sets of instruments.

It is worth noting that study designs are evolving depending on the analysis; for example, Zhao et al. (2019) introduced the 3-sample MR study design, which evolved from the 2-sample MR study design, in which the genetic instruments are generated from a different sample study. Also Burgess et al. (2015a) introduced network MR to untangle the relationship between exposures and complex traits, which is related to the MVMR study design. However, there is a reasoning that bidirectional studies can

evolve into Network MR in the presence of multiple exposures.

2.4 Methods of estimating causal effects in MR analyses

This section examines the methods used to estimate the causal effect in MR, with the statistical properties underlying them.

2.4.1 Ratio method

The ratio of coefficients method, which is the ratio of instrument-outcome and instrument-exposure association, is a simple approach for estimating the causal effect (Wald, 1940). The ratio approach is accurate on the basis of the linearity assumption between the exposure and the outcome, as a result of which the ratio estimate is the linear IV average effect (Didelez et al., 2010). The ratio method is applied to a single IV, if there is more than one IV, we can report an IV estimate for each variant or we can combine the variants as a single IV in an allele score. In MR studies SNPs which are instruments are represented in genetic subgroups which are either in a dichotomous (biallelic) or polytomous form (diallelic). When the IVs are dichotomous (i.e $G = 0, 1$), we measure the causal effect as the average difference between the subgroups indicated below

$$\frac{\Delta Y}{\Delta X} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0}.$$

Alternatively, when SNPs are polytomous we can assume the association between genetic variant and exposure is proportional to the number of alleles in the variant, the causal effect is derived below

$$\hat{\beta} = \frac{\hat{\beta}_{Y|G}}{\hat{\beta}_{X|G}}.$$

When the denominator is closer to zero it is an indication of weak instrument which creates a large variance and unstable estimator. When calculating the ratio estimate, one does not need the full data specification of variables (G, X, Y), summary-level data include instrument-exposure associations and instrument-outcome associations.

From the IV assumptions, a higher association between instrument and exposure (G-X) indicates a strong instrument. Within the linear assumption the strength of the instruments would provide more information on the causal relationship through smaller standard errors and confidence intervals. There is a functional relationship between instrument and confounders, a strong instrument indicates a low effect from the confounders.

The unobserved confounders κ_u have an effect on the outcome (Y) while the causal effect (β) indicates the average causal effect of one unit increase in the exposure X

$$E(Y|X = x, U = u) = \beta x + \kappa_u.$$

Due to this effect, we get a biased estimate if we regress Y directly on X without adjusting for U, hence we perform instrumental variable estimation instead.

To estimate the causal effect within MR, we need to test the instrument-outcome association $(G - Y)$. The rationale follows from IV assumption (IV1), when testing for a null hypothesis of no association the genetic variants would be independent of each other. However there can be other reasons for no association for example low power or a hidden interaction with unobserved groups. The test is to provide evidence and it is indicated as good practice without parametric assumptions (Burgess & Thompson,

2012).

2.4.2 Two-stage methods

Two-stage models follow a two-step procedure and are commonly applied to individual level data. The model incorporates multiple instruments and requires full data from the variables. In a sample population indexed by $i = 1, 2, 3, ..., N$ and a group of instruments J; the first step below shows regression of the exposure (X) on the genetic variant (G)

$$E(X_i|G) = \gamma_0 + \sum_j^J \gamma_j G_{ij} + \varepsilon_{x_i}.$$

The second step shows the outcome variables regresses on the predicted values of the exposure

$$E(Y_i|G) = \alpha + \beta\hat{X} + \varepsilon_{y_i}.$$

The method is asymptotically unbiased for the average causal effect when applied with a single IV, if the instrument is weak, but it is at risk to small sample bias (Bound et al., 1995). However, including more instruments can reduce the bias, additional instruments in the two-stage least squares (TSLS) model estimate the weighted average, the weights are estimated from the first stage regression. Although the TSLS approach derives the correct point estimate, the standard error of the second stage does not compensate for the variability of the first stage, and so the adjustment of replacing $\hat{X}$ with X in the calculation of the second stage residual variance is made in instrumental variable software.

The TSLS model assumes the outcome variable is from a continuous distribution,

the counterpart model when the binary outcomes are distributed (e.g. case-control study) is called two stage predictor substitution (TSPS). Estimates from such an approach would be too precise because they do not account for variability. However, the over-precision would be small if the standard error of the first-stage coefficients were low. Two-stage regression methods for binary outcomes received some criticism due to possible correlation of instruments and outcome (Angrist & Pischke, 2008). The interpretation and validity of the results of the TSPS models depends on the collapsibility. An estimate is collapsible if the constant value in each strata is equal to the marginal value of the analysis, there are implications in MR (see Page 59, Burgess & Thompson (2015)), which has led to adjusted two stage methods. Despite these implications, TSPS is a valid test of the null hypothesis (Vansteelandt et al., 2011). An adjusted two-stage method has been suggested, in which the residuals of the first-stage regression of exposure on IV are included in the second stage, the approach is generally referred to as the two-stage residual inclusion estimator (TSRI). The second stage is derived below by Terza et al. (2008)

$$E(Y_i|G) = \alpha + \beta\hat{R} + \varepsilon_{y_i}.$$

The variable $\hat{R}|G = X - \hat{X}|G$ denotes the residual of the first stage regression. The IV estimate is numerically closer to the conditional log odds ratio, making it recommended for logistic regression (Palmer et al., 2008).

2.4.3 Likelihood methods

The likelihood is derived by including the effects of confounders to correlate the error terms in the two-stages of estimation. This can be formulated as the exposure and outcome following a bivariate normal distribution denoted below

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim MVN \left(\begin{pmatrix} \mu_{x_i} \\ \mu_{y_i} \end{pmatrix}, \Sigma \right).$$

The maximum likelihood estimate of the causal parameters are calculated simultaneously which is referred to as a full information maximum likelihood approach denoted below

$$\mu_{x_i} = \alpha_0 + \sum_k \alpha_k g_{ik}$$
$$\mu_{y_i} = \beta_0 + \beta_1 \mu_{x_i}.$$

A drawback to this approach is it estimates every parameter in each of the equations. An alternative approach is maximizing the likelihood and profiling out other parameters except the causal effect parameter; this approach is the limited information maximum likelihood estimator (LIML). LIML has similar attributes to TSLS making them sensitive to model misspecification and heteroscedasticity. Hahn et al. (2004) notes that the LIML approach has a limitation of undefined moments for any number of instruments. However the median of the estimate distribution is close to unbiased even in the presence of weak instruments; making the approach appealing (Angrist & Pischke, 2008). An alternative likelihood approach can be applied from the Bayesian framework.

The model shows the measured exposure and outcome $(X_i, Y_i)^T$ follow a bivariate distribution. The mean of the exposure distribution is a linear function of the genetic variants and the outcome is a linear function of the mean exposure where β_1 is the causal estimate (Jones et al., 2012). Using uninformative priors, estimation of the posterior distribution is applied in a Markov chain Monte Carlo framework (MCMC),

from which the mean or median of the distribution can be interpreted as the point estimate while the credible interval can be described as the confidence interval. The Bayesian approach has no distributional assumption for the posterior making them robust to weak instruments (Burgess & Thompson, 2012).

2.4.4 Semi-parametric methods

Semi-parametric methods provide non-parametric assumptions in parametric methods with the aim of being robust to model misspecification. Within MR, parametric assumptions are made towards the equation relating the outcome and exposure while error terms are assumed non-parametric. The generalised method of moments is a semi-parametric estimator, the approach is intended as a more flexible TSLS that deals with issues of non-homogeneous error terms and non-linearity (Johnston et al., 2008). Structural mean models are another semi-parametric approach used in instrumental variable estimation (Robins, 1986).

2.5 Robust methods in MR

Pleiotropic effects arise when a gene has multiple effects on different phenotypes, which invalidates the IV assumption. Multiple genetic variants, which are widely used in MR studies, increase the likelihood of pleiotropy. This section will introduce most of the robust estimators (i.e., methods for sensitivity analysis). The instrument violations that these models address are also discussed in this section.

2.5.1 Location parameter for ratio estimates

The standard IVW model is extended to a multiplicative random effects, the model accounts for pleiotropy through an overdispersion parameter incorporated into the variance (Bowden et al., 2017). The IVW model is extended to Egger regression

(MR-Egger) by including an intercept that accounts for directional pleiotropy; the model is based on the assumption of instrument strength independent of direct effects (InSIDE) (Bowden et al., 2015). An extension of MR-Egger through a radial formulation (Radial-MR), in this case the intercept does not include weights, the model is identified as a direct sub-model to IVW model. The Radial-MR model is primarily used as an aid for visualization of valid and invalid instruments however they also produce consistent estimates within InSIDE assumption. The weighted median approach takes the median of the ratio estimates after they have been given probabilistic weights that are inversely proportional to their variances, the model is based on the assumption of the weights derived from greater or equal to 50% valid IVs (Bowden et al., 2016b). Similar weights are used to select the mode of smoothed density function of the ratio estimates. The approach offers reliable estimates provided that the largest single instrument causal effect is from a valid instrument (Hartwig et al., 2017), the modal based estimator was extended through penalized weights (Burgess et al., 2018).

MR analyses with multiple instruments increase the power of an analysis. The ratio estimates are instrument specific with the assumption of constant effect sizes, uneven sizes of ratio estimates indicate weak or invalid instruments we can identify the variants that have similar ratio estimates and obtain the causal effect from the subset (Burgess et al., 2013). As a result, a test of heterogeneity within the ratio estimates has also been defined by Bowden et al. (2019), with the inclusion of extra assumptions of instrument subsets either the majority or mode (in a plurality case) of the ratio estimates to obtain the estimate of the causal effect. While heterogeneity identifies invalid instruments, it does not guarantee that valid instruments exist within a homogeneous subset. An alternative to homogeneous subsets of instruments is using a linear regression where the exposure is the predictor to the outcome, the slope is equal to the causal estimate with residuals including the deviation and the intercept which are assumed independent (InSIDE). The model allows the causal effect estimate to be computed even with

invalid instruments given the validity of InSIDE.

Valid genetic instruments should generally have similar estimates (that is homogeneous), however that is not usually the case as evidence of heterogeneity from combined GWAS studies. Bowden et al. (2016a) introduced measures for between-instrument heterogeneity similar to Q-statistics and I^2. The detection of heterogeneity does not always imply pleiotropy, rather it can relate to violation of instrument variable or model assumptions like no measurement error (NOME). Even if the model assumptions are met Q-statistics can still be inflated due to violation of NOME assumption due to large variability from the instrument-exposure association that cannot be ignored. The large variability can lead to violation of InSIDE assumption making the MR-Egger model suspectible to regression dilution bias, which leads to alternative robust meta-analysis procedures like the weighted median and modal based estimator (Bowden et al., 2016b; Hartwig et al., 2017).

2.5.2 Penalized method and outlier detection

Most times, it is not practical to have complete knowledge of the genetic variants, hence penalized models have been proposed by penalizing the effect of the instruments. The choice of the penalty parameter (λ) follows cross-validation similar to the LASSO method by Tibshirani (1996), tuning λ controls the effect on the model, as a high value of the parameter would estimate most instruments as valid instruments and a low value would estimate most of the instruments as invalid. Kang et al. (2016) introduced the LASSO method for penalizing pleiotropic instruments (SiSVive), by extending the conventional TSLS to include the penalty parameter (λ). The model was extended by estimating the penalty parameter by cross-validation through an adaptive LASSO (Windmeijer et al., 2019). These penalized models showed consistent estimates provided that more than 50% of the instruments are valid, however the models are in the form of individual-level data setting. Adaptive LASSO was extended

into a summary-level data setup where the IVW model is extended by including intercept terms that are SNP-specific and penalizing those pleiotropic intercept terms (Burgess et al., 2016). Another robust method (MR-Robust) was introduced which is a combination of method of moments estimation and Tukey's biweight loss function to downweight the effect of outlying instruments (Burgess et al., 2016). MR-Presso is a test of the residual sums to identify horizontal pleitropic outliers (Verbanck et al., 2018). The outlier test assumes $\geq 50\%$ of the instruments are valid and the InSIDE assumption. Zhao et al. (2018) derived a robust adjusted profile score method (MR-RAPs) that models pleiotropic variants through a random effects distribution, the estimates are robust by penalizing outlying variants by either Huber's loss function or Tukey's biweight loss function.

The estimates of the causal effects can be modelled as a mixture distribution of the effect sizes with the assumption that genetic markers are valid instruments. Qi & Chatterjee (2019) measured the causal effects through a spike detection algorithm assuming valid and invalid instruments known as the MR-Mix model. A similar mixture model approach (conmix) models the ratio estimates which is characterized in clusters of valid and invalid IVs with prespecific variance of invalid instruments (Burgess et al., 2020). From the premise that valid and invalid instruments create clusters, Foley et al. (2021) introduced an algorithm to find clusters and identify variants that indicate different causal pathways.

2.5.3 Bayesian modelling approaches

Bayesian modelling approaches can be considered for relaxing the instrumental variable assumptions. For example, the models can include strategies that account for pleiotropic instruments within the likelihood and/or prior distributions. Some strategies include adding shrinkage priors on the pleiotropic effects through the use of hierarchical priors in a fully Bayesian estimator. In a summary level data setting,

Li (2017) introduced a hierarchial shrinkage approach using a "Spike and Slab" prior algorithm to penalize the effects of weak instruments. Berzuini et al. (2020) applied the horseshoe prior on the effects on the genetic variants to account for and penalize the effects of invalid instruments within an individual-level data setting. Thompson et al. (2017) modelled pleiotropic scenarios within the prior distribution and used Bayesian model averaging to provide estimates robust to those pleiotropic scenarios.

Shapland et al. (2019) introduced Bayesian model averaging methods to produce robust estimators with dependent instruments. Shapland et al. (2020) applied another Bayesian model averaging method where the profile likelihood score is used as the basis for the posterior distribution within a two-sample study design. The research by Bucur et al. (2020) introduced a Bayesian model that accounts for invalid instruments and accommodates reverse causation. Zuber et al. (2020) introduced a two-parameter Bayesian modelling averaging approach, which selects risk factors from a high-throughput multivariate study design for summary-level data.

2.6 MR methods for complex traits

This section reviews MR models that estimate the causal effect of molecular phenotypes known as complex traits. A particular problem in this area is pleiotropy, which arises due to the polygenetic nature of these traits.

GWAS have identified numerous variants associated with complex traits (MacArthur et al., 2017). Using results from GWAS studies to identify causal genes prove to be difficult due to most of associated variants in linkage disequilibrium with the causal marker (Flister et al., 2013). This highlights the importance of transcriptome-wide association studies (TWAS) which integrates expression quantitative trait loci (eQTLs) with GWAS studies to explore gene-trait associations (Nica et al., 2010). Trait-associated SNPs are more likely to be linked to gene expression, which implies

that the instrument-exposure relationship could be mediated by gene expression.

The concept of MR analysis can be applied to gene expression levels, the genetic variant is the instrument, the expression is the exposure of interest and the phenotype is the outcome. Zhu et al. (2016) proposed the use of MR analysis, to search for the most functionally relevant genes at the loci identified in GWAS for complex traits. Since, many human complex traits are polygenic in nature, the amount of variance in the phenotype explained by a single genetic variant is likely to be very small. As a result, a very large sample size is required to detect the effect of a gene on a trait using MR analysis. In practice, such large sample sizes are rarely available; however, large amounts of summary-level data from very large-scale GWAS and eQTL studies are available in the public domain, and these data can be used to perform two-sample MR analyses. To test if the effect of the gene expression is mediated through transcription Porcu et al. (2019) proposed the transcriptome-wide MR (TWMR). TWMR uses the multivariate MR technique for estimating the impact of the gene expression adjusted for transcription on a phenotype. These methods have revealed causal gene-trait associations within complex traits.

2.7 Discussion

The principles, assumptions, and limitations of MR theory have been discussed in this review. Specifically, this review discussed how genetic variants are used as instrumental variables in the MR approach to perform causal inference in epidemiology. The review discussed the need for large studies to improve power, due to the fact that genetic effects explain a small proportion of the variance of the exposure. The review has also shown that the use of genotype summary level data is very common in two-sample MR analyses, which can be performed in MR-Base (Hemani et al., 2018). The review also discussed the limitations of MR in terms of linkage disequilibrium and pleiotropy, in

particular, the different types of pleiotropy and their impact on causal effect estimates. The limitations also include population stratification and genetic heterogeneity.

This review also addressed the types of genetic variants that were used as instruments, with the recommendation that polygenic and cis-variants are the most useful types of instruments. We looked at the strategies used for identifying valid instruments for analysis, which included a combination of statistical and biological processes to identify valid instruments for analysis. We also looked at the lower level definitions used in causal inference which underpin the MR approach. Employing instrumental analysis (in this case, MR) helps to mitigate the effects of confounders that would otherwise bias our causal effect estimates.

The review looked at some of the designs that have been introduced to conduct MR studies. The review observed the similarities and differences between study designs. We anticipate that as more data on instrument-exposure and instrument-outcome relationships become available, MR study designs will continue to evolve.

In this review, the estimators used in MR analyses were critically evaluated, these included the ratio, two-stage, likelihood, semi-parametric, and meta-analysis methods. We discussed their properties in terms of their strengths and limitations. We also covered estimators that are robust to some of the limitations of the conventional MR estimators and the additional assumptions that they require.

Finally, this review discussed the application of MR methods for complex traits. One problem in this field is that genetic variants can be in high linkage disequilibrium with one another. Hence, there is interest in this area to determine whether gene expressions are causally related to phenotypes.

Chapter 3

Bayesian estimation of IVW and MR-Egger models

3.1 Introduction

This chapter introduces `mrbayes`, an R package which implements Bayesian estimation of the IVW, MR-Egger, and Radial MR-Egger models. The models are estimated using Markov chain Monte Carlo(MCMC) methods through an R interface to the JAGS and Stan software (using the `rjags` and `rstan` packages) (Plummer, 2018; Stan Development Team, 2018). Our package includes some specified prior distributions; non-informative, weakly informative, a shrinkage prior on the causal effect estimate (Pseudo-Horseshoe prior), and a joint prior on the intercept and causal effect estimate in the MR-Egger and radial MR-Egger models. The package also allows users to specify their own prior distributions using the JAGS software.

The methods implemented in the package include some prior distributions. The estimates of the models are fitted using different prior distributions on example datasets. Further investigations were conducted into the joint prior distribution in

relation to the InSIDE assumption and some strategies are introduced when choosing prior distributions.

3.2 Methods

From the DAG in figure 3.1 the population size is represented by N, J represents the number of instruments used for MR analysis. In the presence of confounding variables (U), the conditional relationship between the phenotype (X) and outcome (Y) variables is represented by,

$$\begin{aligned} U_i|G_{ij} &= \sum_{j=1}^{J} \kappa_j G_{ij} + \epsilon_u \\ X_i|U_i, G_{ij} &= \sum_{j=1}^{J} \phi_j G_{ij} + \delta_x U_i + \epsilon_x \\ Y_i|X_i, U_i, G_{ij} &= \sum_{j=1}^{J} \Delta_j G_{ij} + \beta X_i + \delta_y U_i + \epsilon_y. \end{aligned} \tag{3.1}$$

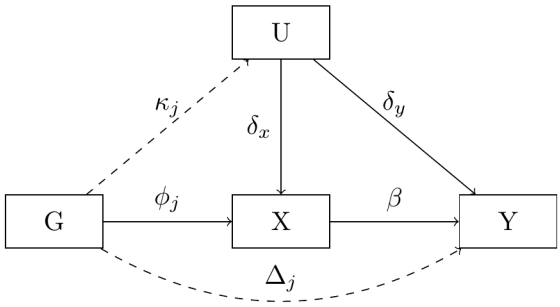

Figure 3.1: Causal directed acyclic graph indicating the instrumental variable assumptions underpinning the Mendelian randomization approach.

The variable G_{ij} denotes the j^{th} genetic instrument for the i^{th} individual in the population, ϕ_j is the parameter for the instrument-phenotype association. The parameter κ_j denotes the pleiotropic (direct) effect of each genetic variant on the outcome. The conditional densities of U, X, and Y denoted below

$$
\begin{aligned}
P(U|G) &\sim N\Big(\kappa_j, \sigma_u^2\Big) \\
P(X|G) &\sim N\Big(\phi_j + \delta_x\kappa_j, \sigma_x^2\Big) \\
P(Y|G) &\sim N\Big(\Delta_j + \delta_y\kappa_j + \beta\Big[\phi_j + \delta_x\kappa_j\Big], \sigma_y^2\Big).
\end{aligned} \tag{3.2}
$$

The variables are extended into summary-level estimates which is a reduced form of (3.2) to obtain estimates and standard errors for instrument-exposure (3.3) and instrument-outcome associations (3.4). The estimate $\hat{\Gamma}_j$ in (3.3) includes the effect of pleiotropy $\alpha_j = \Delta_j + \delta_y\kappa_j$ using valid instruments it can be assumed there is little or no effect of pleiotropy $\alpha_j \approx 0$. The genotype-exposure associations is in a reduced form of $\gamma_j = \phi_j + \delta_x\kappa_j$. and represents the genotype-outcome associations. With the assumption that each instrument has an identical association with the outcome within each sample, it is practical to perform two-sample MR where the genotype-exposure and genotype-outcome variables are obtained from different samples (Bowden et al., 2017), as such we assume that

$$
\hat{\gamma}_j \sim N(\gamma_j, \sigma_{x_j}), \tag{3.3}
$$

and

$$
\hat{\Gamma}_j|\alpha_j, \sigma_{y_j} \sim N(\alpha_j + \beta\gamma_j, \sigma_{y_j}). \tag{3.4}
$$

Using valid instruments, the variable the estimate for β can be represented as a Wald ratio IV estimate

$$
\frac{\hat{\Gamma}_j}{\hat{\gamma}_j} = \beta_j. \tag{3.5}
$$

The inverse variance weighted formula is used to combine the ratio estimates for each instrument with its inverse variance weight of the first order denoted below

$$\frac{\sum_j w_j \hat{\beta}_j}{\sum_j \hat{\gamma}_j} = \beta_{IVW}. \tag{3.6}$$

The IVW estimate also represents the slope of the instrument-outcome associations when regressed to the instrument-exposure associations with no intercept. Including the first order inverse variance (of the instrument-outcome association) weights, where $\sigma^2_{y_j}$ denotes the instrument-outcome variances, we can write the IVW model as

$$\hat{\Gamma}_j = \beta\hat{\gamma}_j + \sigma^2_{y_j}\varepsilon_j; \quad \varepsilon_j \sim N(0,1). \tag{3.7}$$

Note, here the variance of the residuals being constrained to 1 indicates that the IVW model is equivalent to a fixed effect model meta-analysis model.

The MR-Egger model is an extension of IVW model that includes an intercept parameter. The model generates consistent estimates given the InSIDE assumption, from (3.2) InSIDE assumption means $\alpha_j \neq 0$ but $\kappa_j = 0$. The intercept parameter in the MR Egger model represents the average pleiotropic effect. The further the estimate of the mean pleiotropic effect ($\hat{\alpha}$) is from zero, the larger the difference between the IVW estimate and the true causal effect. In the MR-Egger model the variance of the residuals, σ^2, is estimated, which means that this uses a multiplicative random effects meta-analysis model (Higgins & Thompson, 2002). Prior to applying the MR-Egger model the instrument-exposure and instrument-outcome associations must be oriented such that all the instrument-exposures associations are positive

$$\hat{\Gamma}_j = \alpha + \beta\hat{\gamma}_j + \sigma^2_{y_j}\varepsilon_j, \quad \varepsilon_j \sim N(0,\sigma^2). \tag{3.8}$$

The radial MR-Egger model is defined as,

$$\hat{\beta}_j\sqrt{w_j} = \alpha + \beta\sqrt{w_j} + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma^2), \tag{3.9}$$

where the variable w_j denotes the weights and they differ from the MR-Egger model as the weights are applied only to the slope due to the unweighted intercept, it is considered that the IVW model is its sub-model.

Assuming known variance, the likelihood for the MR-Egger estimator follows a univariate Gaussian distribution, again w denotes the weights

$$P(\hat{\Gamma}|\alpha, \beta, \sigma, \hat{\gamma}) = \prod_{j=1}^{J} N(\alpha + \hat{\gamma}_j\beta, \sigma^2 w_j). \tag{3.10}$$

The Bayesian posterior distribution is modelled as;

$$P(\alpha, \beta, \sigma|\hat{\Gamma}_j, \hat{\gamma}_j) \propto P(\hat{\Gamma}_j, \hat{\gamma}_j|\alpha, \beta, \sigma)P(\alpha, \beta, \sigma). \tag{3.11}$$

The default prior distributions for the parameters in the `mrbayes` package are summarised in table 3.1 and discussed in the next section.

Table 3.1: Formula for default prior models in mrbayes. For functions in IVW model, there is no α parameter

Model	Priors
Uninformative Priors	$\alpha \sim N(0, 1000),\ \beta \sim N(0, 1000),\ \sigma \sim U(0.0001, 10)$
Weakly-Informative Priors	$\alpha \sim N(0, 1),\ \beta \sim N(0, 1),\ \sigma \sim U(0.0001, 10)$
Pseudo-Horseshoe Priors	$\alpha \sim N(0, 1),\ \beta \sim C(0, 1),\ \sigma \sim IG(0.5, 0.5)$
Joint Priors	Please see next section

3.3 Prior distributions

The choice of prior distributions is an important factor in Bayesian estimation. This section gives a brief description on the formulation of the different prior distributions included in this package and the JAGS syntax used for implementing them (Plummer, 2018).

3.3.1 Non-informative prior distributions

Non-informative prior distributions are used when there is no prior knowledge about the distribution of a parameter. This type of prior distribution is expected to produce estimates similar to frequentist estimates. There is no "best" choice of non-informative prior but table 3.1 denotes some possible non-informative prior distributions, these have large variances for the average pleiotropic effect (α) and the causal effect (β). Although, an improper prior density was set for the σ, given a large number of instruments ($J > 3$) the prior yields proper posterior densities (Gelman et al., 2006a). In the presence of pleiotropic instruments the use of vague/non-informative prior distributions may lead to estimates with low precision Jones et al. (2012).

3.3.2 Weakly informative prior distributions

The idea of a weakly informative prior, (3.12), is to provide partial information on the parameters to be estimated. Therefore, they are often used when performing regularization. Weakly informative priors could mitigate the effects of winner's curse (when the estimated effect might be exaggerated). These prior distributions are described in below, where the variance is reduced for α and β compared to the non-informative prior distributions,

$$\alpha \sim N(0,1),\ \beta \sim N(0,1),\ \sigma \sim U(0.0001, 10). \tag{3.12}$$

3.3.3 Pseudo-horseshoe prior distribution

The MR-Egger estimator is extended by placing a Cauchy distribution prior on the causal effect $\beta \sim C(0,1)$. The Cauchy distribution was chosen as the prior distribution due to some appealing properties, for example the divergence property of no mean and infinite variances, whereas mode and median which are equal. An investigation into the direction of causality through Bayesian models showed that pleiotropic instruments can give the causal effect a multimodal distribution (Bucur et al., 2020). In the presence of valid instruments $> 50\%$, the divergence property of the Cauchy distribution gives greater weighting to the strong instruments and reduces the effect of outlying instruments. The convergence towards the Gaussian distribution in the presence of a large number of instruments is another useful property of the Cauchy distribution as a shrinkage prior. For efficient mixing and convergence, σ follows an inverse-gamma distribution. The default prior distributions for our `prior = "pseudo"` option, in the `mr_egger_rjags` and `mr_radialegger_rjags` functions,

$$\alpha \sim N(0,1),\ \beta \sim C(0,1),\ \sigma \sim IG(0.5, 0.5). \tag{3.13}$$

3.3.4 Joint prior distribution

A conjugate bivariate normal prior distribution on the slope and intercept in the MR-Egger model has been shown to have good properties (Schmidt & Dudbridge, 2017). α and β follow a bivariate prior distribution,

$$\begin{aligned} \alpha|\sigma^2 &\sim N(\mu_\alpha, \sigma^2\sigma_\alpha) \\ \beta|\sigma^2 &\sim N(\mu_\beta, \sigma^2\sigma_\beta) \\ \sigma^2 &\sim U(1, 10) \\ \text{Cov}(\alpha, \beta|\sigma^2) &= \sigma^2\rho_{\alpha\beta}. \end{aligned} \tag{3.14}$$

Under its accompanying InSIDE assumption, the correlation coefficient can be described as the degree of InSIDE violation when $\sigma_\alpha\sigma_\beta \geq 0$ within the MR-Egger model. The InSIDE assumption is investigated using assumed external information on the values for the hyperparameters denoted below;

$$\begin{aligned} \mu_\alpha, \mu_\beta &= 0 \\ \sigma_\alpha, \sigma_\beta &= 10. \end{aligned} \tag{3.15}$$

I investigate if the magnitude of the correlation coefficient between the intercept and slope of the joint prior distribution influences the estimates, which can help us determine the ideal value of correlation coefficient ρ while conducting an MR analysis. The MR-Egger and radial MR-Egger models are fitted for values of the correlation coefficient (ρ) between $-0.99 \leq \rho \leq 0.99$ under the null and alternative hypothesis. The simulated datasets consist of two-sample study design showing directional pleiotropy when the InSIDE assumption is violated ($\beta = 0.5$).

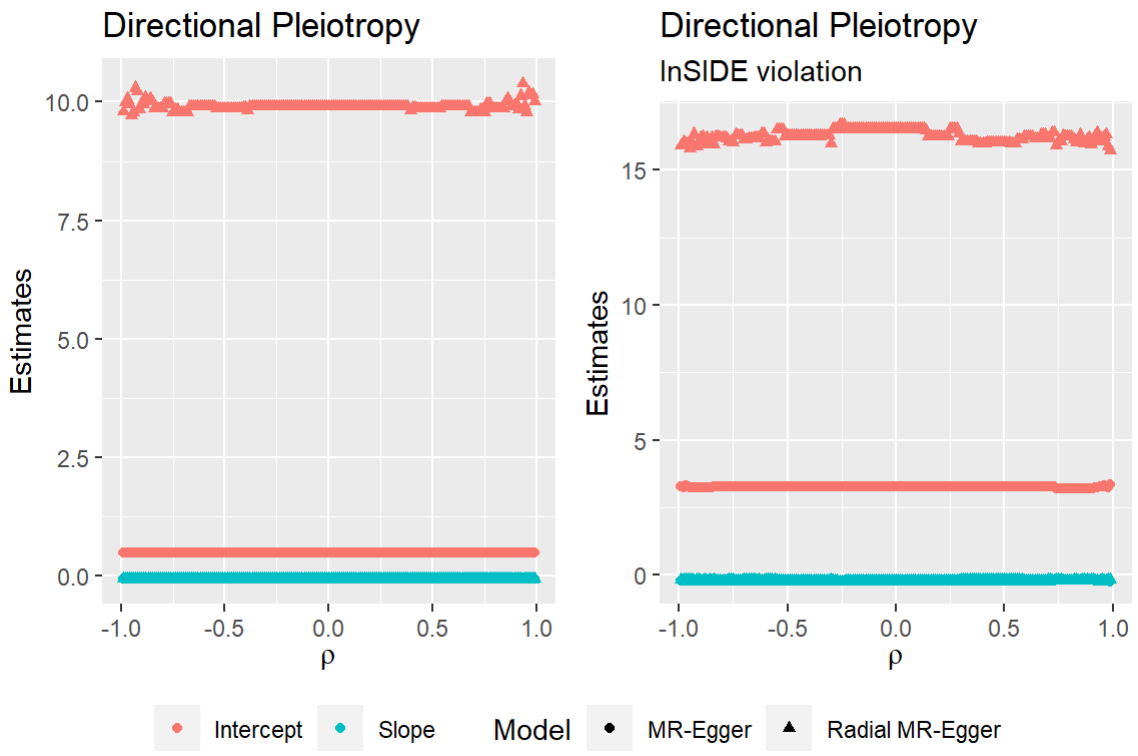

Figure 3.2: Estimates of the causal effect and average pleiotropic effect for different values of ρ in the joint prior distribution.

Figure 3.2 shows the results generated for α and β from the different values of ρ, the values show a similar pattern when the InSIDE assumption is valid or violated. The values of the parameters within the MR-Egger model show no difference when the correlation coefficient changes. The Radial formulation shows a pattern in the intercept parameter where there is a little change when ρ gets closer to ± 1. The simulation illustrates that the correlation coefficient's value has no impact on the causal effect estimate. However, using various distributions of instrument-exposure and instrument-associations in a summary-level data context, more research on the impact of correlation coefficient on the causal effect estimate is needed.

3.4 Implementation

Mrbayes package provides the following functions:

- `mr_format`, a function for setting up the summary-level dataset for analysis.

The functions that use JAGS/Stan software are;

- `mr_ivw_rjags`/`mr_ivw_stan`, a function for estimating causal effects using the Bayesian IVW model, with a choice of prior distributions;
- `mr_egger_rjags`/`mr_egger_stan`, a function for estimating causal effects through the Bayesian MR-Egger model, with a choice of prior distributions;
- `mr_radialegger_rjags`/`mr_radialegger_stan`, a function for performing Bayesian analysis under the radial formulation of MR-Egger.

The package allows users:

- to specify custom prior distributions for the estimate of the causal effect (`betaprior`) and optionally for the residual standard error (`sigmaprior`) for the MR-Egger models (original and radial). This option is only for `_rjags` functions, the prior distributions are written in the JAGS syntax. For more information on how to specify prior distributions see page 34 of JAGS manual; Plummer (2012)
- to choose a random seed for reproducible results and to choose the number of chains for MCMC, each chain should have a different seed;
- to set parameter `rho`, the correlation coefficient between the average pleiotropic effect and causal estimate. This option is only relevant when using the joint prior method;
- to plot the posterior density and investigate the MCMC diagnostics.

The package also includes two summary-level datasets containing:

- 185 SNPs with multiple instrument-phenotype associations for low-density

lipoprotein cholesterol (LDL-c), while the instrument-outcome associations for coronary heart disease (CHD) (Do et al., 2013);

- 14 SNPs with instrument-phenotype associations of body mass index (BMI) and instrument-outcome associations of insulin resistance (Richmond et al., 2017).

3.5 Strategies for choosing priors

Informative prior distributions can help to account for pleiotropy in Bayesian MR analyses (Jones et al., 2012), an approach for informative prior distributions is to use the result from a previous study. Alternatively, for cases where prior estimates cannot be obtained, we can use regularized priors similar to weakly informative prior distributions. I present some strategies when considering informative prior distributions with the emphasis on the slope parameter (which is the causal effect estimate).

The choice of a prior distribution with small standard deviation (e.g. $\beta \sim N(0, 1)$) can be regarded as an ideal option when $\hat{\Gamma}_j$ and $\hat{\gamma}_j$ are standardized which is comparable to the IVW and the original formulation of MR-Egger models. We can set prior distributions for the slope and its standard deviation independently, an example is the normal gamma distribution $\beta \sim N(0, 1/\sigma_\beta)$; $\sigma_\beta \sim G(a, b)$. However, this prior distribution applies to non-standardized error terms similar to the radial MR-Egger. The selection of hyperparameters (a, b) can make the normal-gamma distribution have a similar shape as the Laplace distribution which has stronger regularization. This prior distribution can also be considered a Bayesian version of frequentist LASSO regression (Griffin et al., 2010).

3.6 Example: estimates from some informative prior distributions

The package demonstration using the motivating example by Zhao et al. (2018) which is also the example dataset in `mr.raps` package for estimating the causal effect of body mass index (BMI) on acute ischemic stroke (AIS) Zhao et al. (2018). Two groups of instruments were selected for estimating the causal effect from the summary-level dataset. The first set of instruments were selected from the GWAS p-value threshold as $p \leq 5 \times 10^{-8}$ and the second set included all the instruments. The prior distributions applied in table 3.1 are compared with the frequentist model.

3.6.1 Data description

The data is an excerpt from the `mr.raps` package, details of the instrument-exposure and instrument-outcome associations are within the package documentation. The outcome for this dataset is acute ischemic stroke and the exposure for this analysis is body mass index (BMI). This dataset is created from three genome-wide association studies (GWAS). GWAS on BMI was used for SNP selection by Akiyama et al. (2017), The UK BioBank GWAS of BMI was applied to estimate the SNPs' effect on BMI. The third GWAS study estimates the SNPs' effect on AIS Malik et al. (2018). For the joint prior the correlation between the intercept and slope is assumed 0.5.

Table 3.2: MR estimates from summary-level data (GWAS p-value threshold). CI in the column indicates Confidence/Credible Interval

Method	Model	Coefficient	Estimate	95% CI
Frequentist	IVW	Slope	0.3292	0.1815, 0.4768
Frequentist	MR-Egger	Intercept	0.0047	-0.0057, 0.3423
Frequentist	MR-Egger	Slope	0.3319	0.0393, 0.6245
Frequentist	MR-Egger Radial	Intercept	0.3772	-0.5037, 1.258
Frequentist	MR-Egger Radial	Slope	0.326	0.0147, 0.6373
Weakly Informative	Bayesian IVW	Slope	0.3295	0.1797, 0.4765
Weakly Informative	Bayesian MR-Egger	Intercept	0.0046	-0.0067, 0.0162
Weakly Informative	Bayesian MR-Egger	Slope	0.3329	0.0122, 0.654
Weakly Informative	Bayesian MR-Egger Radial	Intercept	0.3792	-0.4417, 1.2214
Weakly Informative	Bayesian MR-Egger Radial	Slope	0.3266	0.039, 0.6206
Pseudo	Bayesian IVW	Slope	0.3262	0.177, 0.4732
Pseudo	Bayesian MR-Egger	Intercept	0.0046	-0.0066, 0.0156
Pseudo	Bayesian MR-Egger	Slope	0.3175	0.0068, 0.6251
Pseudo	Bayesian MR-Egger Radial	Intercept	0.3763	-0.4997, 1.2425
Pseudo	Bayesian MR-Egger Radial	Slope	0.3123	0.0104, 0.6113
joint	Bayesian MR-Egger	Intercept	0.0045	-0.0033, 0.6585
joint	Bayesian MR-Egger	Slope	0.331	-0.5367, 1.1729
joint	Bayesian MR-Egger Radial	Intercept	0.3454	0.0276, 0.6288
joint	Bayesian MR-Egger Radial	Slope	0.3255	0.1797, 0.4765

The estimates derived from the models are seen in table 3.2 and 3.3 (dataset including all the instruments) figure 3.3 shows a graphical summary. From table 3.2, MR

estimates from the frequentist modes are approximately similar to the estimates from the Bayesian models with weakly informative priors. The slopes show evidence of significance while the intercepts show no evidence of significance. MR estimates from the Bayesian models with pseudo prior distributions have lower estimates this is due to the shrinkage effect of the Cauchy distribution. Similar to the frequentist and weakly informative prior distribution the slope parameter are significant while the intercepts are not significant. MR estimates using the joint prior distribution are similar to frequentist and weakly informative prior distribution. However the estimate of the slope and intercept parameter from the MR-Egger model shows no evidence of significance while the slope and intercept of the radial MR-Egger model indicate significance. The features of the estimates (and significance) in table 3.3 are similar to the 3.2 except for the Bayesian MR-Egger model using joint prior distribution as the slope indicates non-significance while the intercept indicates significance.

Table 3.3: MR Estimates from summary-level dataset when all instruments are included. CI in the column indicates Confidence/Credible Interval

Method	Model	Coefficient	Estimate	95% CI
Frequentist	IVW	Slope	0.3173	0.2117, 0.4229
Frequentist	MR-Egger	Intercept	0.0001	-0.004, 0.3214
Frequentist	MR-Egger	Slope	0.3173	0.0998, 0.5348
Frequentist	MR-Egger Radial	Intercept	0.0015	-0.32, 0.323
Frequentist	MR-Egger Radial	Slope	0.3173	0.0981, 0.5364
Weakly Informative	Bayesian IVW	Slope	0.3175	0.2104, 0.4226
Weakly Informative	Bayesian MR-Egger	Intercept	0.0001	-0.004, 0.0042
Weakly Informative	Bayesian MR-Egger	Slope	0.319	0.1043, 0.5396
Weakly Informative	Bayesian MR-Egger Radial	Intercept	0.0007	-0.3203, 0.3228
Weakly Informative	Bayesian MR-Egger Radial	Slope	0.319	0.1056, 0.5396
Pseudo	Bayesian IVW	Slope	0.316	0.2124, 0.4194
Pseudo	Bayesian MR-Egger	Intercept	0.0001	-0.0039, 0.0042
Pseudo	Bayesian MR-Egger	Slope	0.3095	0.0897, 0.5284
Pseudo	Bayesian MR-Egger Radial	Intercept	0.0025	-0.3105, 0.3208
Pseudo	Bayesian MR-Egger Radial	Slope	0.3094	0.0973, 0.5215
joint	Bayesian MR-Egger	Intercept	0.0001	0.1002, 0.5318
joint	Bayesian MR-Egger	Slope	0.3184	-0.329, 0.3092
joint	Bayesian MR-Egger Radial	Intercept	-0.0013	0.0956, 0.5404
joint	Bayesian MR-Egger Radial	Slope	0.3222	0.2104, 0.4226

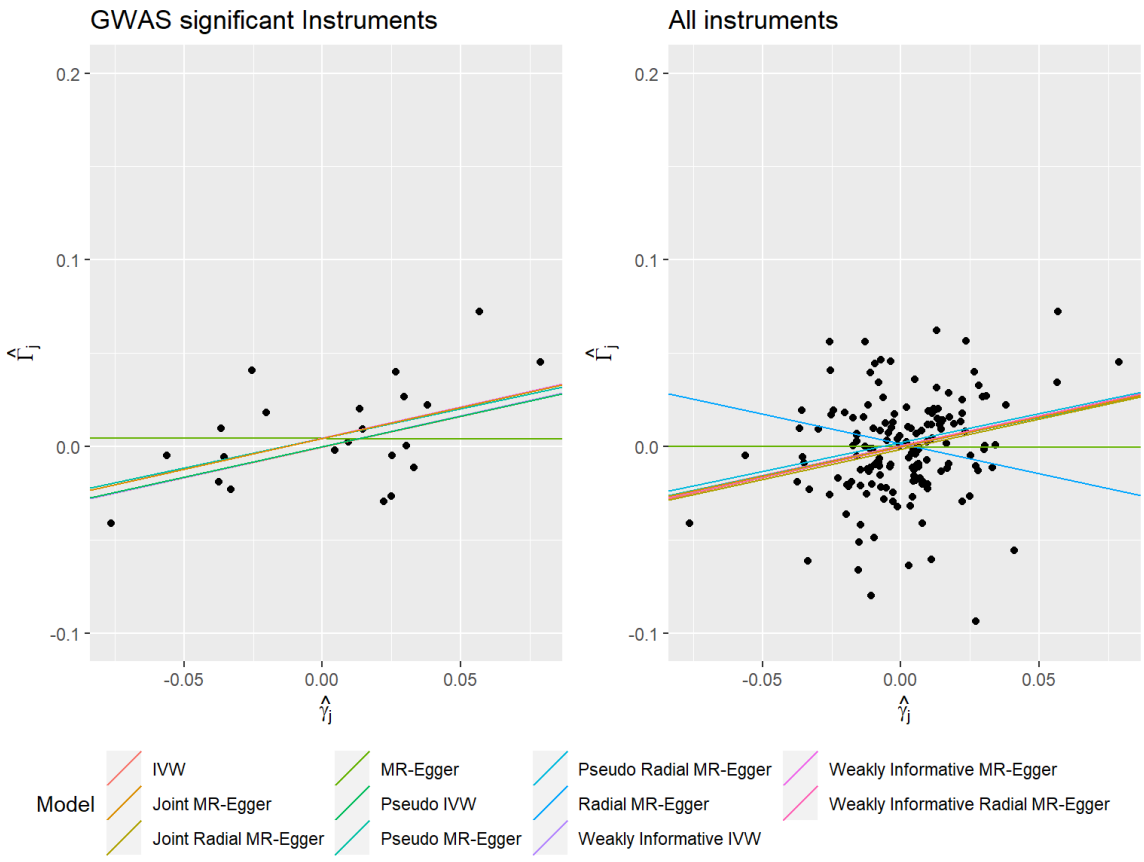

Figure 3.3: Y-axis represents the instrument-outcome associations (acute ischemic stroke) and the x-axis indicates the instrument-exposure associations (body mass index)

3.7 Update including the multivariate IVW and MR-Egger estimators

Bayesian estimation for multivariate IVW and MR-Egger models are included as an update to the package. To investigate the direct causal effects of p exposures, (3.7) extends to a multiple exposure model

$$\hat{\Gamma}_j = \beta_1\hat{\gamma}_{1j} + \beta_2\hat{\gamma}_{2j} + ... + \beta_p\hat{\gamma}_{pj} + \sigma^2_{y_j}\epsilon_j; \quad \epsilon_j \sim N(0,1). \tag{3.16}$$

The MR-Egger model is extended to a multiple exposure model in (3.17), extending (3.16), with the inclusion of an intercept (β_0). We note that for the MVMR-Egger model we have to choose which exposure to orientate the genotype-outcome and genotype-exposure associations. Hence, we can obtain a different causal effect estimate from a model incorporating the same exposures but oriented differently. It is common practice to orient the MVMR-Egger model with respect to the main exposure of interest (Rees et al., 2019)

$$\hat{\Gamma}_j = \beta_0 + \beta_1\hat{\gamma}_{1j} + \beta_2\hat{\gamma}_{2j} + ... + \beta_p\hat{\gamma}_{pj} + \sigma^2_{y_j}\epsilon_j; \quad \epsilon_j \sim N(0, \sigma). \tag{3.17}$$

The distribution of the observed Γ_j and the posterior distribution are given by

$$\Gamma_j|\beta_i, \gamma_{ij}, \sigma^2 \sim N\left(\sum_{i=1}^{p} \gamma_{ij}\beta_i, \sigma^2\right) \tag{3.18}$$

and

$$p(\beta_i, \sigma^2|\Gamma, \gamma) \propto p(\Gamma|\gamma, \beta_i, \sigma^2)p(\beta_i|\sigma^2)p(\sigma^2). \tag{3.19}$$

The likelihood and posterior for MVMR-Egger model would include the intercept parameter in (3.18) and (3.19). The prior distributions are identical to those used in univariate models except the joint prior for the intercept with all of the exposures. The multivariate radial model is not included in this package because there has not been research on selecting which exposures to use in the numerator of the weights, i.e. selecting the parameter γ_j from $w_j = \frac{\gamma_j^2}{\sigma_{y_j}}$.

3.8 Discussion

An R package, `mrbayes`, is presented to perform Bayesian estimation of the IVW and MR-Egger models implemented through the JAGS and Stan software packages. The example demonstrates the use of several different prior distributions to estimate the causal and average pleiotropic effects from these models.

There are several R packages providing functions for MR analyses. The `MendelianRandomization` and `TwoSampleMR` packages implement various two-sample MR methods (Hemani et al., 2018; Yavorska & Burgess, 2017). The `RadialMR` R package implements the radial MR models and visualization of instrument-exposure and instrument-outcome associations through radial plots (Bowden et al., 2018; Spiller & Bowden, 2019). Bayesian methods have not gained popularity in applied studies due to limited availability of user-friendly software (Sheehan & Didelez, 2020). Our package complements previous MR packages by offering a Bayesian perspective. In a Bayesian analysis the prior distributions can have an important impact upon the final parameter estimates. Hence in the `mrbayes` package offers a choice of prior distributions. with the choice of four prior distributions for the causal effect; non-informative, weakly informative, pseudo-horseshoe, and a joint prior distribution for the MR-Egger model's intercept and slope.

Chapter 4

Bayesian hierarchical models in MVMR

4.1 Motivation

Penalized regression is a statistical method commonly used when there are large number of predictors to protect against overfitting (Tibshirani, 1996). Penalized methods include a penalty parameter which is the threshold for the maximum likelihood estimate of each phenotype. Some penalized models in MR analyses have been introduced in section 2.5.2. The aim of this research is to extend the MVMR study design to a Bayesian paradigm using a hierarchical prior distribution applied to high-throughput summary-level data. The prior distributions include: Bayesian Lasso (Park & Casella, 2008), horseshoe (Carvalho et al., 2009) and horseshoe+ priors (Bhadra et al., 2017). The prior distributions will be applied to a two-sample summary level GWAS dataset to investigate the causal effects of metabolites on different outcomes.

4.2 Penalized hierarchical priors applied to MVMR models

Section 3.7 introduced Bayesian estimation in a multivariate MR model. In highthrouput data we assume that there are a large number of exposures, say greater than 10 exposures. The likelihood model in (4.1) includes a parameter λ_i. Therefore, the posterior distribution in (4.2) includes the prior distribution for λ_i conditional on β_i

$$\Gamma_j|\beta_i, \gamma_{ij}, \lambda_i, \sigma^2 \sim N\left(\sum_{i=1}^{p} \gamma_{ij}\beta_i, \lambda_i\sigma^2\right) \tag{4.1}$$

and

$$p(\beta_i, \sigma^2, \lambda_i|\Gamma, \gamma) \propto p(\Gamma|\gamma, \beta_i, \sigma^2)p(\beta_i|\sigma^2, \lambda_i)p(\sigma^2)p(\lambda_i). \tag{4.2}$$

To apply a Bayesian model to a large multivariate dataset, a prior distribution is needed that allows the data to collapse the entire marginal posterior for each phenotype towards either zero or non-zero coefficients. This can be done by assuming a hierarchical model on the parameter of each phenotype which is broken down into stages,

$$\beta_i|\lambda_i \sim p(\beta_i|\lambda_i, \tau)$$
$$\lambda_i|\tau \sim p(\lambda_i|\tau)$$
$$\tau \sim p(\tau).$$

The first prior distribution for each causal effect for each exposure (β_i) is conditioned

with each shrinkage parameter (λ_i), while the distribution of the shrinkage parameter is conditioned on an additional parameter τ. The idea behind the prior distribution is similar to a global-local shrinkage effect, where τ performs a global shrinkage on all the exposures while λ_i is exposure-specific local shrinkage parameter (Polson et al., 2012).

4.2.1 Bayesian Lasso for MVMR

The Lasso model is a form of an L1 regularization method which includes a penalty parameter on the phenotypes (Tibshirani, 1996). The estimate derived from the Lasso applied to the MVMR model is given by the optimization problem

$$\text{argmin} \sum_j \sigma^{-2}(\Gamma_j - \beta_p \hat{\gamma}_{pj})^2 + \lambda \sum_p ||\beta||_1.$$

The principle of regularization methods is to penalise the log-likelihood function with an effective non-decreasing function $\lambda||\beta||$. The shrinkage parameter λ defines the magnitude of the penalty, with greater values indicating a higher penalty. β_i's are assumed independent and can be interpreted from a Bayesian perspective as the posterior mode of each exposure's estimated causal effect. Park & Casella (2008) introduced a fully Bayesian approach using prior distributions, the estimate is generated from the mean of a distribution conditioned on the penalty parameter

$$\begin{aligned} \beta_i &\sim N\left(0, \sigma^2 \lambda_i\right) \\ \lambda_i &\sim \text{Exp}\left(\frac{\tau^2}{2}\right). \end{aligned} \tag{4.3}$$

4.2.2 Horseshoe prior distribution

The horseshoe prior is in the family of multivariate scale mixtures of normal distributions. It has similar features to Bayesian model averaging methods for handling sparsity and large outlying observations. From (4.4), τ represents the global hyperparameter that follow a half-Cauchy distribution that penalises the vector β towards zero and λ_i represents the local hyperparameter that reduces the effect of the shrinkage on each phenotype's causal effect β_i

$$\begin{aligned} \beta_i &\sim N(0, \tau^2\lambda_i^2) \\ \lambda_i &\sim C^+(0, \tau) \\ \tau &\sim C^+(0, \sigma). \end{aligned} \tag{4.4}$$

From (4.4) the density of the local shrinkage parameter conditional to τ is denoted as

$$p(\lambda_i/\tau) = \frac{2}{\pi\tau\Big(1 + (\frac{\lambda_i}{\tau}^2\Big))}$$

which leads to the density of ϕ_i conditional on τ in (4.5)

$$p(\phi_i/\tau) \propto \frac{\tau}{\sqrt{\phi_i(1-\phi_i)}} \frac{1}{(1+\phi_i(\tau^2-1))}. \tag{4.5}$$

The shrinkage profile is plotted in figure 4.1 to indicate the Jacobian terms of the horseshoe prior. The Jacobian indicates the behaviour of the profile to seperate signals ($\phi_i = 0$) from noise ($\phi = 1$). Carvalho et al. (2010) investigated the behaviour of the shrinkage profile (ϕ_i) for the horseshoe prior, which is derived in appendix B.1.

4.2.3 Horseshoe+ prior distribution

The horseshoe+ prior distribution is an extension of the horseshoe prior where the local shrinkage parameters (λ_i)'s are assumed conditionally independent with an extra local shrinkage parameter η_i (Bhadra et al., 2017)

$$\begin{aligned} \beta_i &\sim N(0, \tau^2\lambda_i^2) \\ \lambda_i &\sim C^+(0, \eta_i\tau) \\ \tau &\sim C^+(0, \sigma) \\ \eta_i &\sim C^+(0, 1). \end{aligned} \tag{4.6}$$

The density of the local shrinkage parameter from (4.6) integrated over the extra parameter is

$$p(\lambda_i/\tau) = \frac{4}{\pi^2\tau}\frac{\log(\lambda_i/\tau)}{(\lambda_i/\tau)^2 - 1}$$

that leads to the density of ϕ_i conditional on τ,

$$p(\phi_i/\tau) \propto \frac{\tau}{\sqrt{\phi_i(1-\phi_i)}}\frac{\log\left(1-\phi_i\right)/\phi_i\tau^2}{(1+\phi_i(\tau^2-1))}. \tag{4.7}$$

The horseshoe+ prior offers another horseshoe U-shaped Jacobian component that drives posterior mass to the locations of interest. This provides an extra level of efficiency in the case of several sparse signals.

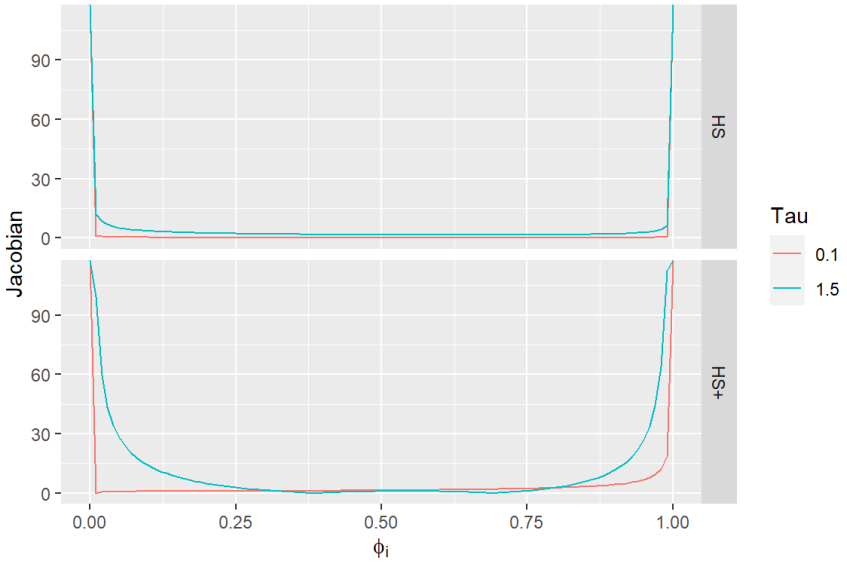

Figure 4.1: Comparing marginal prior densities of ϕ. The top panel shows the density of the horseshoe (HS) prior, and the bottom panel shows the density of the horseshoe+ (HS+) prior.

4.2.4 Estimation including horseshoe prior distributions

In this research the credible intervals of the marginal distribution are used for (exposure) variable selection. For this analysis, the Bayesian models were estimated using the Stan software within R using the RStan package (Stan Development Team, 2018). The sampling technique is based on Hamiltonian Monte Carlo, which uses Hamiltonian dynamics on the derivatives from the density function to produce efficient samples from the posterior distribution (Betancourt & Girolami, 2015; Neal et al., 2012). The posterior distribution is computed from 10 000 iterations, a burn-in of 1 000 samples with 4 chains the threshold for the credible interval is set at 95%.

The formulation for horseshoe and horseshoe+ priors in (4.4) and (4.6) encounters sampling issues when fitting models using Hamiltonian Monte Carlo. This is due to the posterior distribution having an extreme funnel shape which is related to the

thick Cauchy tails of the prior (Betancourt, 2017). To mitigate divergence issues, the half-Cauchy priors on the global-local shrinkage parameters can be written as a mixture of gamma densities

$$\begin{aligned}
\beta_i &\sim N(0, \tau^2\lambda_i^2) \\
\frac{1}{\lambda_i} &\sim G(0.5, \omega) \\
\omega &\sim G(0.5, \tau) \\
\frac{1}{\tau} &\sim G(0.5, \Psi) \\
\Psi &\sim G(0.5, \sigma^2).
\end{aligned}$$

This is similar to the computation of the horseshoe prior in MR by Berzuini et al. (2020). This approach is used for horseshoe and horseshoe+ prior distributions.

4.3 Simulations

The Bayesian models presented above will be investigated in three simulation scenarios and compared with multivariate MR-IVW and MR-Egger models introduced in section 3.7. The simulation will assess the bias, standard deviation of the estimates and coverage of the models. Data generation for the simulation study is based on the variables denoted in (3.1). (4.8) shows the data generating process for the variables of the risk factors and disease outcome. The simulation study indexed instruments as j and individuals indexed as i. The simulation scenarios will be outlined in the following subsections, along with the simulation results. The variable X_{ip} in (4.8) represents the exposure and Y_i is the variable for outcome. The coefficient on the direct effect of the genotypes, i.e., the pleiotropic effect of the genotypes is set to 0.2,

$$
\begin{aligned}
X_{ip} &= \sum \phi_j G_{ij} + U + \epsilon_{x_{ip}} \\
Y_i &= 0.2G_{ij} + \beta_{ip} X_{ip} + U + \epsilon_{y_i} \\
G_{ij} &\sim \text{Binom}(2, 0.3) \\
\phi_j &\sim U(0.05, 0.15) \\
\epsilon_{x_{ip}}, \epsilon_y &\sim N(0, 1).
\end{aligned} \tag{4.8}
$$

4.3.1 Collider scenario

A collider variable has been described in section 1.1.1, the scenario represents a multivariable study design with 4 exposures ($p = 4$) with exposure variables X_2 and X_3 having a null effect on the disease outcome with X_4 colliding with Y as shown in figure 4.2. The simulation scenario uses 10 instruments, 1000 individuals and 5000 iterations.

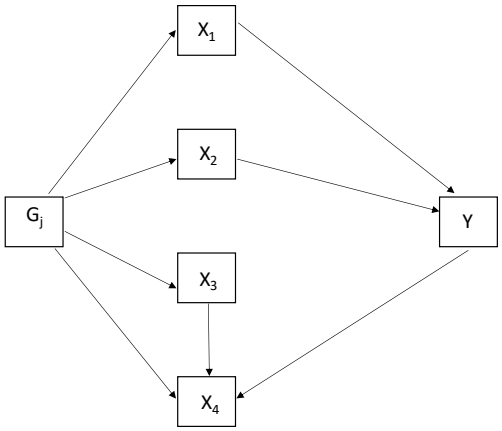

Figure 4.2: DAG representing simulation scenario 1, where X_4 is collider.

From table 4.1, the value of the bias for each model is approximately null. Simulations

from the exposure with the true effect ($\beta_1 = 1$) show larger standard deviations from MVMR-horseshoe and MVMR-horseshoe+ models while MVIVW and MVMR-Egger models have the largest coverage. For the exposure with null effect (β_2), MVIVW and MVMR-Egger models have the largest coverage and largest standard deviations, for β_3 the MVMR-BayesLasso model has the largest standard deviation while MVIVW and MVMR-Egger models have the largest coverage. The simulation results from the collider variable β_4 show the MVMR-Egger model having the largest standard deviations and largest coverage.

Table 4.1: Estimates from the collider scenario simualtions. SD is the standard deviation of the estimates and cov indicates the coverage.

Models	$\beta_1 = 1$			$\beta_2 = 0$			$\beta_3 = 0$			$\beta_4 = 1$		
	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov
MVIVW	-0.0001	0.163	0.99	-0.0001	0.1	0.95	0.0002	0.63	0.95	-0.0001	0.09	0.98
MVMR-Egger	-0.0001	0.136	0.99	0.0001	0.21	0.95	0.0002	0.861	0.95	-0.0001	0.215	0.98
MVMR-BayesLasso	-0.0001	0.1	0.91	-0.00003	0.022	0.76	0.0002	1.06	0.76	-0.00004	0.05	0.82
MVMR-Horseshoe	-0.0001	0.302	0.83	-0.00001	0.005	0.72	0.0002	0.553	0.71	-0.0001	0.128	0.83
MVMR-Horseshoe+	-0.001	0.303	0.82	-0.00002	0.006	0.71	0.0002	0.642	0.71	-0.0001	0.1	0.82

The estimates from the simulation in figure 4.3 shows the MVMR-Egger model has larger variability compared to the other models and the MVMR-horseshoe and MVMR-horseshoe+ having the lowest variability. For $\beta_1 = 1$ more than approximately 75% of the estimates fall below the true value and half of the estimates fall below the null for β_2 and β_3. For the collider variable most of the estimates ($> 95\%$) fall below the true value.

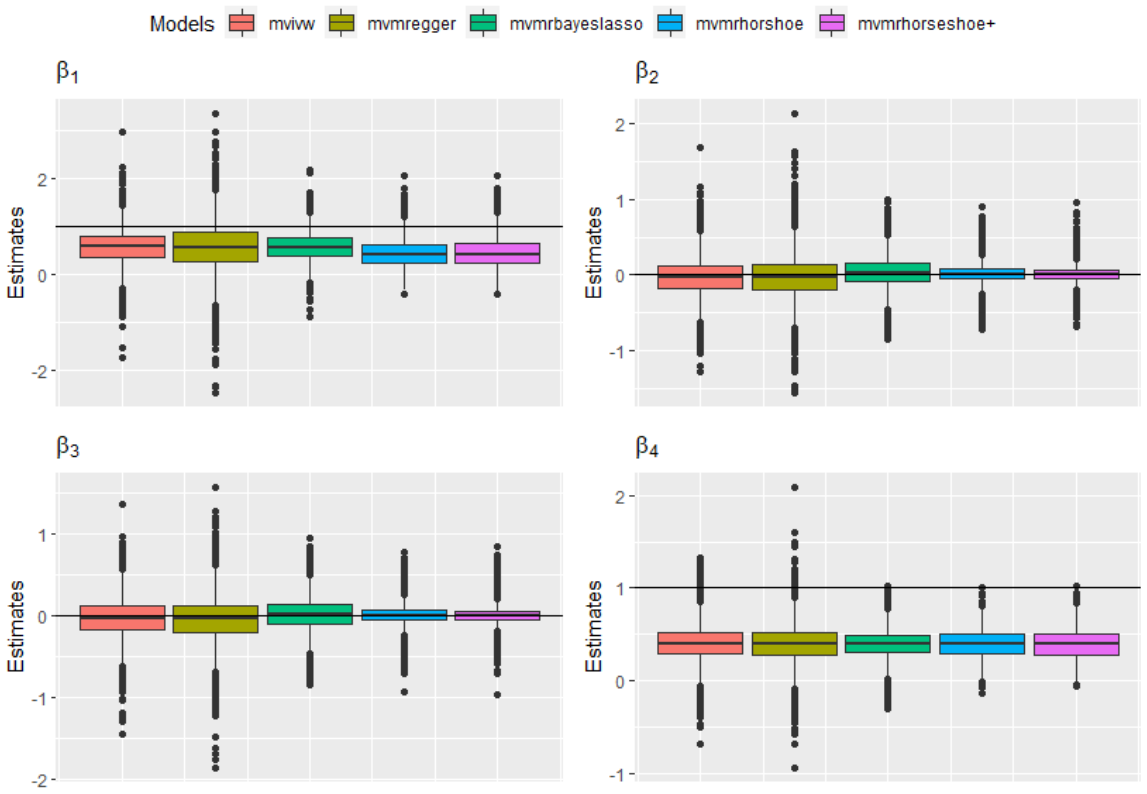

Figure 4.3: Boxplots of the estimates from the collider simulation scenario. Each plot summarises the estimates for an exposure.

4.3.2 Mediation Scenario

Section 1.1.1 describes a mediation scenario. This simulation scenario is similar to the collider scenario with the same number of exposures ($p = 4$) but with a different relationship between the exposure variables. Variables X_2 and X_3 have a null effect on Y while X_4 mediates X_3 and Y depicted in figure 4.4. The simulation scenario uses 10 instruments, 1000 individuals, and 5000 iterations.

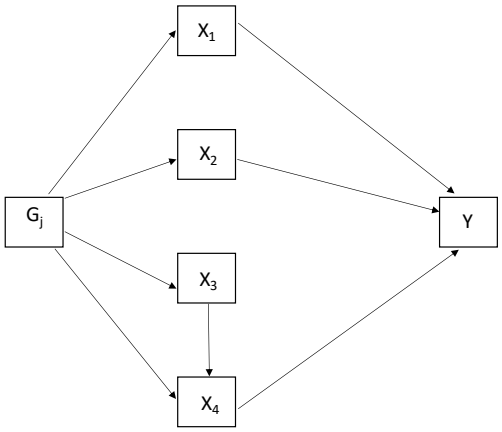

Figure 4.4: DAG representing simulation scenario 1

Table 4.2 shows simulation results for the mediation scenario, the bias for each value of the estimate is approximately null. For the true exposure X_1 the MVMR-Egger model has the largest standard deviation and all the models have the same value of coverage. The exposure variable with null effect (X_2) indicates that the MVMR-horseshoe+ model has the largest standard deviation and that the MVIVW model has the largest coverage. For the exposure variable with null effect (X_3), the MVMR-Egger model has the largest standard deviation and the MVIVW model has the largest coverage. For the X_4 exposure variable MVIVW has the largest standard deviation. Figure 4.5 shows the estimates of the models have lower variability than in the collider scenario and approximately 50% of the estimates fall below the true value for all the variables.

Table 4.2: Estimates the mediation scenario simulations. SD is the standard deviation of the estimates and cov indicates the coverage.

Models	$\beta_1 = 1$			$\beta_2 = 0$			$\beta_3 = 0$			$\beta_4 = 1$		
	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov
MVIVW	0.0001	0.197	0.99	0.0003	2.71	0.99	0.0001	0.442	0.99	0.0002	0.65	0.99
MVMR-Egger	0.0001	0.231	0.99	0.0003	2.51	0.95	0.0002	0.574	0.95	0.0001	0.144	0.99
MVMR-BayesLasso	0.0001	0.11	0.99	0.0003	2.54	0.89	0.0001	0.242	0.88	0.0001	0.348	0.99
MVMR-Horseshoe	0.0001	0.032	0.99	0.0003	2.97	0.88	0.0001	0.224	0.87	0.0002	0.551	0.99
MVMR-Horseshoe+	0.0001	0.052	0.99	0.0004	3.38	0.87	0.0001	0.25	0.86	0.0001	0.477	0.99

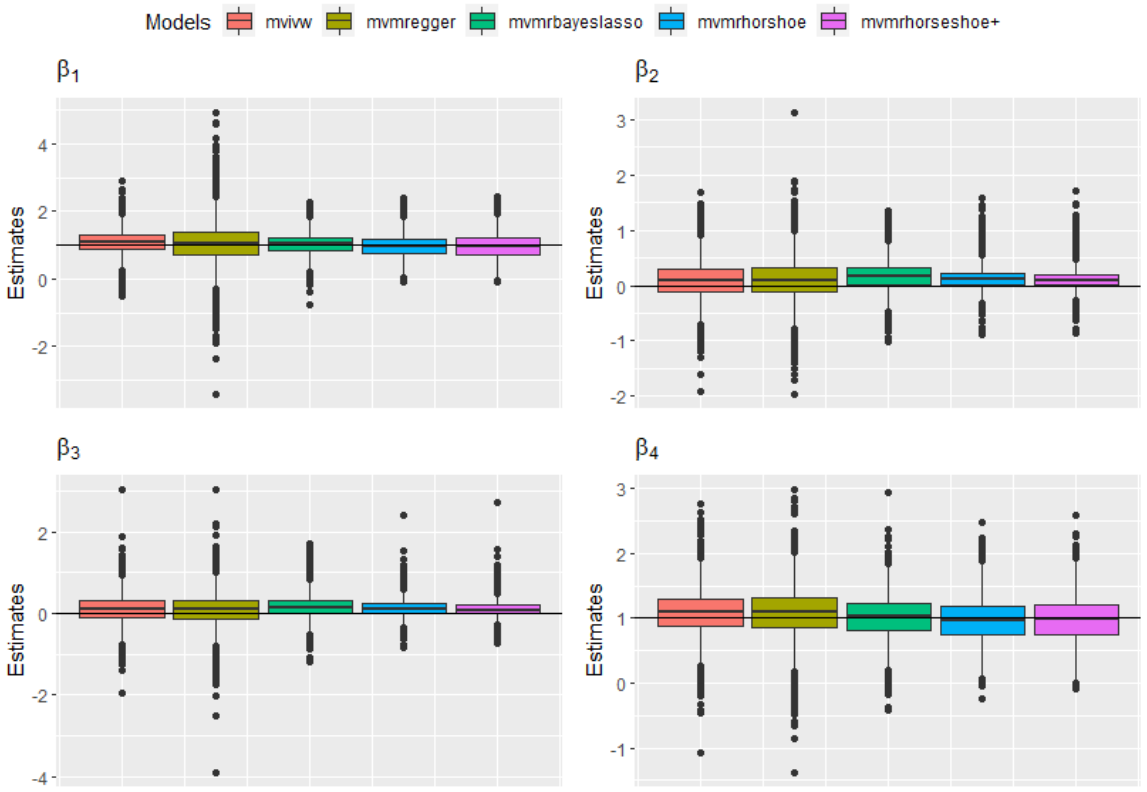

Figure 4.5: Boxplots of the estimates from the mediation simulation scenario. Each plot represents the estimated effect of an exposure

4.3.3 High throughput simulation scenario

In this simulation scenario four exposures have a direct effect on the outcome. The study also incorporates eleven exposures that have no effect on the disease outcome in order to produce a high throughput MVMR study design. For this simulation, 50 instruments were used with 1000 iterations. Figure 4.6 shows a summary of this simulation scenario. The simulation results are shown in table 4.3 and figure 4.3.3.

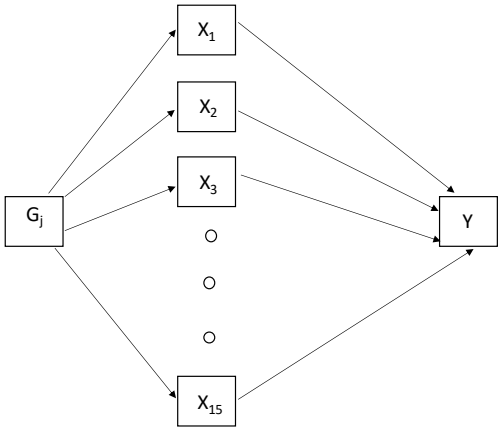

Figure 4.6: DAG representing high throughput simulation scenario

In table 4.3 the models show very low bias for all the estimates (≈ 0), and MVIVW and MVMR-Egger models have the largest coverage. For the first true exposure the MVMR-horseshoe+ model has the largest standard deviation while MVIVW and MVMR-Egger models have the largest standard deviation for the second true exposure. The MVIVW model also has the largest standard deviation for the third and final true exposures. For the exposures with null values, the highest standard deviations vary between MVIVW and MVMR-Egger models except for β_9, β_{12}, β_{13} and β_{14}.

Table 4.3: Estimates from the high throughput simulation. SD is the standard deviation of the estimates and cov indicates the coverage.

Models	$\beta_1 = 1$			$\beta_2 = 1$			$\beta_3 = 1$			$\beta_4 = 1$			$\beta_5 = 0$			$\beta_6 = 0$		
	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov
MVIVW	-0.001	0.371	0.99	-0.003	0.1	0.99	-0.001	2.02	0.99	-0.001	1.5	0.99	0.001	0.59	0.99	-0.0001	0.01	0.99
MVMR-Egger	-0.001	0.363	0.99	-0.0003	0.1	0.99	-0.001	1.14	0.99	-0.001	1.43	0.99	0.001	0.52	0.99	0.0001	0.004	0.99
MVMR-BayesLasso	-0.001	0.476	0.57	-0.0003	0.06	0.58	-0.001	1.62	0.59	-0.001	1.12	0.59	0.001	0.34	0.91	0.0001	0.004	0.89
MVMR-Horseshoe	-0.001	0.588	0.55	-0.0002	0.05	0.55	-0.001	1.12	0.55	-0.001	0.95	0.54	0.001	0.23	0.97	-0.00001	0.00002	0.96
MVMR-Horseshoe+	-0.001	0.61	0.55	-0.0001	0.03	0.54	-0.001	1.06	0.54	-0.001	0.95	0.55	0.001	0.23	0.97	0.00001	0.0002	0.96

Models	$\beta_7 = 0$			$\beta_8 = 0$			$\beta_9 = 0$			$\beta_{10} = 0$			$\beta_{11} = 0$			$\beta_{12} = 0$		
	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov
MVIVW	-0.001	0.28	0.99	-0.0004	0.165	0.99	0.001	0.43	0.99	0.001	0.33	0.99	0.001	0.62	0.99	0.00004	0.002	0.99
MVMR-Egger	-0.001	0.316	0.99	0.0004	0.164	0.99	0.001	0.43	0.99	0.001	0.31	0.99	0.001	0.62	0.99	0.00001	0.0001	0.99
MVMR-BayesLasso	-0.0002	0.043	0.91	0.0003	0.1	0.91	0.001	0.41	0.91	0.001	0.28	0.91	0.001	0.53	0.90	0.0001	0.004	0.90
MVMR-Horseshoe	-0.00003	0.0009	0.97	0.0002	0.051	0.96	0.001	0.48	0.96	0.0004	0.144	0.96	0.001	0.55	0.97	0.0001	0.004	0.95
MVMR-Horseshoe+	-0.0000005	-0.0000002	0.97	0.0002	0.045	0.97	0.001	0.46	0.97	0.0003	0.146	0.96	0.001	0.56	0.96	0.0001	0.005	0.96

Models	$\beta_{13} = 0$			$\beta_{14} = 0$			$\beta_{15} = 0$		
	Bias	SD	Cov	Bias	SD	Cov	Bias	SD	Cov
MVIVW	0.0003	0.06	0.99	0.001	0.478	0.99	0.001	0.667	0.99
MVMR-Egger	0.0002	0.03	0.99	0.001	0.48	0.99	0.001	0.648	0.99
MVMR-BayesLasso	0.0003	0.08	0.92	0.001	0.34	0.92	0.001	0.465	0.91
MVMR-Horseshoe	0.0003	0.07	0.99	0.001	0.64	0.97	0.001	0.33	0.97
MVMR-Horseshoe+	0.0003	0.08	0.98	0.001	0.667	0.97	0.001	0.372	0.97

For the true exposure variables the simulation estimates in figure 4.3.3 show approximately 95% fall below the true value and the Bayesian models have lower variability compared to the MVIVW and MVMR-Egger models. For the null exposure variables, approximately 75% of the estimates are above the true value with similar variability as the true exposure variables.

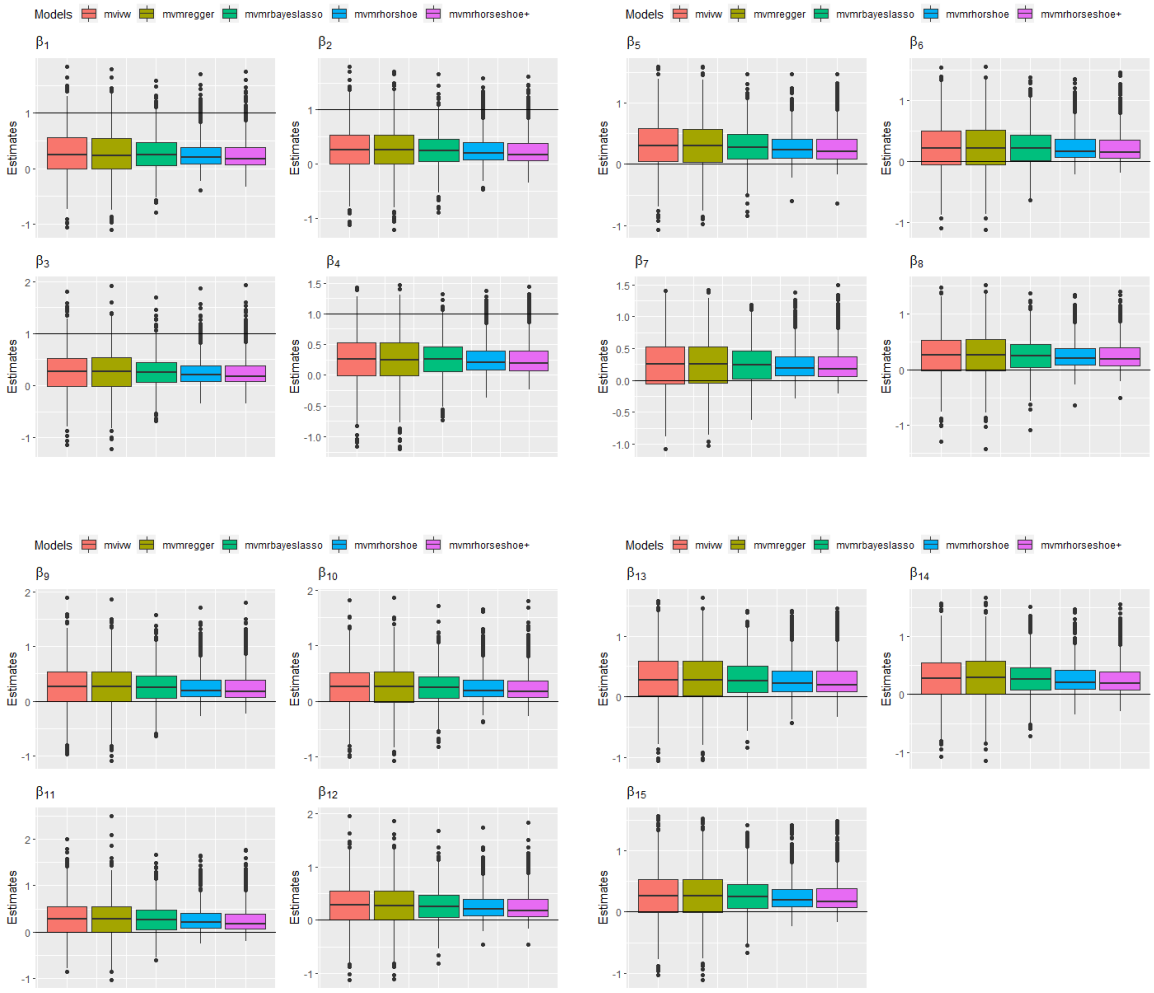

Figure 4.7: Boxplots from estimates from the high throughput simulation scenario

4.4 Data application

For this section, the proposed methods are applied in two two-sample summary level datasets which include; i) a small number of exposures and ii) a large number of exposures. The details of the study designs are described in further sections.

4.4.1 Data description

The models are applied to the summary-level dataset previously published by Kettunen et al. (2016), which was made available in the MRChallenge GitHub repository (https://github.com/WSpiller/MRChallenge2019). The dataset consists of 148 instruments with 118 exposures and 7 disease outcomes. The exposures are grouped into metabolite risk factors. The outcome consists of Ischemic stroke, type-2 diabetes, small vessel stroke, cardioembolic stroke, age-related macular degeneration, Alzheimer's disease, and large artery disease.

4.4.2 Data application 1: Investigating the causal effect of low-density lipoprotein particle sizes on cardioembolic stroke

The multivariate models are applied to the summary-level data described in 4.4.1 to investigate the causal effect of LDL particle sizes on cardioembolic stroke. The particle sizes selected are small very-low density lipoprotein (S.VLDL.P), small low density lipoprotein (S.LDL.P), medium very-low density lipoprotein (M.LDL.P), large very-low density lipoprotein (L.VLDL.P), and large low density lipoprotein (L.LDL.P). Figure 4.8 shows the results of the analysis. The MVIVW and MVMR-Egger models provide evidence of a causal effect of M.LDL.P on cardioembolic stroke, whereas, the Bayesian models show no evidence of a causal effect.

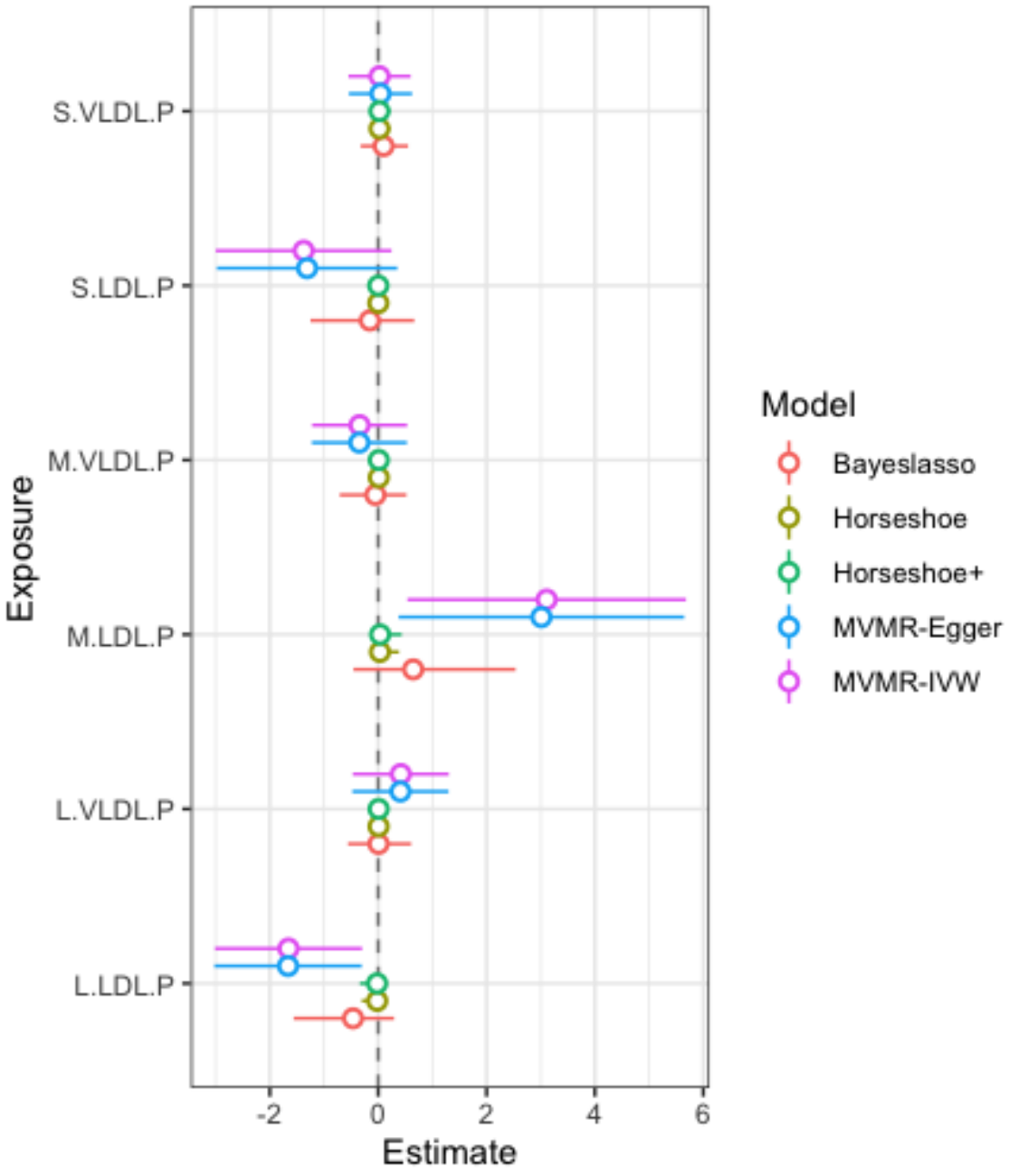

Figure 4.8: Results from multivariate models using cardioembolic stroke as the outcome.

Further assessment of the estimates in table 4.4 indicates that M.LDL.P has a very large estimate and the confidence interval does not include the null using the MVIVW and MVMR-Egger models. However, the MVMR-BayesLasso model has a lower estimate, with a credible interval that does include the null, whilst the MVMR-Horseshoe and

MVMR-Horseshoe+ models show estimates at the null, again with credible intervals including the null. This demonstrates a significant difference in the models, which will be discussed more in the following section.

Table 4.4: Direct effect of M.LDL.P

Models	Estimates (95% CI/CrI)
MVMR-IVW	3.11(0.54,5.68)
MVMR-Egger	3.01(0.38,5.65)
MVMR-BayesLasso	0.64(-0.46,2.53)
MVMR-Horseshoe	0.03(-0.13,0.38)
MVMR-Horseshoe+	0.04(-0.12,0.43)

4.4.3 Data application 2: Investigating the causal effect of cholesterol content, triglyceride content, and particle diameter on Ischemic stroke

The exposures are pre-selected to include only lipoprotein measurements on total cholesterol content, triglyceride content, and particle diameter in order to avoid multi-collinearity among the exposures. Furthermore, risk factors are chosen if they are strongly associated with at least one genetic variant which is included as an instrumental variable, through selecting exposures with p-values less than a certain threshold ($p< 5 \times 10^{-8}$). Figure 4.9 shows the genetic correlation between the selected exposures.

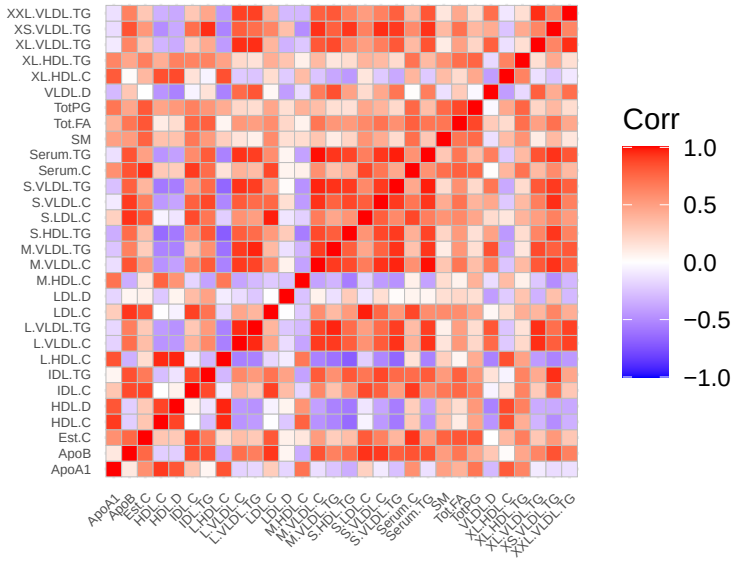

Figure 4.9: Genetic correlation between the selected exposures.

The MVMR models introduced earlier are applied on LDL-C-related lipid subfractions to examine their causal effect on Ischemic stroke. The results of multivariate analysis show the estimates and their confidence/credible intervals from the models in figure 4.10. Table 4.5 shows the estimate of the direct effect of S.LDL.P from the different models. The MVMR-IVW and MVMR-Egger models show evidence of a causal effect, whereas the Bayesian models do not provide evidence in favour of a causal effect.

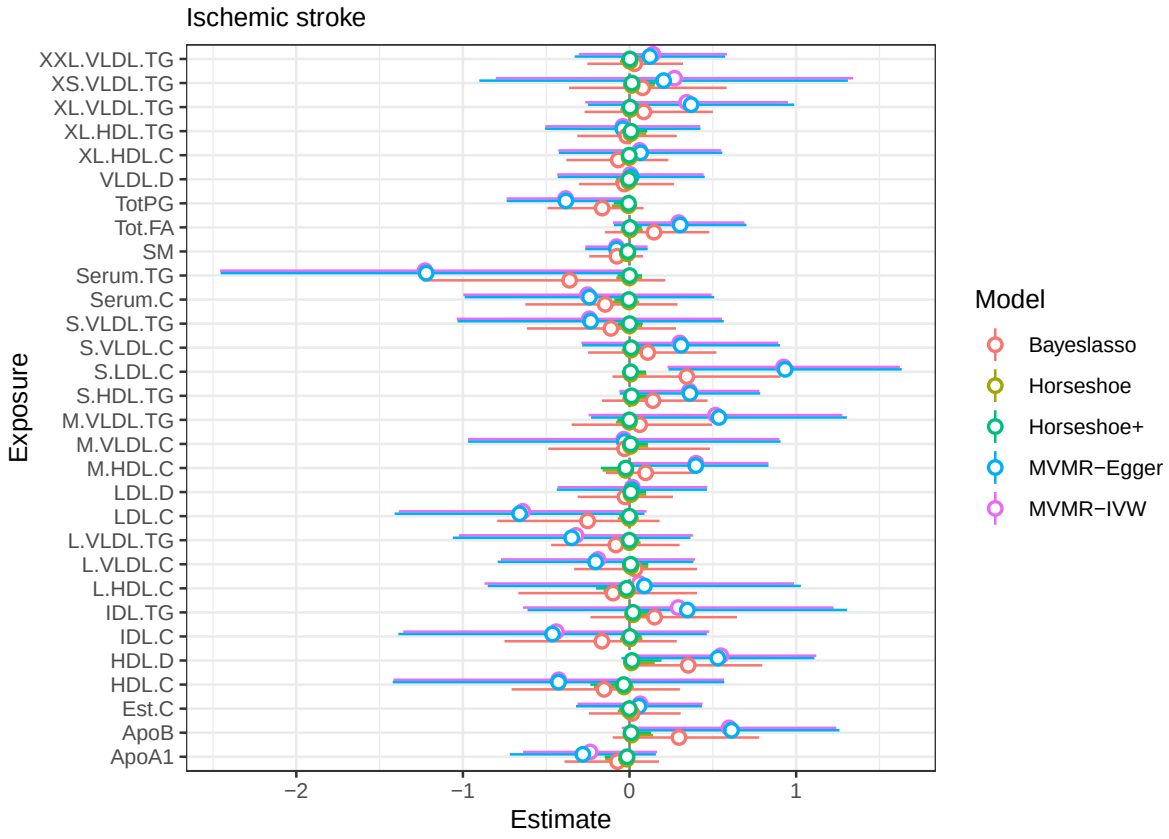

Figure 4.10: Results from multivariate models using Ischemic stroke as the outcome.

Table 4.5: Direct effect of S.LDL.C

Models	Estimates (95% CI/CrI)
MVMR-IVW	0.93 (0.23,1.62)
MVMR-Egger	0.93 (0.23,1.63)
BayesLasso	0.34 (-0.10,0.91)
Horseshoe	0.01 (-0.04,0.10)
Horseshoe+	0.0065 (-0.04,0.1)

4.5 Discussion

This section discusses the differences between the multivariate MR models, potential extensions of shrinkage priors with relation to hyperparameters, as well as alternative techniques for the selection of exposures in the MVMR study designs. Finally, the possibility of big data in MR is discussed.

4.5.1 Penalised versus non-penalised models

The simulation results show MVIVW and MVMR-Egger models having larger coverage than the Bayesian models presented in section 4.2 with similar bias values. In section 4.4, the main difference between the penalised and non-penalized models is highlighted. The non-penalised models (MVIVW and MVMR-Egger) have large and significant estimates, but the penalised models (MVMR-BayesLasso, MVMR-Horseshoe, and MVMR-Horseshoe+) have reduced and non-significant estimates. This raises concerns about the overshrinkage of estimates derived from penalised models. In a later subsection, we will discuss on choice of hyper parameters and variable selection.

4.5.2 Choice of hyperparameters

For the Bayesian models using hierarchical prior distributions (e.g., horseshoe and horshoe+), there has been no consensus on how to derive the inference on the global parameter. An estimate of τ can be input into (4.4) which can be obtained from the maximum marginal likelihood, which has the advantage of lower computational cost. However there is a possibility of collapsing $\tau = 0$ in the presence of larger exposures due to an increased number of exposures with a null effect. Bayesian inference is preferred for the global parameter for computational efficiency and because it can account for posterior uncertainty. The proposed choice of prior follows a half-Cauchy distribution for the global parameter by Carvalho et al. and Gelman et al..

4.5.3 Variable Selection

From our study, we observed that estimates from models using the horseshoe and horseshoe+ priors were heavily penalized which affects the variable selection. Zuber et al. (2020) factored the marginal inclusion probability of each exposure to improve variable selection. We suggest the prior distribution of τ parameter is modified such that the variance is derived from the assumed non-zeros from the vector β_p similar to Piironen & Vehtari (2016). In the presence of correlated exposures the increase in the number of exposures can produce results that are not intuitive. We also suggest further research into applying the regularized horseshoe prior (Piironen et al., 2017).

4.5.4 Conclusion

The increased number of GWAS studies on metabolites and blood lipids provides the opportunity for a summary-level dataset with large number of exposures in an MR study. This work demonstrates the difference between penalised and non-penalized models in the MVMR study design and highlights that over-shrinkage can be problematic.

Chapter 5

Weighted percentile and conditional quantile estimation for summary-level Mendelian randomization analyses

5.1 Introduction

Earlier in section 2.5 we discussed that MR-IVW estimator is susceptible to outliers that may be attributed to pleiotropy. The effect of the outliers could bias the causal estimate. This motivated researchers to develop sensitivity analyzes such as the weighted median, modal estimators, and robust regression methods to mitigate the effects of outlying variants (Bowden et al., 2015; Hartwig et al., 2017; Burgess et al., 2016). Median quantile regression has semi-parametric properties by removing parametric assumptions on the error term making it robust to outliers. The model has been applied in MR settings when investigating a non-linear causal effect between the

exposure and outcome variables (Burgess et al., 2014; Staley & Burgess, 2017) and a sensitivity analysis model in a multivariate MR study design (Grant & Burgess, 2020).

This research extends the conditional and weighted median estimators to investigate the use of different quantiles as a form of further sensitivity analysis. The approach is akin to extracting estimates of different percentiles from the MR-median and weighted median estimators.

5.1.1 Quantile regression in Mendelian randomization

The motivation for looking into different quantiles will be explained in this section. Quantile regression is widely used in epidemiology, particularly in the research of growth trajectories for example Wei et al. (2019), but do they have the same meaning when it comes to a Mendelian randomization analysis? Figure 5.1 illustrates a hypothetical scenario in which a summary-level dataset is analysed using quantile regression. The instrument-exposure and instrument-outcome associations from two summary-level datasets include both valid and invalid instruments. In the left panel, the IVW estimate is biased by the invalid instruments. Whereas, the conditional median estimate using quantile regression model is close to the true value of the causal effect. The right panel shows the estimate from the 1^{st} quartile (0.25) is closer to the true value. This is a good reason to look at the other percentiles for weighted median and conditional-median, the differences of the medians will be explained in 5.2.3.

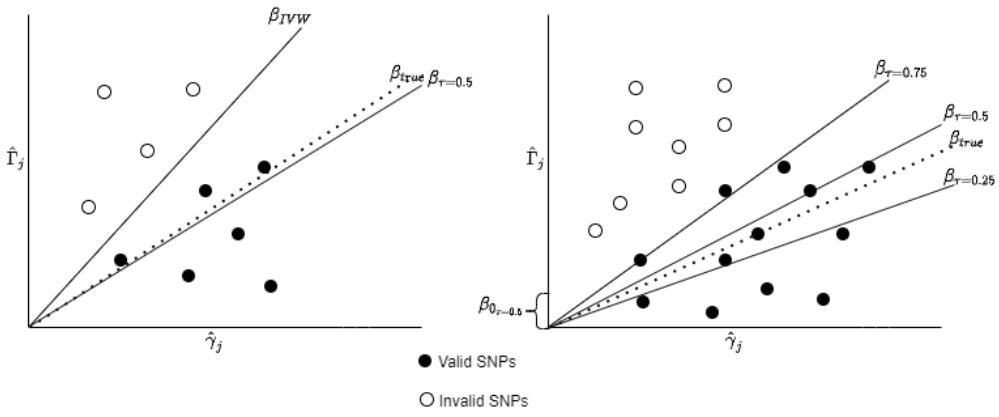

Figure 5.1: Scatter plots of instrument-exposure and instrument-outcome associations for hypothetical datasets in which the MR-median estimate is approximately unbiased (left panel) and where the MR-median estimate is biased (right panel). β_{IVW}: IVW estimate, β_{true}: True value in these hypothetical examples, $\beta_{\tau=0.5}$: MR-Median estimate, $\beta_{\tau=0.25}$: 0.25 quantile estimate

5.2 Percentile and quantile models in MR

The weighted percentile and MR-quantile models are presented in this section. This section also explains the difference between the weighted percentile and MR-quantile models, as well as the rationale behind the adjusted quartile estimator and modal quantile estimator.

5.2.1 Weighted percentile estimator

The weighted median assumes a median estimator with equal weights and it is consistent if only up to 50% of the instruments are invalid (Kang et al., 2016). The contribution of the instrument (j) and the weights to the distribution of the causal effect are proportional. The weighted percentile estimator is described below, when $\tau = 0.5$ the equation is equivalent to the weighted median estimator

$$s_j = \sum_{k=1}^{j} w_k$$
$$\theta = \frac{s_j - w_k \tau}{s_j}$$
$$\text{WP} = \hat{\beta}_k + (\hat{\beta}_{k+1} - \hat{\beta}_k)\theta.$$

The variables represent;

- k: the number of values less than the percentile
- $\hat{\beta}_k = \frac{\hat{\Gamma}_k}{\hat{\gamma}_k}$: ordered ratio estimates
- $w_k = \frac{\gamma_k}{\sigma_{y_k}}$: ordered corresponding weights
- $\tau \in (0, 1)$ represents the quantile level.

5.2.2 Quantile regression models

Quantile regression summarizes the univariate probability distribution function. It is not influenced by the tail of distributions which provides robust estimates in the presence of outliers (Koenker, 2017). The following is a brief description of the quantile distribution:

- Given a discrete random variable (X), for any $\tau \in (0, 1)$ the τth quantile of X is any value (ρ_τ) such that $\Pr(X < \rho_\tau) \leq \tau \leq \Pr(X \geq \rho_\tau)$.
- If X is a continuous random variable with cumulative distribution function then for every X the probability would be $\Pr(X < x) = \Pr(X \leq x) = F(x)$ the τth quantile is any value (ρ_τ) such that $F(\rho_\tau) = \tau$.
- Given a probability density $p(x, y)$ the quantile function for the conditional

distribution can be defined as $Q(\tau|x) = F^{-1}(\tau|X) = \inf\left[y : P(Y \leq y|x)\right]$

Koenker & Bassett Jr (1978) provide a method to estimate the conditional quantile for any $\tau \in (0, 1)$, the quantile function for Mendelian randomization

$$Q_{\Gamma}(\tau|\gamma) = \beta\gamma_j + w_j F^{-1}(\tau), \tag{5.1}$$

when $\tau = 0.5$ this is equivalent to conditional median model.

Quantile regression in MR uses a linear programming algorithm

$$\hat{\beta}(\tau) := \text{argmin} \sum_{j=1}^{J} \rho_{\tau}(\tilde{\Gamma}_j - \beta\tilde{\gamma}_j^T)$$

for estimation, the variable w_j represents the first order weights.

The variables $\tilde{\Gamma}_j$ and $\tilde{\gamma}_j$ represents the instrument associations divided by the first order weights. The loss function variable $\rho_{\tau}(.)$, is defined as

$$\rho_{\tau}(.) = \begin{cases} \tau u & u \geq 0 \\ -(1-\tau)v, & v < 0. \end{cases}$$

The quantile estimator uses an absolute loss function, which does not directly estimate a standard error, however the bootstrap method is used to estimate the standard errors, from which confidence interval limits are constructed. The standard form of the linear programme is elaborated in section C.1.

5.2.3 Ranked ratio and conditonal quantile estimates

Ranked ratio estimates are estimates derived from the weighted percentile model while quantile conditional quantile estimates are derived from the quantile regression model.

This section briefly discuss the differences between the estimators. In a research by Grant & Burgess (2020), the multivariate version of weighted median is assumed to be multivariate quantile regression model. In some cases the estimates can be closely similar, however, these approaches are different because the weighted percentile is the product of ranked ratio estimates with their respective weights, whereas the quantile regression is the conditional densities at each quantile. Also, for quantile regression, the lengths of the solution intervals for the quantiles are irregular and depend on the study design, the role of order statistics is now performed by pairs of points to define the approximate linear conditional quantile functions. In contrast to the ordinary sample quantiles which are equally spaced on the quantile interval $\tau \in [0, 1]$, with each separate order statistic occupying equal intervals of length $\frac{1}{n}$ (Koenker, 2005). This means the order of the estimates from the conditional densities would be random as the quantile level changes, table 5.1 summarises the difference between the estimators.

Table 5.1: Difference between weighted quantile and conditional quantile

Weighted quantile	Conditional quantile
Extension of weighted median for other centiles	Quantile regression for other quantiles
Ranked ratio estimates with their respective weights	Conditional densities at each quantile
Estimates are linear with the centile value	Estimates depend on study design in (5.1)

5.3 Alternative quantile estimators

This section will introduce two proposed estimators extended from conditional quantile models in (5.1) in the following subsections.

5.3.1 Adjusted quartile estimator

Given that greater than 50% of the instruments are valid, the median of the centile estimators has been proved to be consistent (Bowden et al., 2016b; Grant & Burgess, 2020). In figure 5.2, we hypothesise that, for example, the range of valid instruments lie between the 0.25 and 0.75 percentiles. Based on this intuition, we propose an adjusted centile estimator utilising the quartile estimates

$$\beta_{adj_\tau} = \frac{\beta_{0.25} + 2\beta_{0.5} + \beta_{0.75}}{4}. \tag{5.2}$$

We derive its standard error by bootstrapping methods.

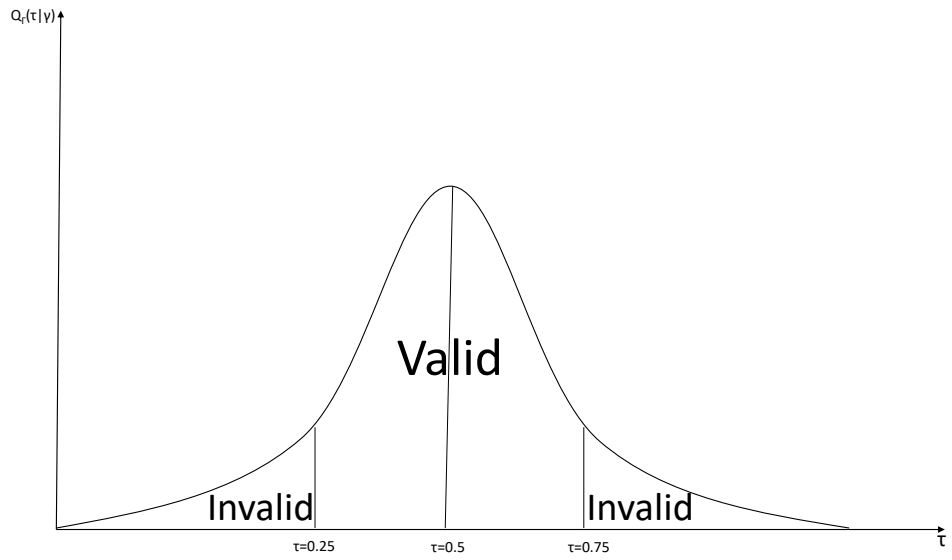

Figure 5.2: Hypothetical density plot indicating conditional density plot in summary-level Mendelian Randomization

5.3.2 Modal quantile estimator

From section 5.2.3, we discuss why the estimates from the conditional quantile are usually irregular due to the study design. The modal quantile estimator is proposed

as an alternative estimator, this involves estimating the mode of the quantile space (that is the mode of estimates from (5.1) within the quantile space). Consider the kernel density function of the quantile estimates

$$f(x) = \frac{1}{h\sqrt{2\pi}} \sum_{i=1}^{n_\tau} \exp\left[-\frac{1}{2}\left(\frac{x - \hat{\beta}_\tau}{h}\right)^2 \right] \tag{5.3}$$

where h is the smoothing bandwidth parameter that regulates bias-variance trade off. Silverman's rule of thumb is used in selecting the value h (Silverman, 2018). The mode from the kernel density of the quantile is value of x that will maximise $f(\beta_{MQE}) = \max[f(x)]$. Figure 5.3 gives a summary of the proposed quantile estimator.

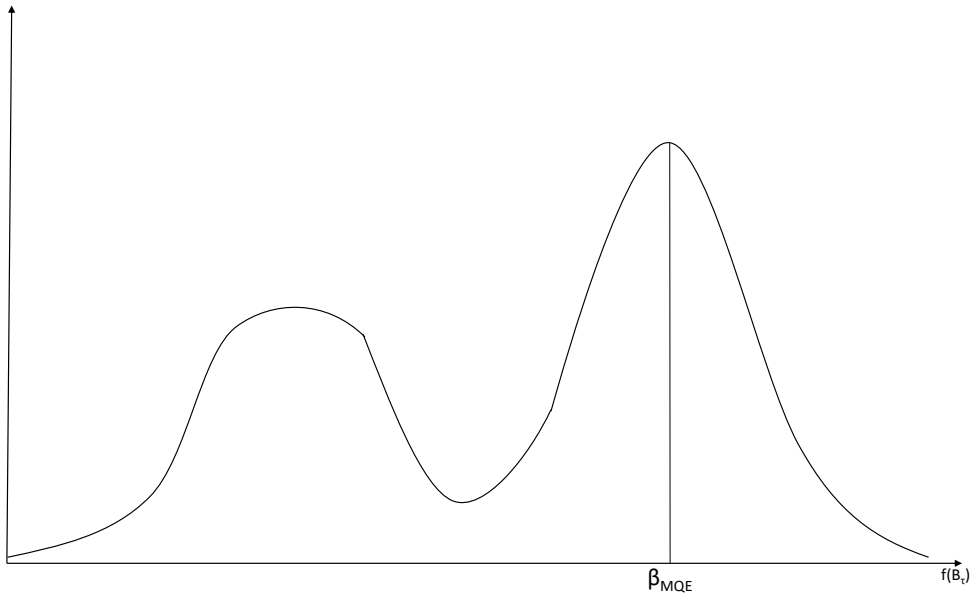

Figure 5.3: Hypothetical density plot from the distribution of quantile estimates. The x-axis denotes the distribution function of the conditional quantile estimates.

5.3.3 Instrument validity within quantiles

The research will also investigate the influence of outlying variants in the various quantiles using the generalized Cook's distance for the different quantiles (Benites et al., 2015). The method uses EM algorithm to derive the maximum likelihood of the regression quantile estimates and develop a case-deletion diagnostics analysis. This method is applied using `quokar` package in R statistical software (Wang et al., 2017).

5.4 Simulation studies

The properties of the quantile models are investigated in simulation scenarios. The setup for the simulation is in a two sample summary-level dataset generated from an individual level data with sample size of ($n = 10,000$), and large number of genetic variants $j = 200$. The simulation studies will investigate the models assuming a null effect and evidence of causal effect ($\beta = 0.2$) between the exposure and outcome variables. The variables are derived from (3.1), (3.3), and (3.4). The parameters are generated by

$$\begin{aligned} G_{ij} &\sim \text{Binomial}(2, \text{maf}) \\ \text{maf} &\sim U(0.1, 0.5) \\ \gamma_j &\sim U(0.5, 4) \\ \varepsilon^u &\sim N(0, 2) \\ \varepsilon^x,\ \varepsilon^y &\sim N(0, 1) \end{aligned} \tag{5.4}$$

where maf represents minor allele frequency.

The simulation scenarios include Balanced, Directional and Directional (InSIDE satisfied) pleiotropy. The scenarios depend on the parameters for genetic effect from

the confounder (κ_j) and pleiotropy (α_j) link to (3.4), Table 5.2 indicates the data generation model of each variable in relation to the scenario. The simulation studies would be used to assess the precision and accuracy of the estimates.

Table 5.2: The different distributions of the pleiotropy parameter in the simulation scenarios.

Scenario	Distribution of pleiotropy parameter
Balanced Pleiotropy	$\alpha_j \sim U(-0.2, 0.2)$
Directional Pleiotropy(InSIDE)	$\alpha_j \sim U(0, 0.2)$
Directional Pleiotropy	$\alpha_j \sim U(0, 0.2), \kappa_j \sim U(0, 0.1)$

5.4.1 Simulations assuming null effect

This section discusses the simulation results assuming a null effect. Table 5.3 shows the mean estimates and mean standard error from the simulation scenarios. Figure 5.4 denotes the boxplot of the estimates generated from the simulation scenarios to indicate the accuracy of the estimates. The estimators are accurate within the balanced pleiotropy scenario although the precision reduces with the increase of invalid instruments. From the directional pleiotropy scenario, the adjusted quartile estimator produced more accurate estimates, however, as the proportion of invalid instruments increases most of the estimates generated fall above the null value making them less accurate than the balanced pleiotropy scenario. Following the InSIDE assumption within the directional pleiotropy scenario, the generated estimates are more accurate than within the directional pleiotropy but less accurate as the proportion of invalid instruments increases. For the directional pleiotropy scenarios, the boxplots show the adjusted quartile estimator produce more accurate values within those scenarios.

Table 5.3: Mean estimates and standard errors within simulation scenarios assuming null causal effect

Proportion of invalid instruments	MR-median	Weighted Median	Adjusted quartile	MQE
Balanced				
0.1	0(0.002)	0(0.0022)	0(0.0013)	0(0.0022)
0.3	0(0.0024)	0(0.0024)	0(0.0018)	0(0.0029)
0.5	0(0.0031)	0(0.0026)	0(0.0026)	0(0.0042)
Directional				
0.1	0.0017(0.002)	0.0016(0.0022)	0.0021(0.0014)	0.0002(0.0023)
0.3	0.0072(0.0032)	0.0066(0.0026)	0.0108(0.0029)	0.0017(0.0031)
0.5	0.0195(0.0066)	0.0177(0.0031)	0.0235(0.0039)	0.0086(0.0051)
Directional (InSIDE)				
0.1	0.0015(0.002)	0.0014(0.0022)	0.0018(0.0014)	0.0004(0.0022)
0.3	0.0056(0.0026)	0.005(0.0024)	0.0073(0.0021)	0.0017(0.0028)
0.5	0.0124(0.0041)	0.011(0.0027)	0.0153(0.0027)	0.0057(0.0037)

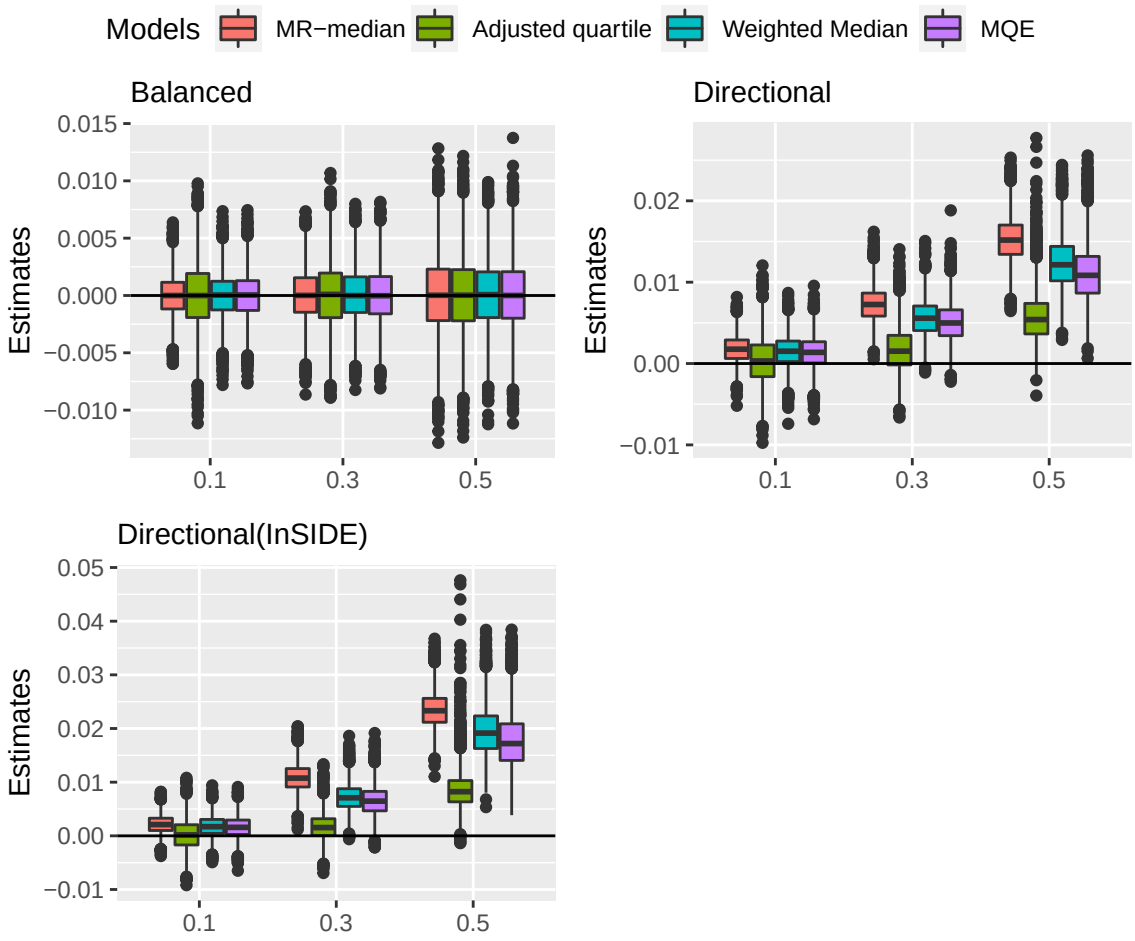

Figure 5.4: Boxplot of the estimates within the simulation scenarios. MR-median represents the quantile regression model in MR.

The mean absolute error (MAE) is measured from the simulation scenarios denoted in figure 5.5. For the pleiotropic scenarios, the modal quantile estimator(MQE) has the largest value of MAE for all proportion of invalid instruments. At 10% of invalid instruments the adjusted quartile estimator has the lowest value, within 30% of invalid instruments the adjusted quartile and MR-median estimators have the lowest MAE and at 50% of invalid instruments the MR-median and weighted median have the lowest MAE.

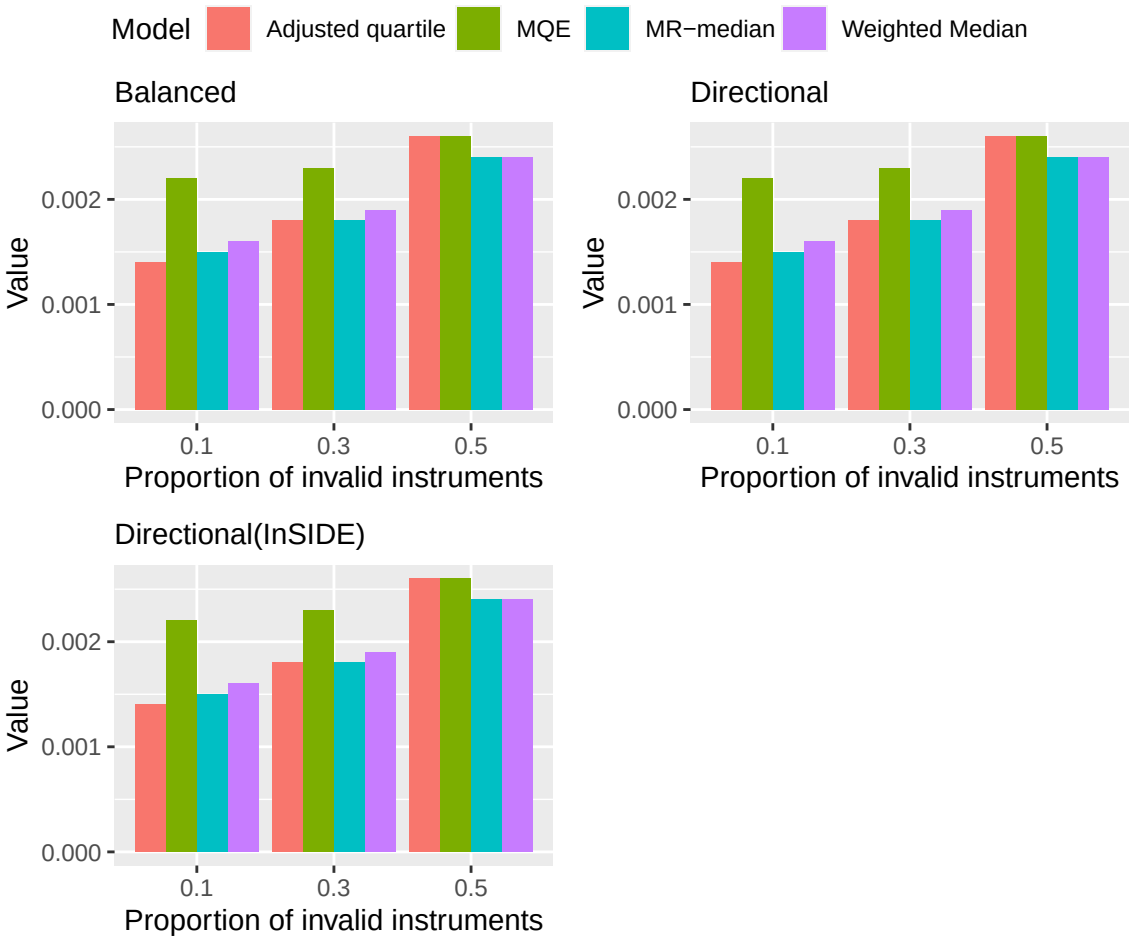

Figure 5.5: Barplot of the mean absolute error of the estimates within the simulation scenarios

Simulations also investigated the behaviour of the weighted percentile and MR-quantile estimators when the quantile (τ) lies between (0.1, 0.9). The results from the directional pleiotropy scenario (InSIDE) denoted in figure 5.6 were compared using simulations within different quantiles. Figure 5.6 shows that as the proportion of invalid instruments increases, the estimates from lower percentiles ($\tau \leq 0.4$) fall below the true value.

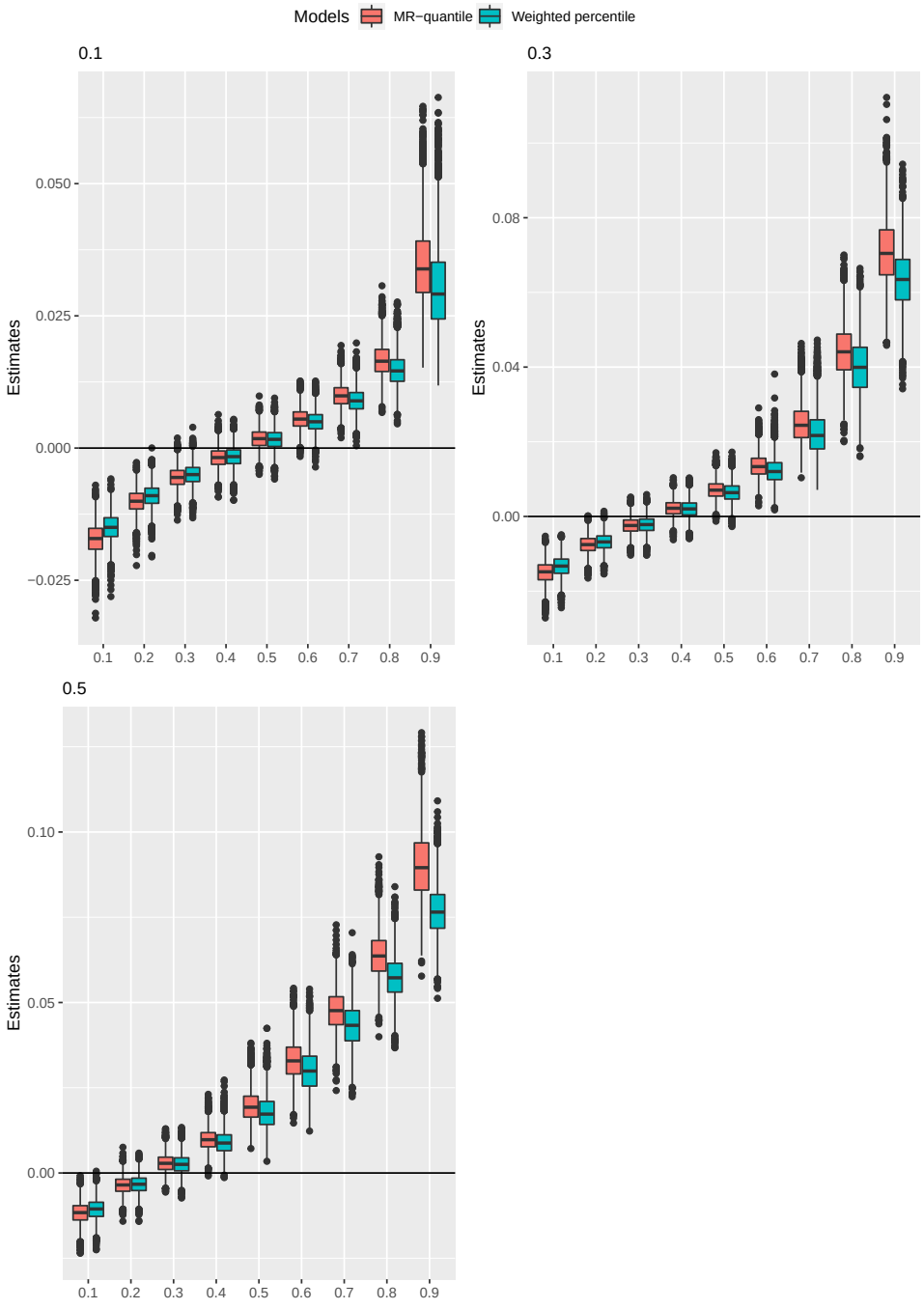

Figure 5.6: Boxplot of the estimates within different quantiles in a directional pleiotropy scenario. The title of each plot indicates the proportion of invalid instruments.

5.4.2 Simulations assuming a causal effect

Table 5.4 shows the mean estimates and mean standard error generated within the simulation scenarios. For the pleiotropic scenarios, the mean estimates from the adjusted quartile estimator are closer to the true value followed by the MR-median model, the weighted median model and the modal quantile estimator. In the directional pleiotropic scenarios, the mean estimates increases with the proportion of invalid instruments.

Figure 5.7 denotes the boxplot of estimates generated within the simulation scenario. Generally, the MR-median model produces more accurate estimates than other models and the estimators performed better within the directional (InSIDE) pleiotropy scenario more surprisingly with increased invalid instruments. Figure 5.7 also indicate that within a balanced pleiotropic scenario, the proportion of invalid instruments have little or no effect on the estimates as they remain the same within the proportion of invalid instruments. The mean absolute error (MAE) in figure 5.8 show the same features within all simulation scenarios, with the modal quantile and weighted median estimator having the higher mean absolute error and the adjusted quartile having the least MAE.

Table 5.4: Mean estimates and standard errors within simulation scenarios assuming a causal effect

Proportion of invalid instruments	MR-median	Weighted Median	Adjusted quartile	MQE
Balanced				
0.1	0.165(0.0147)	0.1484(0.0069)	0.1684(0.0098)	0.145(0.0146)
0.3	0.1648(0.0149)	0.1483(0.0069)	0.1684(0.0099)	0.1449(0.0148)
0.5	0.1645(0.0151)	0.1481(0.0069)	0.1684(0.01)	0.1444(0.0151)
Directional				
0.1	0.1702(0.0148)	0.1534(0.0069)	0.174(0.0099)	0.15(0.0147)
0.3	0.1813(0.0153)	0.1634(0.0072)	0.1857(0.0103)	0.1599(0.0154)
0.5	0.1921(0.0155)	0.1737(0.0074)	0.1972(0.0105)	0.1702(0.0158)
Directional (InSIDE)				
0.1	0.1684(0.0148)	0.1516(0.0069)	0.172(0.0098)	0.1482(0.0146)
0.3	0.1751(0.0151)	0.1577(0.0071)	0.1792(0.01)	0.1542(0.015)
0.5	0.1826(0.0153)	0.1643(0.0072)	0.1873(0.0102)	0.1613(0.0152)

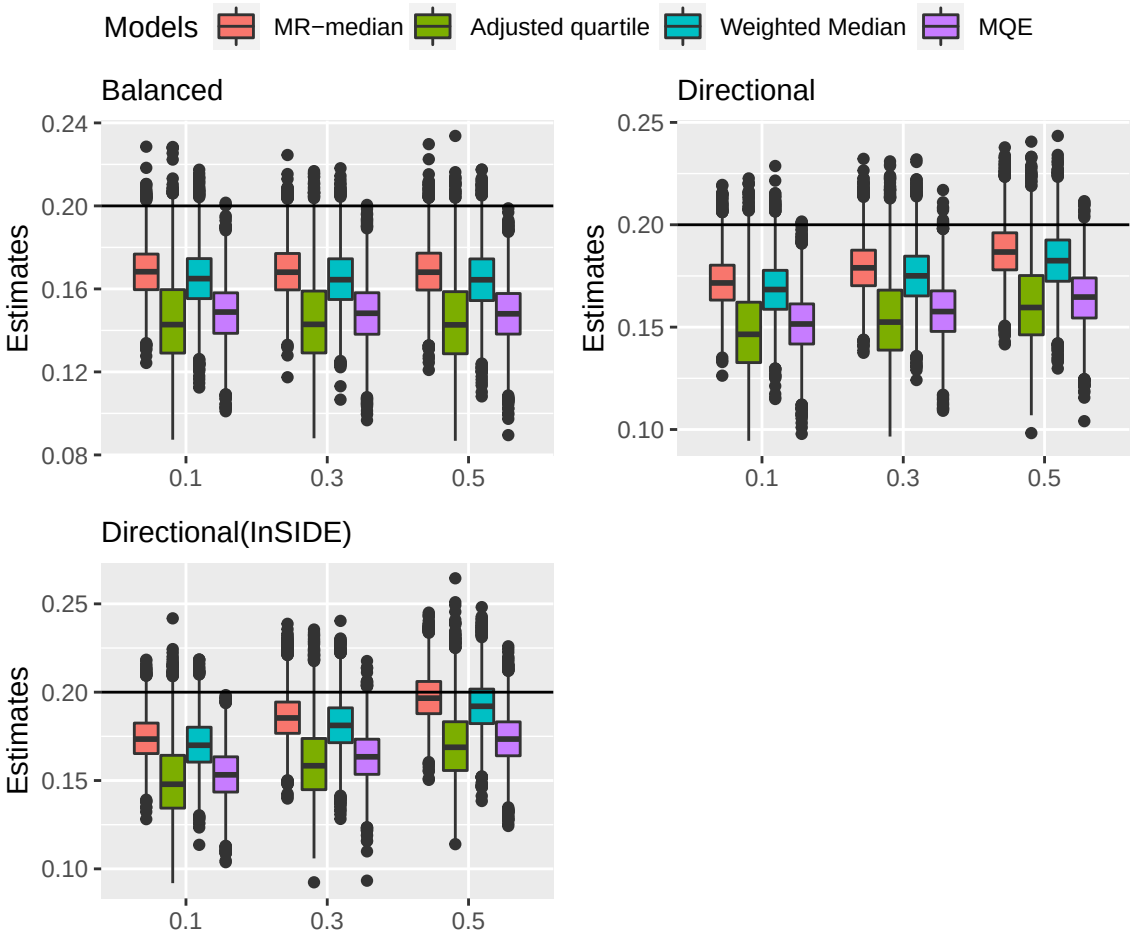

Figure 5.7: Boxplot of the estimates within the simulation scenarios. MR-median represents the quantile regression model in MR.

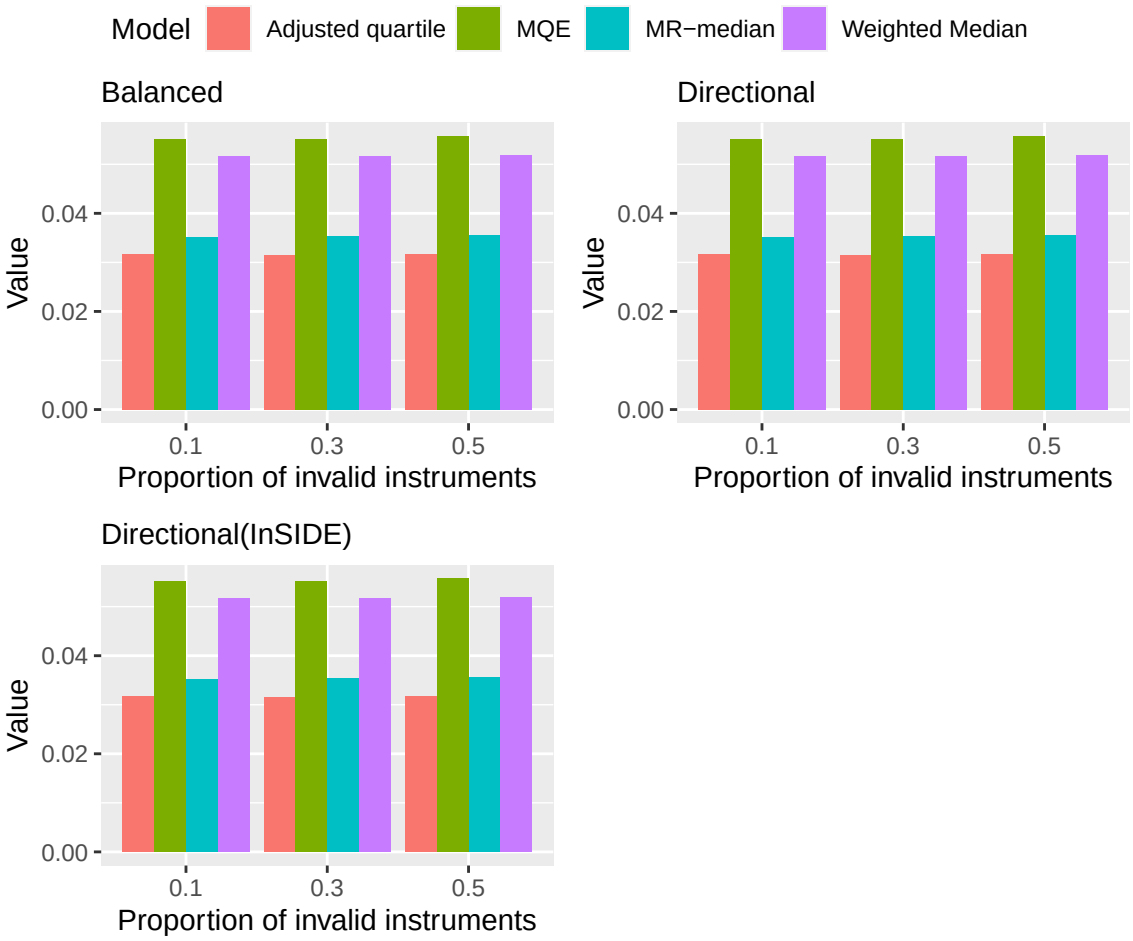

Figure 5.8: Barplot of the mean absolute error of the estimates within the simulation scenarios

The results from the directional pleiotropy scenario (InSIDE) denoted in figure 5.9 were compared using simulations within different quantiles. From figure 5.9, in the presence of varying invalid instruments, estimates from the higher quantile/percentile have a better precision towards the true value, especially within the 0.6 and 0.7 quantile.

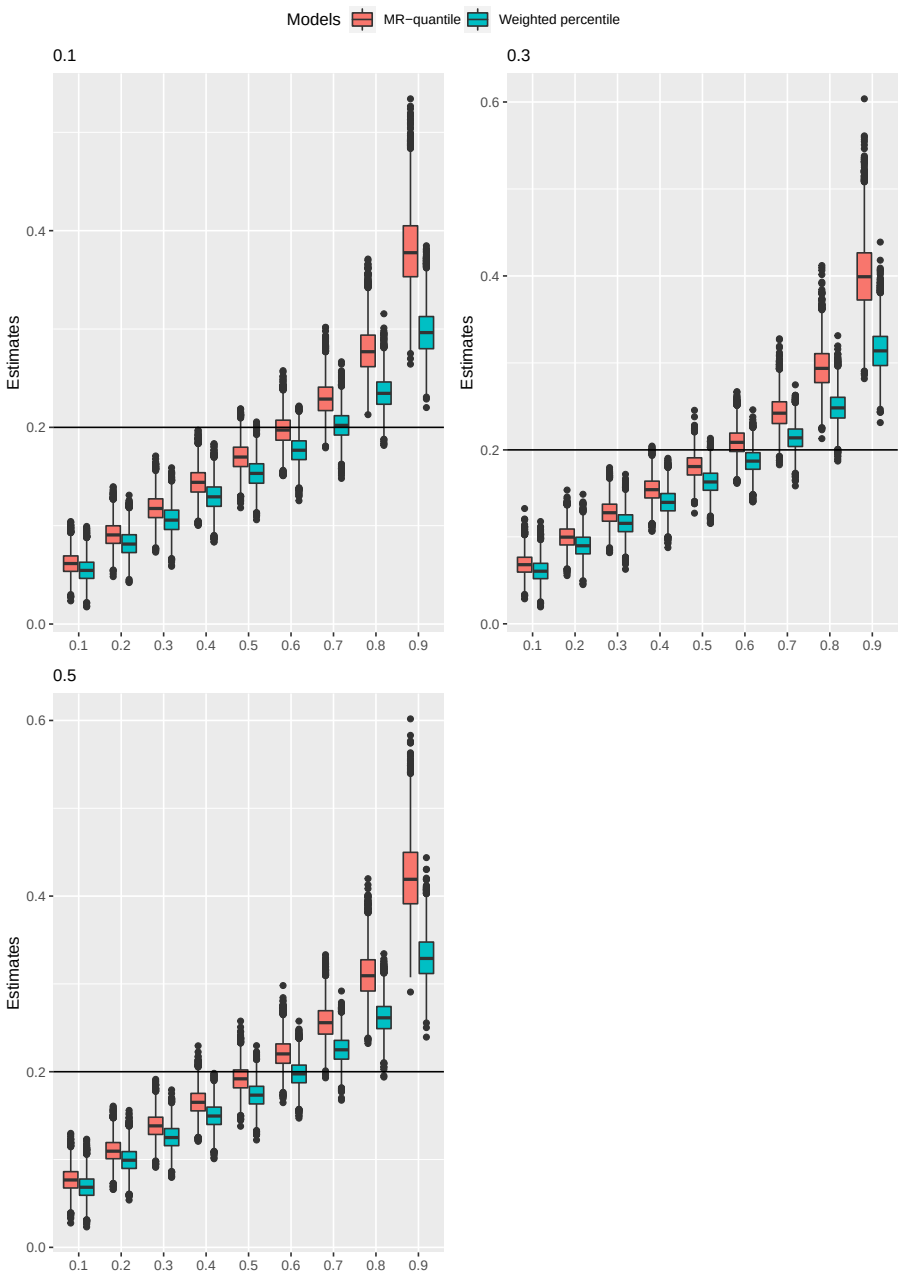

Figure 5.9: Boxplot of the estimates within different quantiles in a directional pleiotropy scenario. The title of each plot indicates the proportion of invalid instruments.

5.5 Data application

The quantile models are used to estimate the causal effect of lipid levels on coronary heart diseases. Summary-level estimates and standard errors for the data were generated with genetic variants for cholesterol levels (Waterworth et al., 2010). The summarized data includes instrument-exposure associations of LDL-C and HDL-C; the instrument-outcome association of coronary heart disease was obtained from `MendelianRandomization` package (Yavorska & Burgess, 2017). The analysis is implemented into two separate categories based on instrument selection, the categories are instruments within p-value threshold (p-value $\leq 5 \times 10^{-8}$) for each instrument-exposure associations and using all the instruments for the same analysis.

5.5.1 Analysis using instruments from p-value threshold

The estimates from the quantile models are summarised in table 5.5. Results from investigating the causal effect of LDL-C of coronary heart disease show the evidence of causal effect which is similar to previous research works. Within LDL-C, when using all the instruments also show evidence of causal effect. There is an inflation of the estimates in the presence of more invalid instruments which is similar in the simulation scenario. For the HDL-C exposure there are negative causal estimates which are significant except for the weighted median. The result from the analysis is denoted in a scatter plot for selected instruments in figure 5.10 and all instruments in figure 5.14.

Table 5.5: Estimates and confidence interval of the quantile models

Models	P-value threshold	All instruments
LDL-C		
Weighted Median	0.4588(0.3329,0.5846)	0.4606(0.3377,0.5834)
MR-median	0.4565(0.3368,0.5763)	0.4571(0.3393,0.575)
Adjusted quartile	0.4555(0.3799,0.5311)	0.4571(0.3854,0.5289)
MQE	0.4566(0.3381,0.5751)	0.4571(0.3381,0.5751)
HDL-C		
Weighted Median	-0.0687(-0.2069,0.0696)	-0.0689(-0.204,0.0663)
MR-median	-0.1758(-0.342,-0.0095)	-0.2222(-0.3931,-0.0514)
Adjusted quartile	-0.1641(-0.2747,-0.0535)	-0.1962(-0.3107,-0.0816)
MQE	-0.2209(-0.3665,-0.0753)	-0.2382(-0.3665,-0.0753)

The analysis is extended to investigate the causal effect within the quantile region for instruments selected from the p-value threshold of the genotype-exposure associations. Table 5.6 summarises the estimates from the conditional and weighted quantile estimators. Under the LDL-C exposure variable, the weighted percentile model has an increasing causal estimate with the percentile, whereas the mr-quantile model has the same estimate within the quantile space. The weighted percentile estimates for HDL-C exposure increase as the quantile level increases, whereas the MR-quantile estimates decrease.

Table 5.6 indicates how each approach differs as earlier stated in section 5.2.3. The estimate from weighted percentile is expected to correlate with the quantile level (i.e. the quantile level increases/decreases with the estimate), whereas estimates from the MR-quantile model vary within the quantiles. For insight into the influence of outlying

variants within the quantile space the Cook's distance for each genetic variant are measured for each exposure the results are denoted in figure 5.12 and 5.13 .

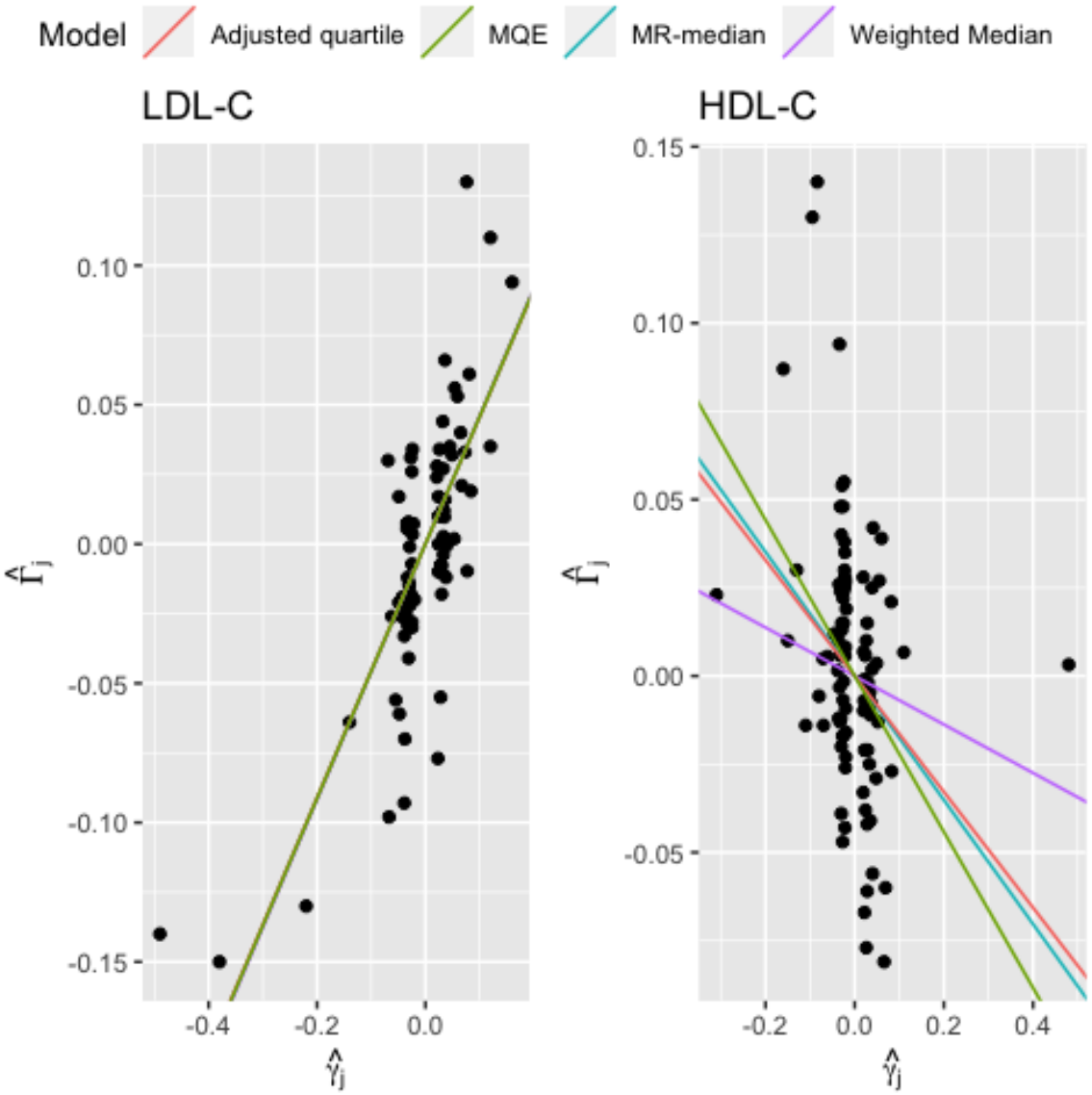

Figure 5.10: Scatter plot of the summary-level data including the genetic variants within the p-value threshold.

The estimates of the weighted percentile in table 5.6 was used to fit the density plot including the estimates as vertical lines in figure 5.11. Estimates lie below or

approximate the median except for the MR-Egger estimate.

Table 5.6: Estimates and confidence interval for weighted and conditional quantile space

	LDL-C		HDL-C	
Quantile	Weighted Percentile	MR-IVW quantile	Weighted Percentile	MR-IVW quantile
0.1	-0.1254(-0.3328,0.0821)	0.4571(0.314,0.6003)	-1.3333(-1.5828,-1.0839)	-0.0742(-0.2691,0.1207)
0.2	0.2821(0.1347,0.4295)	0.4571(0.3355,0.5787)	-0.6068(-0.7977,-0.416)	-0.0742(-0.2489,0.1005)
0.3	0.3001(0.1639,0.4363)	0.4571(0.3414,0.5729)	-0.355(-0.5366,-0.1733)	-0.0902(-0.254,0.0737)
0.4	0.4489(0.3188,0.5791)	0.4571(0.3397,0.5746)	-0.177(-0.3281,-0.0259)	-0.1758(-0.3407,-0.0108)
0.5	0.4606(0.3377,0.5834)	0.4571(0.3367,0.5776)	-0.0689(-0.204,0.0663)	-0.2222(-0.3958,-0.0486)
0.6	0.588(0.4571,0.7189)	0.4571(0.3411,0.5732)	0.001(-0.131,0.1329)	-0.2235(-0.4054,-0.0416)
0.7	0.6021(0.4594,0.7448)	0.4571(0.3266,0.5877)	0.0622(-0.0747,0.1991)	-0.25(-0.4388,-0.0612)
0.8	0.7668(0.6025,0.9311)	0.4571(0.3191,0.5951)	0.1408(-0.0119,0.2935)	-0.2625(-0.5061,-0.0189)
0.9	1.0526(0.8339,1.2712)	0.4571(0.3085,0.6058)	0.3858(0.1864,0.5853)	-0.2722(-0.5542,0.0097)

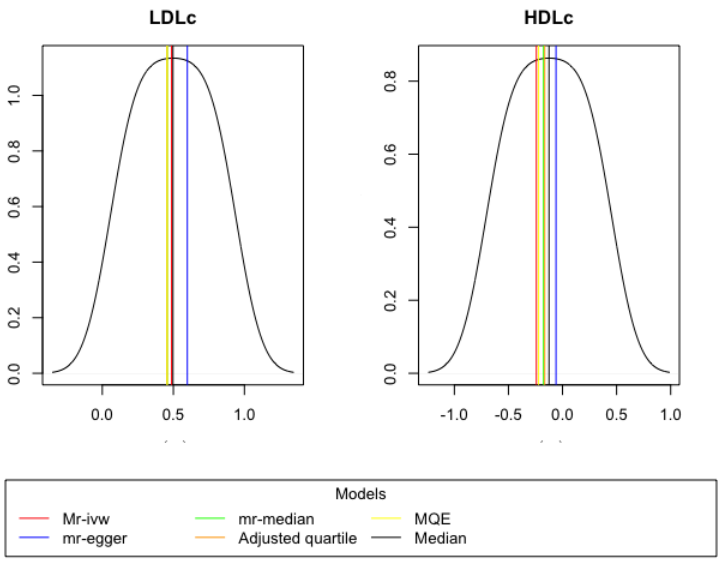

Figure 5.11: Density plot of weighted percentile the vertical lines indicate estimates from the models

Figure 5.12 shows the model may be affected by outlying instruments especially within lower quantile, however there is no evidence of outlying variants around the upper quantile ($\tau = 0.6, 0.7$ & 0.8). Tracing the estimates from the quantile level in table 5.6 show similar and significant values indicating the model is not affected by outliers. The results from investigating outlying instruments within the different quantiles for HDL-C as shown in figure 5.13 shows no outlying variant $\tau = 0.4$.

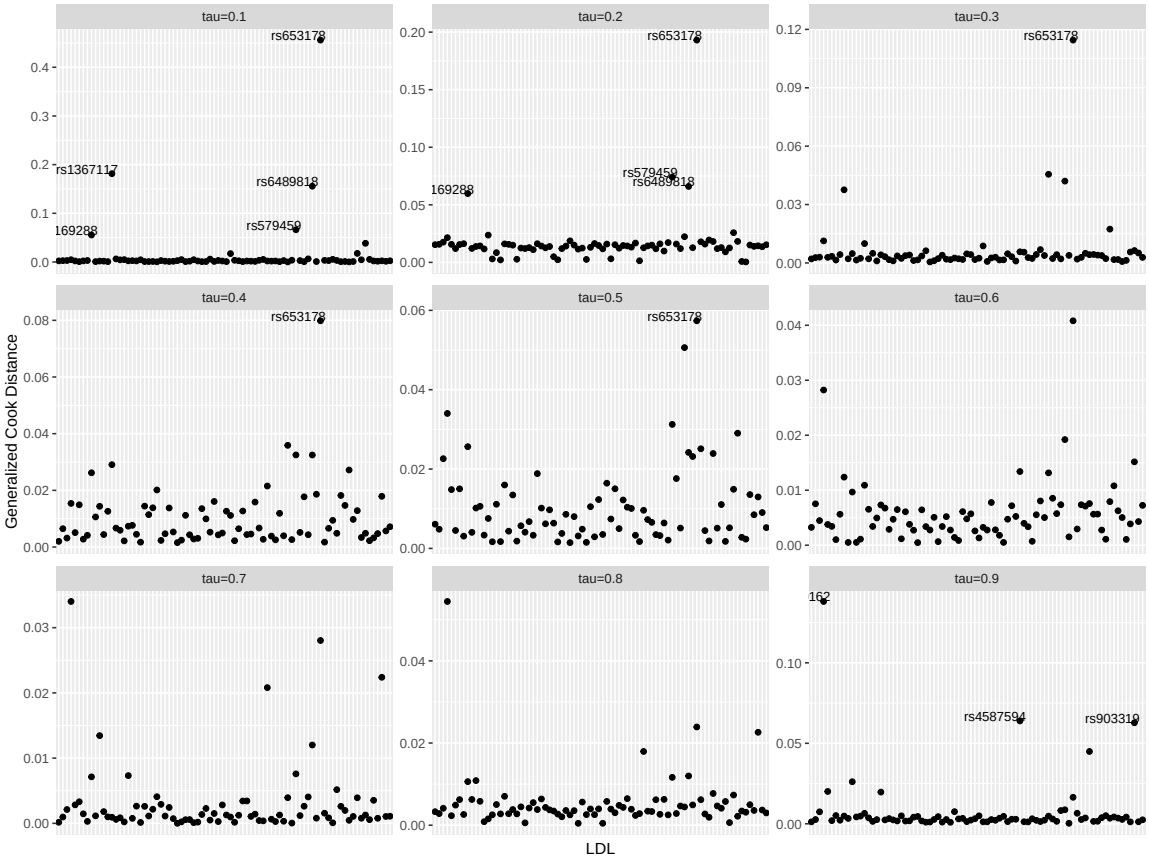

Figure 5.12: Diagnostic plots of each quantile using instruments related to LDL-C selected from p-value threshold

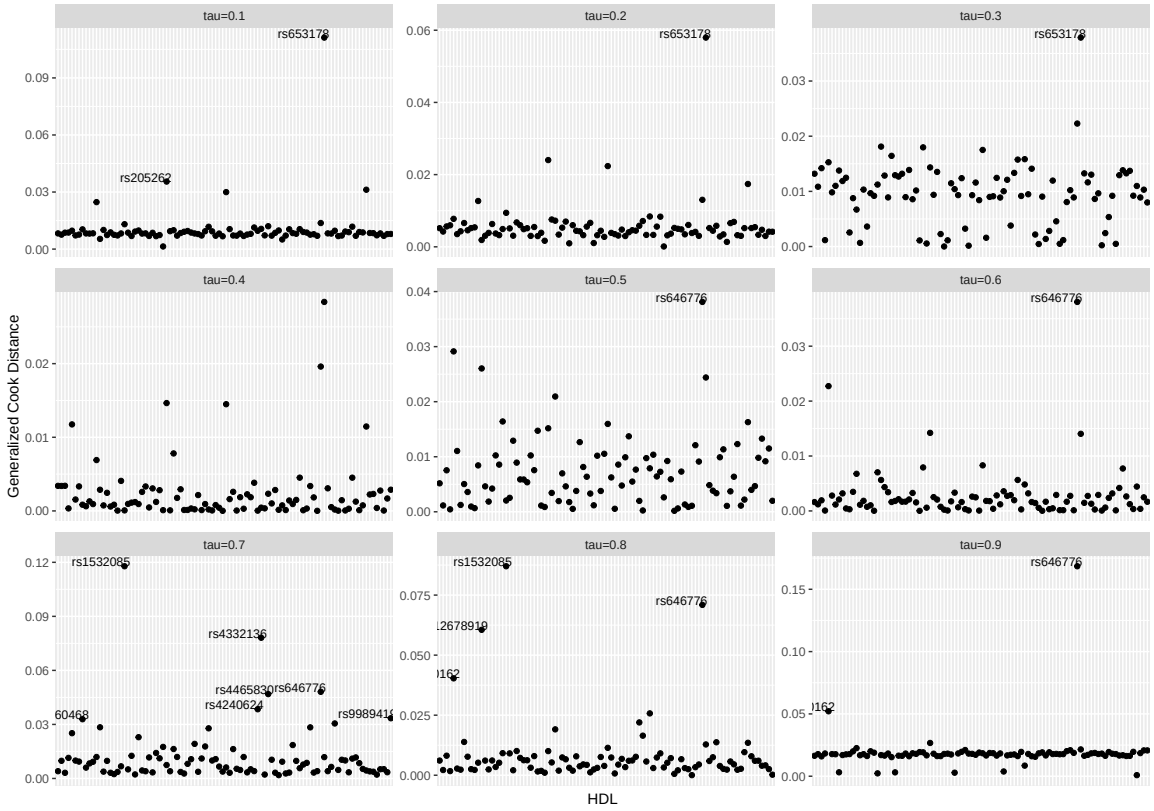

Figure 5.13: Diagnostic plots of each quantile using instruments related to HDL-C selected from p-value threshold

5.5.2 Analysis using all the instruments

This section discusses the findings of the analysis using all of the instruments within the quantile space, having discussed estimates from the quantile model in the earlier section. Figure 5.14 gives a graphical summary of the estimates, looking at the scatter plot indicates similar estimates within each exposure, leading to further investigation within the quantile space. Estimates from the weighted percentile model in table 5.7 has similar features with table 5.6 as the estimates correlates with quantile level. Estimates from MR-quantile models show changes as the quantile level increases.

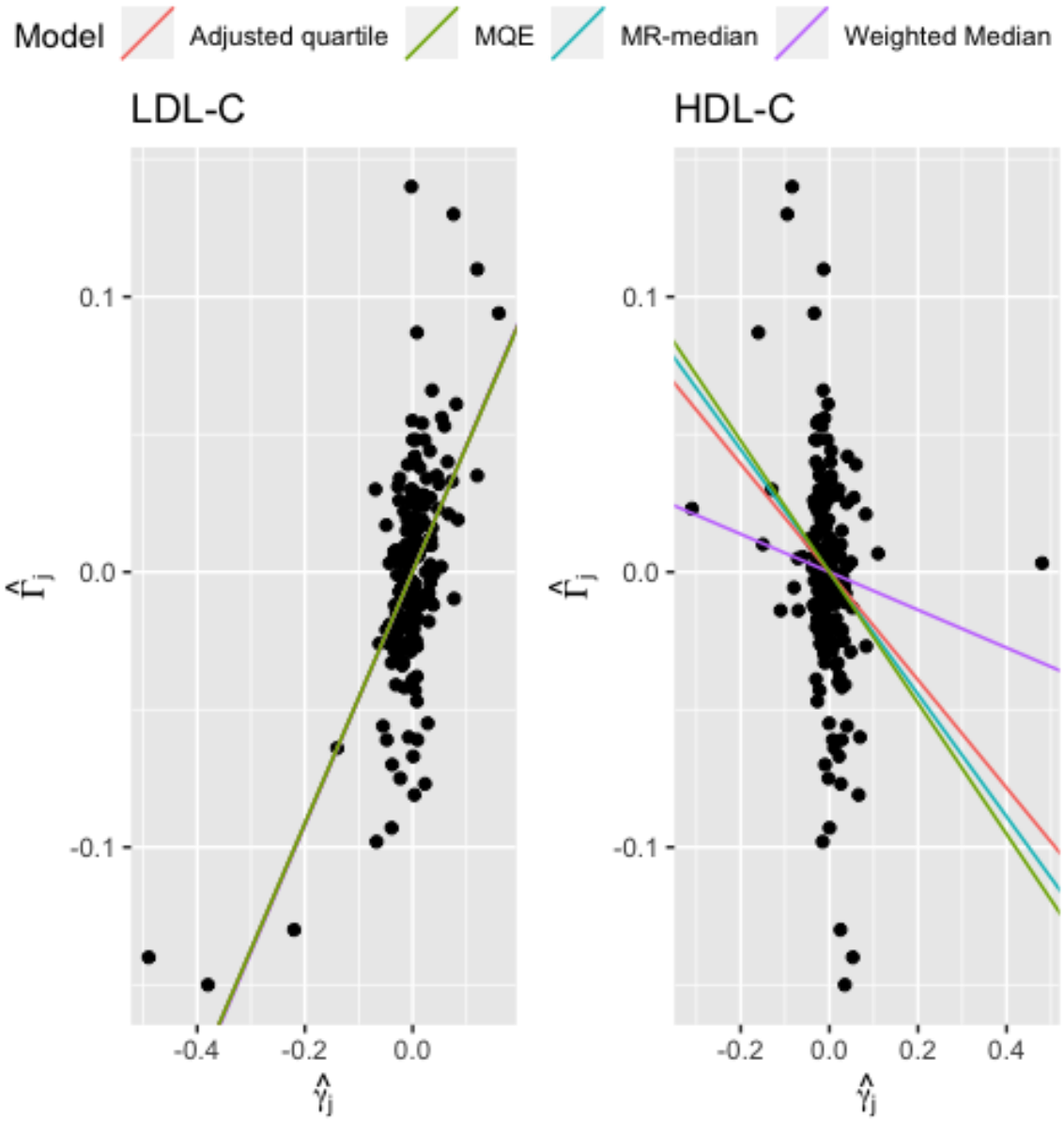

Figure 5.14: Scatter plot of the summary-level data including all instruments

Table 5.7: Estimates and confidence interval for weighted and conditional quantile space using all the instruments

	LDL-C		HDL-C	
Quantile	Weighted Percentile	MR-IVW quantile	Weighted Percentile	MR-IVW quantile
0.1	-0.1053(-0.3076,0.097)	0.4524(0.3148,0.59)	-1.2735(-1.5281,-1.0188)	-0.0742(-0.2392,0.0908)
0.2	0.2834(0.1305,0.4363)	0.4524(0.3235,0.5813)	-0.6015(-0.7963,-0.4067)	-0.0742(-0.253,0.1047)
0.3	0.3007(0.162,0.4395)	0.4524(0.3299,0.5749)	-0.3293(-0.4937,-0.1649)	-0.0902(-0.2661,0.0858)
0.4	0.4482(0.3197,0.5767)	0.4565(0.3304,0.5827)	-0.1354(-0.2846,0.0139)	-0.1324(-0.3057,0.041)
0.5	0.4588(0.3329,0.5846)	0.4565(0.3428,0.5702)	-0.0687(-0.2069,0.0696)	-0.1758(-0.3484,-0.0031)
0.6	0.5877(0.4534,0.7219)	0.4565(0.3496,0.5634)	-0.022(-0.1561,0.1122)	-0.2222(-0.4073,-0.0372)
0.7	0.5948(0.4504,0.7392)	0.4565(0.3242,0.5889)	0.0621(-0.0711,0.1953)	-0.2235(-0.4279,-0.0192)
0.8	0.756(0.5965,0.9155)	0.4565(0.3136,0.5994)	0.1326(-0.0197,0.2849)	-0.2308(-0.4616,1e-04)
0.9	1.036(0.8136,1.2583)	0.4571(0.3014,0.6129)	0.3407(0.1463,0.5351)	-0.25(-0.5198,0.0198)

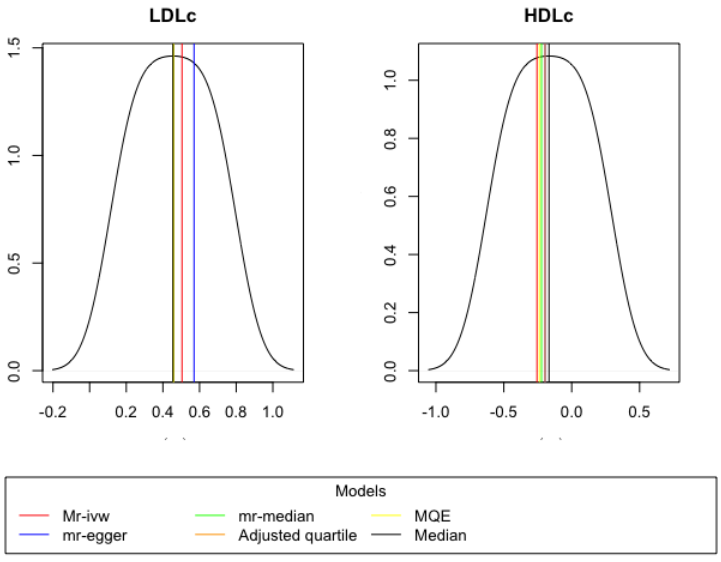

Figure 5.15: Density plot of weighted percentile the vertical lines indicate estimates from the models

Further investigation of the outlying effects within the quantile space shows more variability due to the inclusion of more instruments especially in figure 5.16.

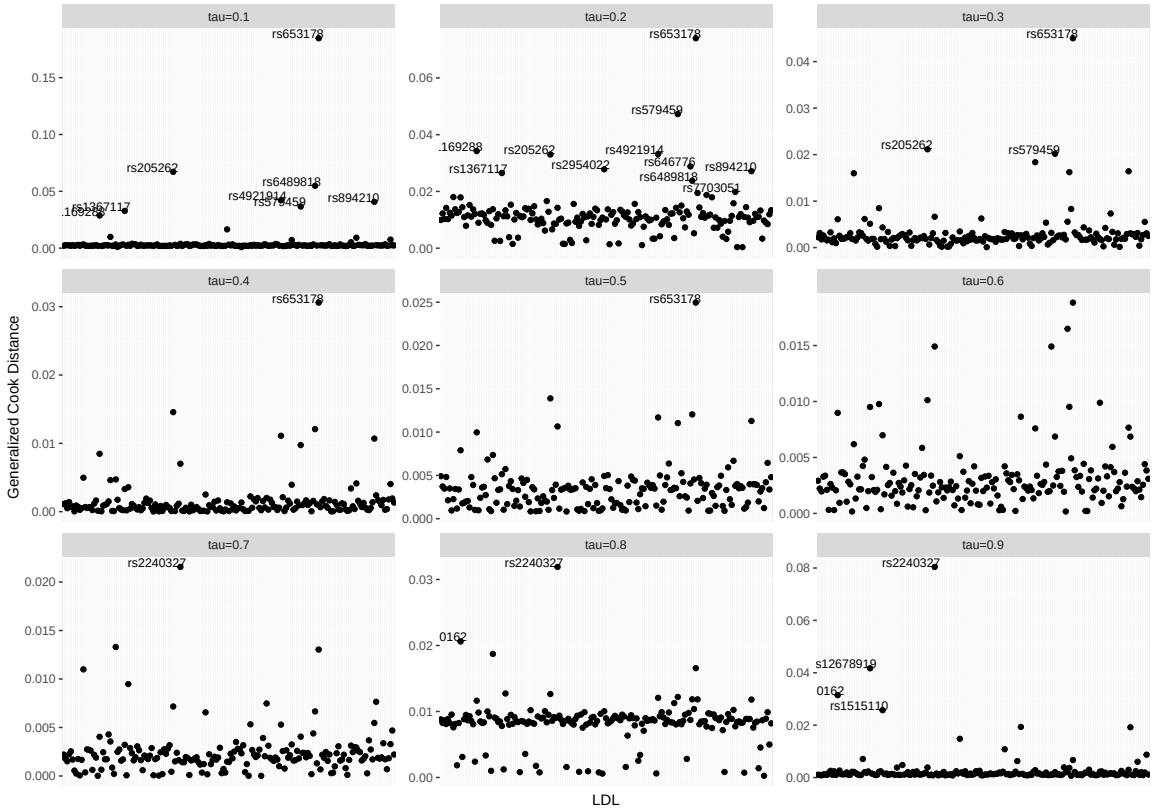

Figure 5.16: Diagnostic plots of each quantile in LDL-C using all instruments

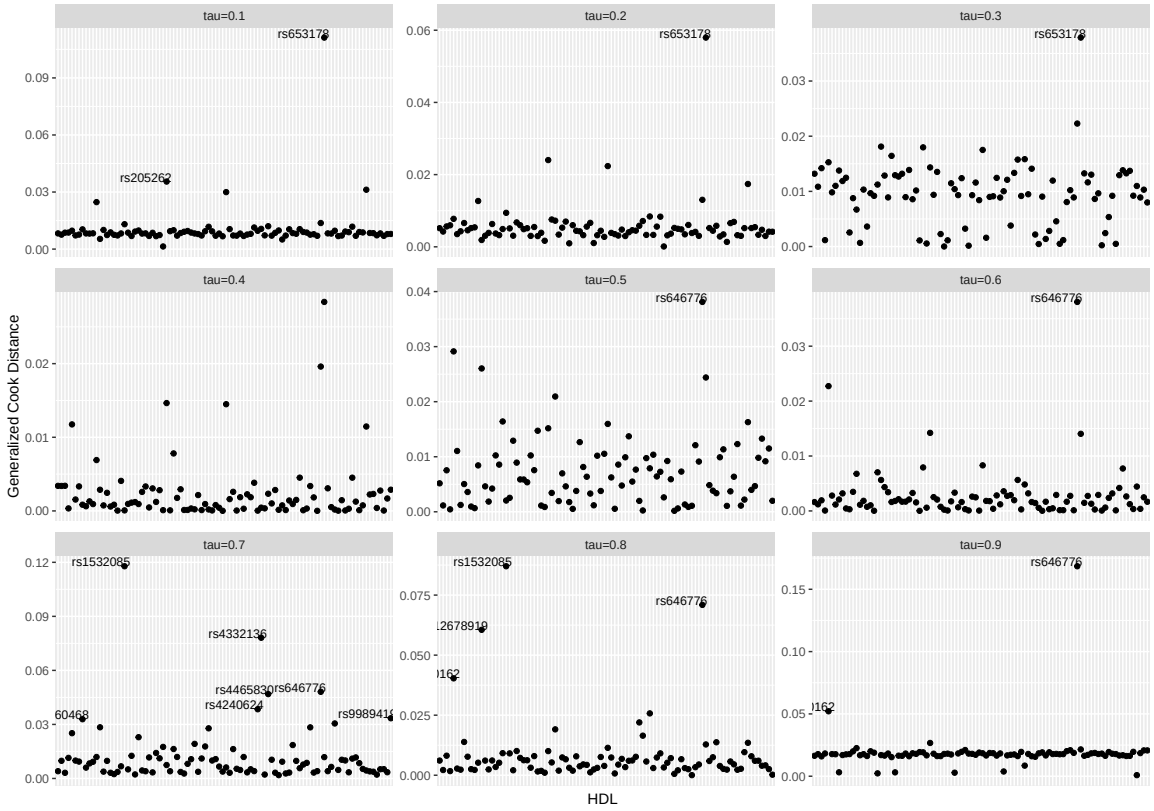

Figure 5.17: Diagnostic plots of each quantile related to HDL-C using all instruments

5.6 Discussion

This research investigates quantile models applied to summary level Mendelian randomization analyses. The MR-median estimator is known to have good properties, from this alternative estimators were proposed. The properties of the estimators are investigated using simulations and data applications. The simulations show consistent estimates in the presence of pleiotropy and invalid instruments. From the data application, this research also introduced the use of the generalised Cook's distance to investigate the presence of outlying instruments.

Grant & Burgess (2020), in a simulation study, introduced the multivariate quantile median estimator as a multivariate form of the weighted median estimator and is implemented in the `MendelianRandomization` package (Yavorska & Burgess, 2017). Results from simulation within quantiles show the estimates from the models differ especially in pleiotropic scenarios, this is also supported by the estimates in the data application section. According to the data application, the estimates generated from the quantile space of the MR-quantile model do not correspond to the quantile level in a linear fashion. The estimators proposed here are designed to take advantage of the quartiles and mode of estimates found within the conditional quantile space of estimates. Compared to weighted estimators, quantile regression models produce consistent estimates. The mode and quartiles estimators are being used to support the median of the quantile regression model as a sensitivity analyses.

Further research could extend the MVMR-Median model by Grant & Burgess (2020) to include a penalty parameter in the design of a high-throughput MVMR model. Additionally, due to the limitations of the parametric bootstrap to estimate the standard error of these estimates, there is scope for further research into semi-parametric methods. For example, assuming the instrument-outcome association follows an asymmetric Laplace distribution to derive the maximum likelihood estimate (Benites et al., 2015).

Finally, conditional median density and weighted median are established sensitivity analyses in MR studies. The advantages of investigating the different quantiles is that they can produce consistent estimates in the presence of heterogeneity and give a broader understanding of the conditional distribution of the effects of the exposure on a disease outcome.

Chapter 6

Discussion & conclusions

6.1 Discussion

This thesis investigated and proposed statistical models to apply in a MR analysis using summary-level data. The research consists of a literature review, the introduction of software to apply Bayesian summary-level data models, introducing hierarchical Bayesian models suitable for high-throughput summary-level data, and the investigation of quantile estimators to summary-level data MR analyses. Each area is discussed, and suggestions are made for future research.

6.1.1 Bayesian implementation of estimators for summary level MR analyses: mrbayes package

Chapter 3 is a report on the package `mrbayes` which implements Bayesian models for two-sample summary-level data MR analyses. The package implements a range of prior distributions for the IVW, MR-Egger, and Radial Egger models, and their multivariate versions. The Monte-Carlo Markov chain sampling is performed within either JAGs or Stan. Introducing this package gives the opportunity for more applied

Bayesian analysis within summary-level MR research. Extension to the package could include incorporating additional likelihood models and creating a web application for univariate and multivariate Bayesian MR models similar to the frequentist estimators implemented within the MR-Base web application https://www.mrbase.org/ (Hemani et al., 2018).

6.1.2 Bayesian models in MVMR

Chapter 4 shows the effectiveness of hierarchical shrinkage priors when applied to models fitted to high-throughput data. The setup of hierarchical models in MVMR study designs is efficient in estimating causal effects especially for pleiotropic exposures. The data application highlights how the inclusion of large numbers of exposures can affect multivariate models in such data applications. The research concluded with recommendations relating to regularizing the hierarchical shrinkage priors and further research towards establishing predictive models within summary-level data MR analyses.

6.1.3 Efficient information from quantile estimates

The information from the quantile estimation is used in this analysis to gain a better understanding of the conditional distribution of the effects of exposure on disease outcomes in chapter 5. The conditional median is a robust estimator and useful for sensitivity analysis in MR. This research investigated the estimates from the conditional quantile space using quantile regression in an MR analysis to estimate the causal effect, and introduced the use of Cook's distance to help identify outlying instruments. From the data example the estimates were approximately constant in all the quantile spaces. To make the most of the estimates from the quantiles, we proposed two estimators as alternatives to the conditional median for sensitivity analysis. The first alternative estimator is derived from the median of the upper and lower quartile estimates and the

second estimator is the mode of the estimates between the 0.1 and 0.9 quantiles. The estimators were shown to have low bias in the presence of invalid instruments, although bootstrapping was required to calculate their standard errors. Further research could include investigating the use of semi-parametric methods for quantile estimation.

6.2 Future directions

As more GWAS studies are conducted especially in large consortia with high dimensional datasets, methods for MR will continue to be developed. I will briefly mention potential MR models to consider for future use.

6.2.1 MR models for time-varying exposures

There is a limitation in using MR to investigate lifetime effects of certain exposures, for example, body mass index on disease outcomes because data for MR studies are mostly measured at one time point which do not capture information at multiple time points (Davey Smith & Ebrahim, 2003). Labrecque & Swanson (2019) shows the potential bias of estimates from MR models using time-varying exposures. When considering time-varying exposures, functional data analysis methods have been proposed to estimate causal effects, for example, Yao et al. (2005) developed a method known as principal analysis by conditional expectation (PACE) to recover the underlying trajectories of time-varying exposures. Cao et al. (2016) extended the PACE method within an individual-level data study design and proposed two models one of which assumes the time-varying exposure has a cumulative effect on the outcome. The second model uses functional regression methods to satisfy the assumption of the genetic effect on the time-varying exposure changing with time. Functional regression models could be extended to summary-level data.

6.2.2 Using predictive models in high-throughput MR

The increased number of published GWAS studies has produced more summary level estimates of instrument-exposure relationships, including blood lipids and metabolites. These metabolites and lipids are linked to a variety of outcomes, and evidence of causality can aid in disease prediction. Howey et al. (2020) applied Bayesian networks to assess causality in complex data. We propose extending Bayesian networks by incorporating machine learning techniques such as using predictive power to select causal exposures or phenotypes on a large scale.

6.3 Conclusion

This thesis has described and reviewed the MR approach and then developed and assessed Bayesian methods for genotype summary level data for application in MR analyses. This research shows how prior distributions can be used to make MR models more robust to the standard IV assumptions.

Bibliography

Akiyama, Masato, Okada, Yukinori, Kanai, Masahiro, Takahashi, Atsushi, Momozawa, Yukihide, Ikeda, Masashi, Iwata, Nakao, Ikegawa, Shiro, Hirata, Makoto, Matsuda, Koichi, et al. 2017. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nature Genetics, **49**(10), 1458–1467.

Angrist, Joshua D, & Pischke, Jörn-Steffen. 2008. Mostly Harmless Econometrics: An empiricist's companion. Princeton University Press.

Baigent, Colin, Landray, Martin, Leaper, Craig, Altmann, Paul, Armitage, Jane, Baxter, Alex, Cairns, Hugh S, Collins, Rory, Foley, Robert N, Frighi, Valeria, et al. 2005. First United Kingdom Heart and Renal Protection (UK-HARP-I) study: biochemical efficacy and safety of simvastatin and safety of low-dose aspirin in chronic kidney disease. American Journal of Kidney Diseases, **45**(3), 473–484.

Banks, Emily, Reeves, Gillian, Beral, Valerie, Bull, Diana, Crossley, Barbara, Simmonds, Moya, Hilton, Elizabeth, Bailey, Stephen, Barrett, Nigel, Briers, Peter, et al. 2004. Influence of personal characteristics of individual women on sensitivity and specificity of mammography in the Million Women Study: cohort study. BMJ, **329**(7464), 477.

Benites, Luis E, Lachos, Víctor H, & Vilca, Filidor E. 2015. Case-deletion diagnostics for Quantile regression using the asymmetric Laplace distribution. arXiv preprint arXiv:1509.05099.

Berzuini, Carlo, Guo, Hui, Burgess, Stephen, & Bernardinelli, Luisa. 2020. A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. Biostatistics, **21**, 86–101.

Betancourt, Michael. 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434.

Betancourt, Michael, & Girolami, Mark. 2015. Hamiltonian Monte Carlo for hierarchical models. Current Trends in Bayesian Methodology with Applications, **79**(30), 2–4.

Bhadra, Anindya, Datta, Jyotishka, Polson, Nicholas G, Willard, Brandon, et al. 2017. The horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis, **12**(4), 1105–1131.

Bound, John, Jaeger, David A, & Baker, Regina 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, Jack, Davey Smith, George, & Burgess, Stephen. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology, **44**(2), 512–525.

Bowden, Jack, Del Greco M, Fabiola, Minelli, Cosetta, Davey Smith, George, Sheehan, Nuala A, & Thompson, John 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, Jack, Davey Smith, George, Haycock, Philip C, & Burgess, Stephen. 2016b. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic Epidemiology, **40**(4), 304–314.

Bowden, Jack, Del Greco M, Fabiola, Minelli, Cosetta, Davey Smith, George, Sheehan, Nuala, & Thompson, John. 2017. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Statistics in Medicine, **36**(11), 1783–1802.

Bowden, Jack, Spiller, Wesley, Del Greco M, Fabiola, Sheehan, Nuala, Thompson, John, Minelli, Cosetta, & Davey Smith, George. 2018. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression. International Journal of Epidemiology, **47**(4), 1264–1278.

Bowden, Jack, Del Greco M, Fabiola, Minelli, Cosetta, Lawlor, Debbie, Zhao, Qingyuan, Sheehan, Nuala, Thompson, John, & Davey Smith, George. 2019. Improving the accuracy of two-sample summary data Mendelian randomization: moving beyond the NOME assumption. International Journal of Epidemiology, **48**(3), 728–742.

Brennan, Paul. 2004. Commentary: Mendelian randomization and gene–environment interaction. International Journal of Epidemiology, **33**(1), 17–21.

Brion, Marie-Jo A, Shakhbazov, Konstantin, & Visscher, Peter M. 2013. Calculating statistical power in Mendelian randomization studies. International Journal of Epidemiology, **42**(5), 1497–1501.

Bucur, Ioan Gabriel, Claassen, Tom, & Heskes, Tom. 2020. Inferring the direction of a causal link and estimating its effect via a Bayesian Mendelian randomization approach. Statistical Methods in Medical Research, **29**(4), 1081–1111.

Burgess, Stephen, & Thompson, Simon G. 2012. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. Statistics in Medicine, **31**(15), 1582–1600.

Burgess, Stephen, & Thompson, Simon G. 2015. Mendelian randomization: methods for using genetic variants in causal estimation. CRC Press.

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

Burgess, Stephen, Butterworth, Adam, & Thompson, Simon G. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic Epidemiology, **37**(7), 658–665.

Burgess, Stephen, Davies, Neil M, & Thompson, Simon G. 2014. Instrumental variable analysis with a nonlinear exposure–outcome relationship. Epidemiology, **25**(6), 877.

Burgess, Stephen, Daniel, Rhian M, Butterworth, Adam S, Thompson, Simon G, & Consortium, EPIC-InterAct. 2015a. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. International Journal of Epidemiology, **44**(2), 484–495.

Burgess, Stephen, Dudbridge, Frank, & Thompson, Simon G. 2015b. Re:"Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects". American Journal of Epidemiology, **181**(4), 290–291.

Burgess, Stephen, Scott, Robert A, Timpson, Nicholas J, Davey Smith, George, Thompson, Simon G, Consortium, EPIC-InterAct, et al. 2015c. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. European Journal of Epidemiology, **30**(7), 543–552.

Burgess, Stephen, Bowden, Jack, Dudbridge, Frank, & Thompson, Simon G. 2016. Robust instrumental variable methods using multiple candidate instruments with application to Mendelian randomization. arXiv preprint arXiv:1606.03729.

Burgess, Stephen, Zuber, Verena, Gkatzionis, Apostolos, & Foley, Christopher N. 2018. Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid. International Journal of Epidemiology, **47**(4), 1242–1254.

Burgess, Stephen, Smith, George Davey, Davies, Neil M, Dudbridge, Frank, Gill, Dipender, Glymour, M Maria, Hartwig, Fernando P, Holmes, Michael V, Minelli, Cosetta, Relton, Caroline L, et al. 2019. Guidelines for performing Mendelian randomization investigations. Wellcome Open Research, **4**.

Burgess, Stephen, Foley, Christopher N, Allara, Elias, Staley, James R, & Howson, Joanna MM. 2020. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nature Communications, **11**(376).

Cao, Ying, Rajan, Suja S, & Wei, Peng. 2016. Mendelian randomization analysis of a time-varying exposure for binary disease outcomes using functional data analysis methods. Genetic Epidemiology, **40**(8), 744–755.

Cardon, Lon R, & Palmer, Lyle J. 2003. Population stratification and spurious allelic association. The Lancet, **361**(9357), 598–604.

Carvalho, Carlos M, Polson, Nicholas G, & Scott, James G. 2009. Handling sparsity via the horseshoe. Pages 73–80 of: Artificial Intelligence and Statistics.

Carvalho, Carlos M, Polson, Nicholas G, & Scott, James G. 2010. The horseshoe estimator for sparse signals. Biometrika, **97**(2), 465–480.

Chen, Zhengming, Chen, Junshi, Collins, Rory, Guo, Yu, Peto, Richard, Wu, Fan, & Li, Liming. 2011. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. International Journal of Epidemiology, **40**(6), 1652–1666.

Danesh, J, et al. 2008. Collaborative pooled analysis of data on C-reactive protein gene variants and coronary disease: judging causality by Mendelian randomisation. European Journal of Epidemiology, **23**(8), 531–540.

Davey Smith, George. 2006. Randomised by (your) god: robust inference from an observational study design. Journal of Epidemiology & Community Health, **60**(5), 382–388.

Davey Smith, George, & Ebrahim, Shah. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology, **32**(1), 1–22.

Davey Smith, George, & Ebrahim, Shah. 2008. A cautionary note on the use of Mendelian randomization to infer causation in observational epidemiology: Reply. International Journal of Epidemiology, **37**(2), 416–417.

Debat, Vincent, & David, Patrice. 2001. Mapping phenotypes: canalization, plasticity and developmental stability. Trends in ecology & evolution, **16**(10), 555–561.

Didelez, Vanessa, & Sheehan, Nuala. 2007. Mendelian randomization as an instrumental variable approach to causal inference. Statistical Methods in Medical Research, **16**(4), 309–330.

Didelez, Vanessa, Meng, Sha, Sheehan, Nuala A, et al. 2010. Assumptions of IV methods for observational epidemiology. Statistical Science, **25**(1), 22–40.

Do, Ron, Willer, Cristen J, Schmidt, Ellen M, Sengupta, Sebanti, Gao, Chi, Peloso, Gina M, Gustafsson, Stefan, Kanoni, Stavroula, Ganna, Andrea, Chen, Jin, et al. 2013. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nature Genetics, **45**(11), 1345.

Elwood, Mark. 2017. Critical appraisal of epidemiological studies and clinical trials. Oxford University Press.

Flister, Michael J, Tsaih, Shirng-Wern, O'Meara, Caitlin C, Endres, Bradley, Hoffman, Matthew J, Geurts, Aron M, Dwinell, Melinda R, Lazar, Jozef, Jacob, Howard J, & Moreno, Carol. 2013. Identifying multiple causative genes at a single GWAS locus. Genome Research, **23**(12), 1996–2002.

Foley, Christopher N, Mason, Amy M, Kirk, Paul DW, & Burgess, Stephen. 2021. MR-Clust: clustering of genetic variants in Mendelian randomization with similar causal estimates. Bioinformatics, **37**(4), 531–541.

Frayling, Timothy M, Timpson, Nicholas J, Weedon, Michael N, Zeggini, Eleftheria, Freathy, Rachel M, Lindgren, Cecilia M, Perry, John RB, Elliott, Katherine S, Lango, Hana, Rayner, Nigel W, et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science, **316**(5826), 889–894.

Gelman, Andrew, et al. 2006a. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis, **1**(3), 515–534.

Gelman, Andrew, et al. 2006b. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis, **1**(3), 515–534.

Grant, Andrew J, & Burgess, Stephen. 2020. Pleiotropy robust methods for multivariable Mendelian randomization. arXiv, 2008.11997.

Gray, Richard, & Wheatley, Keith. 1991. How to avoid bias when comparing bone marrow transplantation with chemotherapy. Bone Marrow Transplantation, **7**, 9–12.

Greenland, Sander. 2000. An introduction to instrumental variables for epidemiologists. International Journal of Epidemiology, **29**(4), 722–729.

Greenland, Sander, Pearl, Judea, & Robins, James M. 1999. Causal diagrams for epidemiologic research. Epidemiology, 37–48.

Griffin, Jim E, Brown, Philip J, et al. 2010. Inference with normal-gamma prior distributions in regression problems. Bayesian Analysis, **5**(1), 171–188.

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

Hardy, Godfrey H, et al. 1908. Mendelian proportions in a mixed population. Science, **28**(706), 49–50.

Hartwig, Fernando Pires, Davies, Neil Martin, Hemani, Gibran, & Davey Smith, George. 2016. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. International Journal of Epidemiology, **45**, 1717–1726.

Hartwig, Fernando Pires, Davey Smith, George, & Bowden, Jack. 2017. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. International Journal of Epidemiology, **46**(6), 1985–1998.

Hemani, Gibran, Zheng, Jie, Wade, Kaitlin H, Laurin, Charles, Elsworth, Benjamin, Burgess, Stephen, Bowden, Jack, Langdon, Ryan, Tan, Vanessa, Yarmolinsky, James, et al. 2016. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. BioRxiv, 078972.

Hemani, Gibran, Zheng, Jie, Wade, Kaitlin H, Laurin, Charles, Elsworth, Benjamin, Burgess, Stephen, Bowden, Jack, Langdon, Ryan, Tan, Vanessa, Yarmolinsky, James, A, Hashem, Timpson, Nicholas J, Evans, David M, Relton, Caroline, Martin, Richard M, Smith, George Davey, Gaunt, Tom R, & Haycock, Philip C. 2018. The

MR-Base platform supports systematic causal inference across the human phenome. eLife, **7**, e34408.

Hernán, Miguel Angel. 2004. A definition of causal effect for epidemiological research. Journal of Epidemiology & Community Health, **58**(4), 265–271.

Higgins, Julian, & Thompson, Simon G. 2002. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, **21**(11), 1539–1558.

Hirschhorn, Joel N, Lohmueller, Kirk, Byrne, Edward, & Hirschhorn, Kurt. 2002. A comprehensive review of genetic association studies. Genetics in Medicine, **4**(2), 45–61.

Howey, Richard, Shin, So-Youn, Relton, Caroline, Davey Smith, George, & Cordell, Heather J. 2020. Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data. PLoS Genetics, **16**(3), e1008198.

Johnston, KM, Gustafson, P, Levy, AR, & Grootendorst, P. 2008. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. Statistics in Medicine, **27**(9), 1539–1556.

Jones, EM, Thompson, JR, Didelez, V, & Sheehan, NA. 2012. On the choice of parameterisation and priors for the Bayesian analyses of Mendelian randomisation studies. Statistics in Medicine, **31**(14), 1483–1501.

Kang, Hyunseung, Zhang, Anru, Cai, T Tony, & Small, Dylan 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, MartijnB. 1986. Apoupoprotein E isoforms, serum cholesterol, and cancer. The Lancet, **327**(8479), 507–508.

Kettunen, Johannes, Demirkan, Ayşe, Würtz, Peter, Draisma, Harmen HM, Haller, Toomas, Rawal, Rajesh, Vaarhorst, Anika, Kangas, Antti J, Lyytikäinen, Leo-Pekka, Pirinen, Matti, et al. 2016. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nature Communications, **7**, 11122.

Khaw, Kay-Tee, Bingham, Sheila, Welch, Ailsa, Luben, Robert, Wareham, Nicholas, Oakes, Suzy, & Day, Nicholas. 2001. Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. The Lancet, **357**(9257), 657–663.

Koenker, Roger. 2005. Quantile Regression. Cambridge University Press.

Koenker, Roger. 2017. Quantile regression: 40 years on. Annual Review of Economics, **9**, 155–176.

Koenker, Roger, & Bassett Jr, Gilbert. 1978. Regression quantiles. Econometrica, 33–50.

Labrecque, Jeremy A, & Swanson, Sonja A. 2019. Interpretation and potential biases of Mendelian randomization estimates with time-varying exposures. American Journal of Epidemiology, **188**(1), 231–238.

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

Li, Sai. 2017. Mendelian randomization when many instruments are invalid: hierarchical empirical Bayes estimation. arXiv preprint arXiv:1706.01389.

Lippman, Scott M, Klein, Eric A, Goodman, Phyllis J, Lucia, M Scott, Thompson, Ian M, Ford, Leslie G, Parnes, Howard L, Minasian, Lori M, Gaziano, J Michael, Hartline, Jo Ann, et al. 2009. Effect of selenium and vitamin E on risk of prostate

cancer and other cancers: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). Journal of the American Medical Association, **301**(1), 39–51.

Little, Julian, & Khoury, Muin J. 2003. Mendelian randomisation: a new spin or real progress? The Lancet, **362**(9388), 930–930.

Lock, Robert Heath, Doncaster, Leonard, & Woolf, Bella Sidney. 1916. Recent progress in the study of variation, heredity, and evolution. Dutton.

Lyngdoh, Tanica, Vuistiner, Philippe, Marques-Vidal, Pedro, Rousson, Valentin, Waeber, Gérard, Vollenweider, Peter, & Bochud, Murielle. 2012. Serum uric acid and adiposity: deciphering causality using a bidirectional Mendelian randomization approach. PLoS One, **7**(6), e39321.

MacArthur, Jacqueline, Bowler, Emily, Cerezo, Maria, Gil, Laurent, Hall, Peggy, Hastings, Emma, Junkins, Heather, McMahon, Aoife, Milano, Annalisa, Morales, Joannella, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research, **45**(D1), D896–D901.

Malik, Rainer, Chauhan, Ganesh, Traylor, Matthew, Sargurupremraj, Muralidharan, Okada, Yukinori, Mishra, Aniket, Rutten-Jacobs, Loes, Giese, Anne-Katrin, Van Der Laan, Sander W, Gretarsdottir, Solveig, et al. 2018. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nature Genetics, **50**(4), 524–537.

Neal, Radford M, et al. 2012. MCMC using Hamiltonian dynamics. arXiv, 1206.1901.

Nica, Alexandra C, Montgomery, Stephen B, Dimas, Antigone S, Stranger, Barbara E, Beazley, Claude, Barroso, Inês, & Dermitzakis, Emmanouil T. 2010. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genetics, **6**(4), e1000895.

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

Park, Trevor, & Casella, George. 2008. The bayesian lasso. Journal of the American Statistical Association, **103**(482), 681–686.

Pauling, Linus, Itano, Harvey A, Singer, Seymour J, & Wells, Ibert C. 1949. Sickle cell anemia, a molecular disease. Science, **110**(2865), 543–548.

Pearl, Judea. 2009. Causality. 2 edn. Cambridge, UK: Cambridge University Press.

Piironen, Juho, & Vehtari, Aki. 2016. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. arXiv preprint arXiv:1610.05559.

Piironen, Juho, Vehtari, Aki, et al. 2017. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics, **11**(2), 5018–5051.

Plummer, Martyn. 2012. JAGS Version 3.3.0 user manual. International Agency for Research on Cancer, Lyon, France.

Plummer, Martyn. 2018. rjags: Bayesian Graphical Models using MCMC. R package version 4-8.

Polson, Nicholas G, Scott, James G, et al. 2012. On the half-Cauchy prior for a global scale parameter. Bayesian Analysis, **7**(4), 887–902.

Porcu, Eleonora, Rüeger, Sina, Lepik, Kaido, Santoni, Federico A, Reymond, Alexandre, & Kutalik, Zoltán. 2019. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nature Communications, **10**(1), 1–12.

Qi, Guanghao, & Chatterjee, Nilanjan. 2019. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. Nature Communications, **10**(1), 1–10.

Rees, Jessica MB, Wood, Angela M, Dudbridge, Frank, & Burgess, Stephen. 2019. Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. PloS One, **14**(9).

Relton, Caroline L, & Davey Smith, George. 2012. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. International Journal of Epidemiology, **41**(1), 161–176.

Richmond, Rebecca, Wade, Kaitlin, Corbin, Laura, Bowden, Jack, Hemani, Gibran, Timpson, Nicholas, & Davey Smith, George. 2017. Investigating the role of insulin in increased adiposity: Bi-directional Mendelian randomization study. bioRxiv, 155739.

Robins, James. 1986. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling, **7**(9-12), 1393–1512.

Rothman, Kenneth, Greenland, Sander, & Timothy, Lash. 2008. Modern Epidemiology. Cambridge University Press.

Salanti, Georgia, Amountza, Georgia, Ntzani, Evangelia E, & Ioannidis, John PA. 2005. Hardy–Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. European Journal of Human Genetics, **13**(7), 840–848.

Schmidt, AF, & Dudbridge, F. 2017. Mendelian randomization with Egger pleiotropy correction and weakly informative Bayesian priors. International Journal of Epidemiology, **47**(4), 1217–1228.

Shapland, Chin Yang, Thompson, John R, & Sheehan, Nuala A. 2019. A Bayesian approach to Mendelian randomisation with dependent instruments. Statistics in Medicine, **38**(6), 985–1001.

Shapland, Chin Yang, Zhao, Qingyuan, & Bowden, Jack. 2020. Profile-likelihood Bayesian model averaging for two-sample summary data Mendelian randomization in the presence of horizontal pleiotropy. BioRxiv.

Sheehan, Nuala A, & Didelez, Vanessa. 2020. Epidemiology, genetic epidemiology and Mendelian randomisation: more need than ever to attend to detail. Human Genetics, **139**(1), 121–136.

Silverman, Bernard W. 2018. Density estimation for statistics and data analysis. Routledge.

Spiller, Wes, & Bowden, Jack. 2019. RadialMR: A package for implementing radial inverse variance weighted and MR-Egger methods. R package version 0.4.

Staiger, Douglas, & Stock, James H. 1994. Instrumental variables regression with weak instruments. Tech. rept. National Bureau of Economic Research.

Staley, James R, & Burgess, Stephen. 2017. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. Genetic Epidemiology, **41**(4), 341–352.

Stan Development Team. 2018. RStan: the R interface to Stan. R package version 2.18.2.

Sudlow, Cathie, Gallacher, John, Allen, Naomi, Beral, Valerie, Burton, Paul, Danesh, John, Downey, Paul, Elliott, Paul, Green, Jane, Landray, Martin, et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Medicine, **12**(3).

Terza, Joseph V, Basu, Anirban, & Rathouz, Paul J. 2008. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. Journal of Health Economics, **27**(3), 531–543.

Thomas, Duncan C, Lawlor, Debbie A, & Thompson, John R. 2007. Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al. Annals of Epidemiology, **7**(17), 511–513.

Thompson, John R, Minelli, Cosetta, Bowden, Jack, Del Greco, Fabiola M, Gill, Dipender, Jones, Elinor M, Shapland, Chin Yang, & Sheehan, Nuala A. 2017. Mendelian randomization incorporating uncertainty about pleiotropy. Statistics in Medicine, **36**(29), 4627–4645.

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

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

Verbanck, Marie, Chen, Chia-yen, Neale, Benjamin, & Do, Ron. 2018. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature Genetics, **50**(5), 693–698.

Vineis, Paolo. 2004. A self-fulfilling prophecy: are we underestimating the role of the environment in gene–environment interaction research? International Journal of Epidemiology, **33**(5), 945–946.

Wald, Abraham. 1940. The fitting of straight lines if both variables are subject to error. The Annals of Mathematical Statistics, **11**(3), 284–300.

Wang, Wenjing, Cook, Di, & Wang, Earo. 2017. quokar: Quantile Regression Outlier Diagnostics with K Left Out Analysis. R package version 0.1.0.

Wardle, Jane, Carnell, Susan, Haworth, Claire MA, Farooqi, I Sadaf, O'Rahilly, Stephen, & Plomin, Robert. 2008. Obesity associated genetic variation in FTO is associated with diminished satiety. The Journal of Clinical Endocrinology & Metabolism, **93**(9), 3640–3643.

Waterworth, Dawn M, Ricketts, Sally L, Song, Kijoung, Chen, Li, Zhao, Jing Hua, Ripatti, Samuli, Aulchenko, Yurii S, Zhang, Weihua, Yuan, Xin, Lim, Noha, et al. 2010. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. Arteriosclerosis, Thrombosis, and Vascular Biology, **30**(11), 2264–2276.

Wei, Ying, Kehm, Rebecca D, Goldberg, Mandy, & Terry, Mary Beth. 2019. Applications for Quantile regression in epidemiology. Current Epidemiology Reports, **6**(2), 191–199.

Weinberg, Wilhelm. 1908. ber den Nachweis der Vererbung beim Menschen. Jahres. Wiertt. Ver. Vaterl. Natkd., **64**, 369–382.

Welsh, Paul, Polisecki, Eliana, Robertson, Michele, Jahn, Sabine, Buckley, Brendan M, de Craen, Anton JM, Ford, Ian, Jukema, J Wouter, Macfarlane, Peter W, Packard, Chris J, et al. 2010. Unraveling the directional link between adiposity and inflammation: a bidirectional Mendelian randomization approach. The Journal of Clinical Endocrinology & Metabolism, **95**(1), 93–99.

Wheatley, Keith, & Gray, Richard. 2004. Commentary: Mendelian randomization—an update on its use to evaluate allogeneic stem cell transplantation in leukaemia. International Journal of Epidemiology, **33**(1), 15–17.

Windmeijer, Frank, Farbmacher, Helmut, Davies, Neil, & Davey Smith, George. 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.

Yao, Fang, Müller, Hans-Georg, & Wang, Jane-Ling. 2005. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association, **100**(470), 577–590.

Yavorska, Olena O, & Burgess, Stephen. 2017. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. International Journal of Epidemiology, **46**(6), 1734–1739.

Zhao, Qingyuan, Wang, Jingshu, Hemani, Gibran, Bowden, Jack, & Small, Dylan S. 2018. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. arXiv preprint arXiv:1801.09652.

Zhao, Qingyuan, Chen, Yang, Wang, Jingshu, & Small, Dylan 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.

Zhu, Zhihong, Zhang, Futao, Hu, Han, Bakshi, Andrew, Robinson, Matthew R, Powell, Joseph E, Montgomery, Grant W, Goddard, Michael E, Wray, Naomi R, Visscher, Peter M, et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics, **48**(5), 481–487.

Zuber, Verena, Colijn, Johanna Maria, Klaver, Caroline, & Burgess, Stephen. 2020. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. Nature Communications, **11**(1), 1–11.

Appendix A

Installation and application of mrbayes package in R

A.1 Links to install mrbayes R package

To install the `mrbayes` package please use the CRAN link below, and to install the development version of the package please see my GitHub repository linked below;

- CRAN link (https://cran.r-project.org/package=mrbayes)
- GitHub repository link (https://github.com/okezie94/mrbayes)

A.2 Results from more informative priors

From the example dataset, we applied an informative prior distribution on the parameter on the causal effect for MR-Egger model (3.8) as follows

$$\begin{aligned} \beta &\sim N\left(0, \phi\right) \\ \phi &\sim IG(0.5, 0.5) \\ \sigma &\sim IG(0.5, 0.5). \end{aligned} \tag{A.1}$$

The hierarchical prior distributions were used for the Radial MR-Egger model (3.9) on the parameter for the causal effect estimate (β) as follows,

$$\begin{aligned} \beta &\sim N\left(0, \frac{\tau}{\lambda}\right) \\ \lambda &\sim IG(0.5, 0.5) \\ \tau &= 0.025 \\ \sigma &\sim IG(0.5, 0.5). \end{aligned} \tag{A.2}$$

The posterior distribution was sampled using the rjags software. The estimates are shown in Table A.1.

Table A.1: Estimates from informative prior distributions

Model	Coefficient	Estimate	95% CrI
Bayesian MR-Egger	Intercept	0.0038	-0.0058, 0.0137
Bayesian MR-Egger	Slope	0.3271	0.0412, 0.6153
Bayesian MR-Egger Radial	Intercept	0.3178	-0.4154, 1.051
Bayesian MR-Egger Radial	Slope	0.3248	0.0592, 0.598

Estimates in table A.1 show shrinkage towards the null and the credible interval spans zero for the slope parameter. The estimates show the effects of different priors when using Bayesian models in MR with summary-level data.

In summary, it is helpful to compare estimates from models fitted with both uninformative and partially informative prior distributions.

Appendix B

Shrinkage profile and bounds for Horseshoe and Horseshoe+ priors

B.1 Shrinkage profile

The shrinkage effect of the Horseshoe prior can be described through the shrinkage profile ϕ_i which is the weight of the posterior mean of each exposure (β_i), and the posterior mean is derived as,

$$\begin{aligned} E\left(\beta_{ij}|\Gamma_j, \lambda_i^2\right) &= \frac{n\sigma^{-2}}{\tau^{-2}\lambda_j^{-2} + n\sigma^{-2}}\beta_{ij} \\ \left(1 - \frac{1}{1 + n\sigma^{-2}\tau^2\lambda_j^2}\right)\beta_{ij} &= (1-\phi_j)\beta_{ij} \\ \text{where } \phi_i &= \frac{1}{1 + n\sigma^2\lambda_i^2\tau^2}. \end{aligned} \tag{B.1}$$

B.2 Bounds for horseshoe prior

The Horseshoe density does not have an analytic form, however, tight bounds are available. Given fixed scale values $\sigma^2 = \tau^2 = 1$ the marginal density is derived as

$$p_{hs}(\beta) = \int_0^\infty \frac{1}{2\pi\lambda^2}^{\frac{1}{2}} \exp\left(\frac{-\beta^2}{2\lambda^2}\frac{2}{\pi(1+\lambda^2)}\right) d\lambda, \tag{B.2}$$

differentiating by parts

$$\begin{aligned} u &= \frac{1}{\lambda^2} \\ \frac{du}{d\lambda} &= -2\lambda^{-3} \\ d\lambda &= \frac{du}{-2\lambda^{-3}}, \end{aligned} \tag{B.3}$$

and including (B.3) in (B.2), gives

$$\frac{1}{2\pi^{\frac{1}{2}}} \int_0^\infty \frac{\lambda^2}{\pi(1+\lambda^2)} \exp\left(\frac{-\beta^2}{2\lambda^2}\right) du. \tag{B.4}$$

Recall from (B.3)

$$\frac{1}{2\pi^{\frac{1}{2}}} \int_0^\infty \frac{1}{1+u} \exp\left(\frac{-\beta^2 u}{2}\right) du \tag{B.5}$$

then differentiating by parts

$$\begin{aligned} v &= 1+u \\ \frac{dv}{du} &= 1 \\ du &= dv \end{aligned} \tag{B.6}$$

and subsituting into (B.5) we have

$$k \exp\left(\frac{\beta^2}{2}\right) \int_1^\infty \frac{1}{v} \exp\left(\frac{-v\beta^2}{2}\right) dv$$

where $k = \frac{1}{\sqrt{(2\pi^3)}}$. We can rewrite this in terms of the exponential integral i.e.,

$$E_1(v) = \int_1^\infty \frac{e^{-v}dv}{v}.$$

The reduced form is

$$ke^{\frac{\beta^2}{2}}E_1\frac{\beta^2}{2}. \tag{B.7}$$

From the exponential integral we set the bounds as;

$$\frac{K}{2}\log\left(1+\frac{4}{\beta^2}\right) < p_{hs}(\beta) < K\log\left(1+\frac{2}{\beta^2}\right).$$

B.3 Bounds for horseshoe+ estimator

The marginal density for the horseshoe+ is

$$p_{hs+}(\beta) = \int_0^\infty \frac{4}{\pi^2\sqrt{2\pi}\lambda^2}e^{\frac{-\beta^2}{2\lambda^2}}\frac{\log(\lambda)}{\lambda^2-1}d\lambda. \tag{B.8}$$

Substituting (B.3) into (B.8) gives

$$\frac{1}{\pi^2\sqrt{2\pi}}\int_0^\infty e^{\frac{-\beta^2 u}{2}}\frac{\log(u)}{u-1}du.$$

To setup the bounds (Bhadra et al. 2017) followed the strategy of $\frac{\log(u)}{U-1} \leq \frac{1}{\sqrt{u}}$ for $U > 0$ for the upper bound. Subsituting we have;

$$\int_0^\infty e^{\frac{-\beta^2 u}{2}} \frac{\log(u)}{u-1} du \leq \int_0^\infty \frac{-1}{\sqrt{U}} e^{\frac{-\beta^2 u}{2}} du$$
$$\frac{\Gamma(1/2)}{\sqrt{\beta^2/2}} = \frac{\sqrt{2\pi}}{|\beta|}.$$

For the lower bound $\frac{\log(u)}{U-1} \geq \frac{2}{1+U}$ for $U > 0$;

$$\int_0^\infty e^{\frac{-\beta^2 u}{2}} \frac{\log(u)}{u-1} du \geq \int_0^\infty \frac{2}{U+1} e^{\frac{-\beta^2 u}{2}} du$$
$$= 2e^{\frac{\beta^2}{2}} E_1\left(\frac{\beta^2}{2}\right).$$

Recall the upper limit of an exponential integral and combining both the bounds for the horseshoe+ estimator is

$$\frac{1}{\pi^2\sqrt{2\pi}} \log\left(1 + \frac{4}{\beta^2}\right) < p_{hs+}(\beta) \leq \frac{\sqrt{2\pi}}{|\beta|}. \tag{B.9}$$

We see from (B.9), the bounds are sharper than for the horseshoe estimator.

B.4 Data Application of the MVMR models

The abbreviations of the selected exposures are denoted in table B.1.

Table B.1: Summary of selected exposures

Abbreviation	Name
ApoA1	ApoA1
ApoB	ApoB
Est.C	Esterified cholesterol
HDL.C	Total cholesterol in HDL
HDL.D	HDL diameter
IDL.C	Total cholesterol in IDL
IDL.TG	Triglycerides in IDL
L.HDL.C	Total cholesterol in large HDL
L.VLDL.C	Total cholesterol in large VLDL
L.VLDL.TG	Triglycerides in large VLDL
LDL.C	Total cholesterol in LDL
LDL.D	LDL diameter
M.HDL.C	Total cholesterol in medium HDL
M.VLDL.C	Total cholesterol in medium VLDL
M.VLDL.TG	Triglycerides in medium VLDL
S.HDL.TG	Triglycerides in small HDL
S.LDL.C	Total cholesterol in small LDL
S.VLDL.C	Total cholesterol in small VLDL
S.VLDL.TG	Triglycerides in small VLDL
Serum.C	Serum total cholesterol
Serum.TG	Serum total triglycerides
SM	Sphingomyelins
Tot.FA	Total fatty acids
TotPG	Total phosphoglycerides
VLDL.D	VLDL diameter
XL.HDL.C	Total cholesterol in very large HDL
XL.HDL.TG	Triglycerides in very large HDL
XL.VLDL.TG	Triglycerides in very large VLDL
XS.VLDL.TG	Triglycerides in very small VLDL
XXL.VLDL.TG	Triglycerides in chylomicrons and extremely large VLDL

Appendix C

Additional simulation results for quantile models

C.1 Linear programme for quantile regression

The standard form for linear programs is

$$\min_{z} C^T z \tag{C.1}$$

subject to $Az = b, z \geq 0$.

All variables minimising z should be positive to arrive at a linear programme on a standard form. Therefore it is decomposed to a positive and negative part using slack variables that is $\varepsilon_j = u_j - v_j$, where;

$$u_j = \max(0, \varepsilon_j) = |\varepsilon_j|[\varepsilon_j \geq 0]$$

$$v_j = \max(0, -\varepsilon_j) = |\varepsilon_j|[\varepsilon_j < 0].$$

The sum of residuals assigned weights by the check function

$$\sum_{j=1}^{J} \rho_\tau(\varepsilon_j) = \sum_{j=1}^{J} \tau u_j + (1-\tau)v_j = \tau u + (1-\tau)v;$$

where $u = (u_1, \ldots, u_j)^T$ and $v = (v_1, \ldots, v_j)^T$.

The residuals must satisfy the J constraints that $\tilde{\Gamma}_j - \tilde{\gamma}_j^T \beta = \varepsilon_j = u_j - v_j$, this results in the formulation as a linear programme

$$\min_{\beta \in \mathbb{R}^J, u \in \mathbb{R}_+^J, v \in \mathbb{R}_+^J} \{\tau u + (1-\tau)v | \tilde{\Gamma}_j = \tilde{\gamma}_j \beta + u_j - v_j, \quad j = 1, \ldots, J\}. \tag{C.2}$$

However $\beta \in \mathbb{R}^J$ is still not restricted to be positive as required for the standard form. Hence $\beta = \beta^+ - \beta^-$ where again $\beta^+ = \max(0, \beta)$ and $\beta^- = \max(0, -\beta)$. The J constraints can be written as;

$$\Gamma := \begin{bmatrix} \Gamma_1 \\ \vdots \\ \Gamma_j \end{bmatrix} = \begin{bmatrix} \gamma_1^T \\ \vdots \\ \gamma_j^T \end{bmatrix} \left(\beta^+ - \beta^-\right).$$

b is defined as $b := \Gamma$ and the design matrix γ for storing independent variables as

$$\gamma := \begin{bmatrix} \gamma_1^T \\ \vdots \\ \gamma_j^T \end{bmatrix}.$$

The constraint can be rewritten as

$$\begin{aligned} b &= \gamma\left(\beta^+ - \beta^-\right) + u - v \\ &= [\gamma, -\gamma] \begin{bmatrix} \beta^+ \\ \beta^- \\ u \\ v \end{bmatrix}. \end{aligned} \tag{C.3}$$

The variables in (C.3) can be reduced to $A := [\gamma, -\gamma]$ and β^+andβ^- are minimised and they are part of z, making (C.3)

$$b = [\gamma, -\gamma] \begin{bmatrix} \beta^+ \\ \beta^- \\ u \\ v \end{bmatrix} = Az.$$

The variables β^+andβ^- only affect the minimization problem through the constraint 0, the coefficient vector c can be defined as;

$$c = \begin{bmatrix} \beta^+ \\ \beta^- \\ u \\ v \end{bmatrix}.$$

This ensures that all the variables c, A and b are specified

$$c^T z = 0^T \underbrace{\left(\beta^+ - \beta^-\right)}_{0} + \tau u + (1-\tau)v = \sum_{j=1}^{J} \rho_\tau(\varepsilon_j).$$

C.2 Simulation results from quantile estimates

The results of the simulation study of the quantile levels in Chapter 5 from the balanced and directed pleiotropic scenarios are discussed in this section. Under the null hypothesis, the median estimates are the most accurate for all proportions, 10%, 30%, and 50%, of invalid instruments, as shown in Figure C.1.

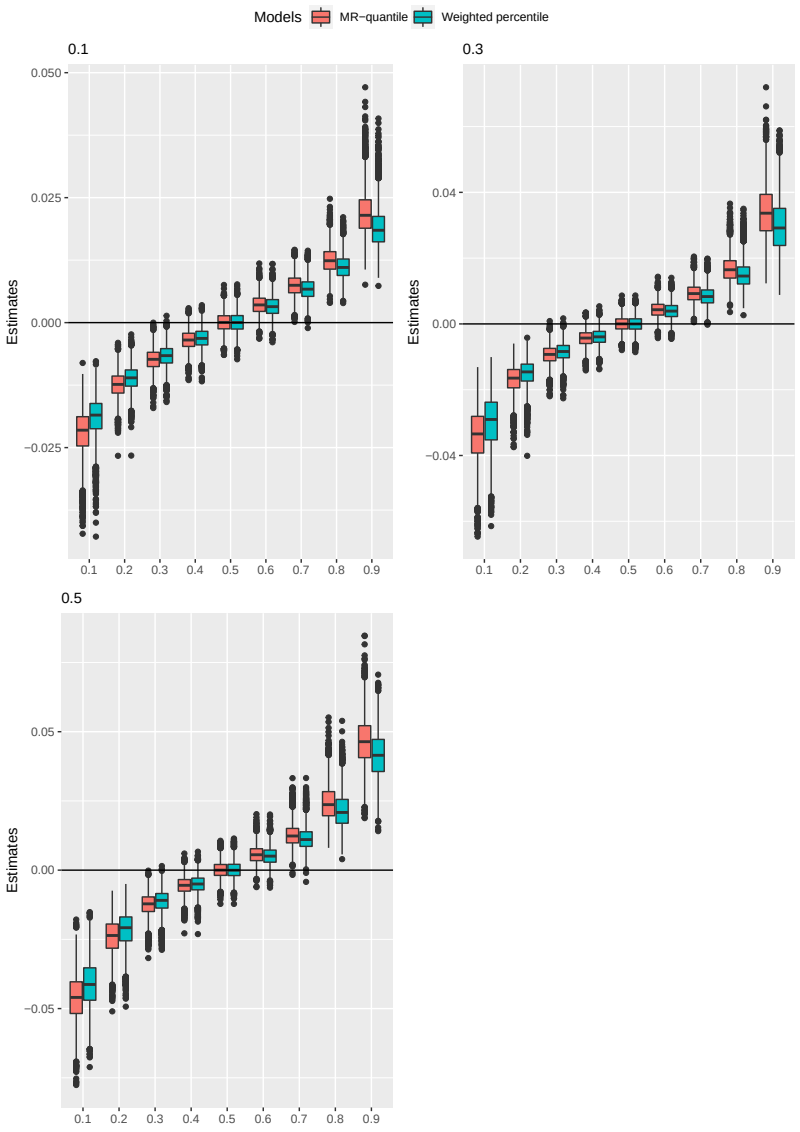

Figure C.1: Boxplot of the estimates within different quantiles in a balanced pleiotropy scenario assuming null effect. The title of each plot is according to proportion of invalid instruments

The simulation estimates from the MR-quantile imply better accuracy in 0.6 for the alternative hypothesis, however the quantile level of 0.7 shows better accuracy for the weighted percentile estimator based on figure C.2.

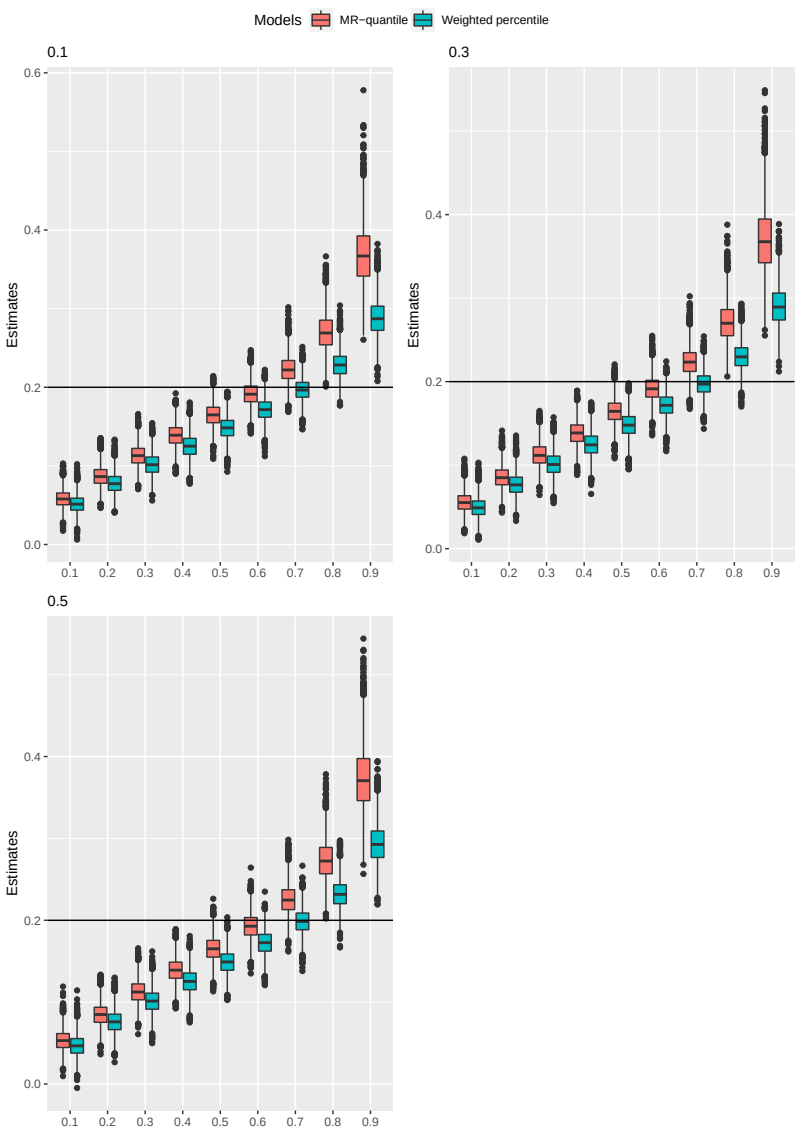

Figure C.2: Boxplot of the estimates within different quantiles in a balanced pleiotropy scenario assuming causal effect. The title of each plot is according to proportion of invalid instruments

The simulation studies depicted in figure C.3 show that the estimations from the weighted percentile and MR-quantile models are accurate within the lower quantiles when the null effect is assumed. For the alternative assumption depicted in figure C.4, estimates from the MR-quantile model show accuracy within the 0.6 quantile level, whereas estimates from the weighted percentile model show more accuracy at the 0.7

quantile level.

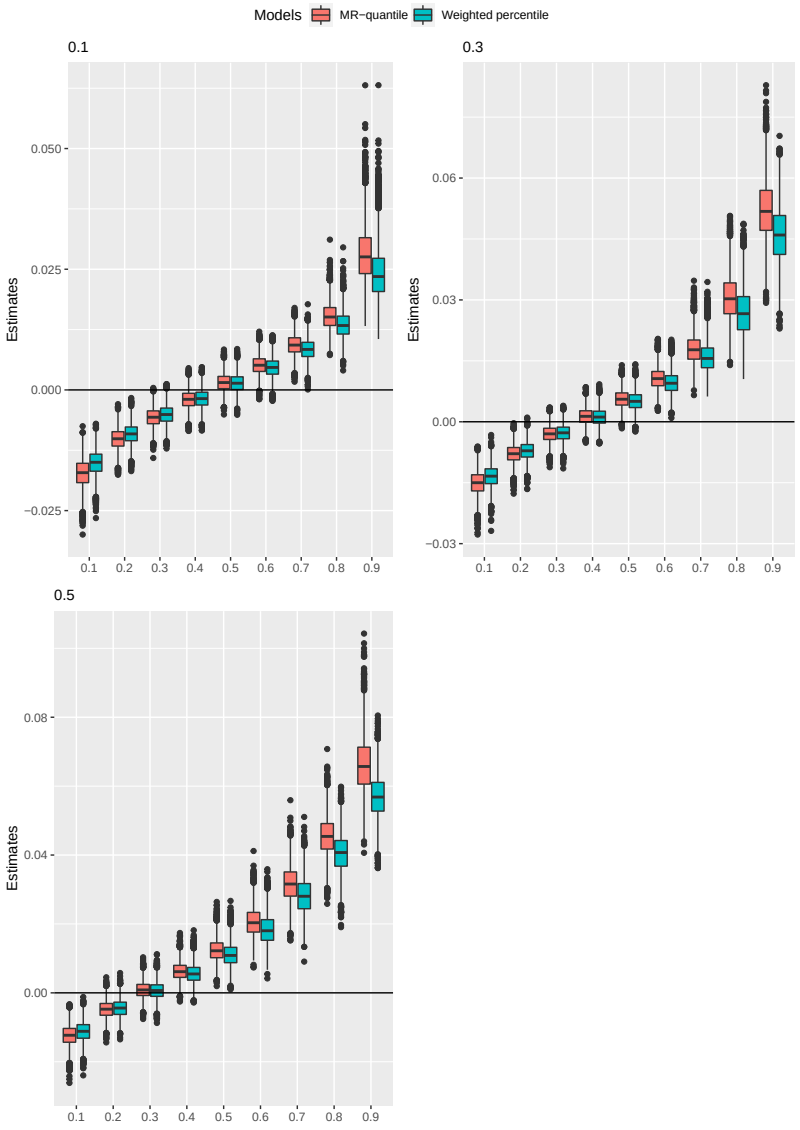

Figure C.3: Boxplot of the estimates within different quantiles in a balanced pleiotropy scenario assuming null effect. The title of each plot is according to proportion of invalid instruments

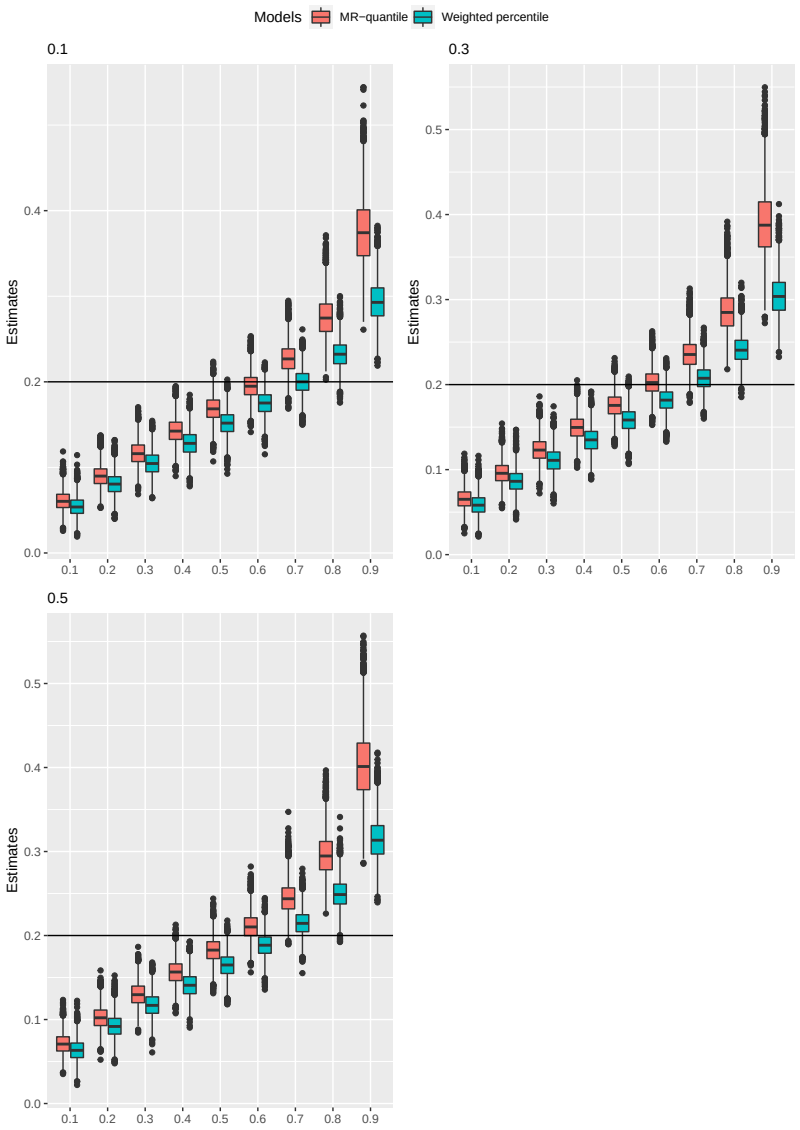

Figure C.4: Boxplot of the estimates within different quantiles in a balanced pleiotropy scenario assuming causal effect. The title of each plot is according to proportion of invalid instruments