Weak-Instrument and Pleiotropy-Robust Methods for Mendelian randomisation, with Applications to Mental Health

Submitted by Vasileios Karageorgiou, to the University of Exeter

as a thesis for the degree of Doctor of Philosophy in Medical Studies, July 2023.

I certify that all material in this thesis which is not my own work has been identified and that any material that has previously been submitted and approved for the award of a degree by this or any other University has been acknowledged.

January 14, 2024

Abstract

This PhD dissertation focused on developing and applying new methods for Mendelian Randomisation (MR), a technique that uses genetic variants as instrumental variables in order to assess causal effects of exposures on health outcomes. The major focus of the applied research is psychiatric research and mental health, with a range of analyses that address the topic of causal risk factors for depression with the use of these genetics-informed methods.

The first contribution of this dissertation is the development of new methods for pleiotropy-robust MR by leveraging sex specificity of phenotypes. These methods allow for more accurate and robust estimation of causal effects by cancelling out potential pleiotropic effects of genetic instruments. The second contribution is a new method for appraising high-dimensional correlated variables in multivariable MR. This method allows for the inclusion of multiple correlated variables as exposures in MR analyses, through a transformation to groups of exposures that have attractive statistical properties and biological meaning. Finally, the dissertation provides an applied analysis of how inflammation and BMI affect a range of depression phenotypes with cutting-edge methods. This analysis replicates previous results on the harmful effects of overweight on mood and challenges the independent effect of inflammation as proxied by CRP. The introduction of the dissertation is divided into two parts. The first part provides a walkthrough of the epidemiological concepts of bias, randomisation, and causal inference with observational data. The second part is a specific introduction to MR, including its underlying assumptions and limitations, as well as detailed discussion of developments that make it more robust. Overall, this dissertation contributes new methods and applied analyses to the field of MR, with potential implications for researchers and practitioners.

Contents

List of Tables

List of Figures

List of Publications

- **Karageorgiou V**, Gill D, Bowden J, & Zuber V. (2023). Sparse dimensionality reduction approaches in Mendelian randomisation with highly correlated exposures. In eLife (Vol. 12). eLife Sciences Publications, Ltd. doi.org/10.7554/elife.80063
- **Karageorgiou V**, Tyrrell J, Mckinley T J, & Bowden, J. (2023). Weak and pleiotropy robust sex-stratified Mendelian randomization in the one sample and two sample settings. In Genetic Epidemiology (Vol. 47, Issue 2, pp. 135–151). Wiley. doi.org/10.1002/gepi.22512
- **Karageorgiou V**, Casanova F, O'Loughlin J, Green H, McKinley T J, Bowden J, & Tyrrell J. (2023). Body mass index and inflammation in depression and treatment-resistant depression: a Mendelian randomisation study. In BMC Medicine (Vol. 21, Issue 1). Springer Science and Business Media LLC. doi.org/10.1186/s12916-023-03001-7

Author's Declaration

This dissertation is submitted to the University of Exeter in fulfillment of the requirements for the degree of Doctor of Philosophy.

In accordance with the regulations of the University of Exeter, I hereby grant permission for this dissertation to be made publicly available in whole or in part, with the understanding that the intellectual property rights associated with this research will be duly acknowledged, and that no quotation may be published without proper acknowledgement.

I declare that the research presented in this dissertation has been conducted in accordance with the ethical guidelines set forth by UK Biobank Ethics and Governance Council (EGC). Throughout the research process, strict adherence to confidentiality and data protection protocols was maintained. Any potential conflicts of interest that may have arisen during the course of this research have been disclosed and addressed accordingly.

The presented work is original and all sources of information used are appropriately referenced. Collaboration with other researchers and colleagues has been duly acknowledged, with the nature and extent of their contributions outlined as follows. In Chapters 1 and 2, I performed the original review and wrote the draft and my supervisors read, reviewed and edited the chapters. In Chapters 3-5, I conducted the

investigations and wrote the original drafts of the papers. While I am primarily responsible for the presented investigations, the open and collaborative culture at the University of Exeter has enabled these and I am thankful for the many contributions that shaped this work. Professor Jack Bowden conceptualised the initial model of Chapter 3. Dr Verena Zuber and Dr Dipender Gill were responsible for the original project proposal and its applied focus, which formed the basis of my work in Chapter 4. Professor Bowden helped me to turn this into a coherent and rigorous piece of research, in particular by advising on the design of an extensive simulation study. Dr Jess Tyrell led the supervision of the applied analysis on the genetics of mental health in Chapter 5, with Professor Bowden specifically advising on the use of collider correction and mediation analysis. Dr Trevelyan J McKinley provided statistical advice for Chapters 3 and 5 and reviewed and edited the dissertation. I also received software support and valuable scientific discussions from Dr Francesco Casanova, Dr Jessica O'Loughlin, Dr Ninon Mounier, Dr Xiaoran Liang, Dr Luke C. Pilling, Dr Tanimola Martins, Dr Harry Green, and Dr Robin Beaumont (particularly for Chapter 5).

I take responsibility for the content and integrity of the dissertation, and for adhering to the principles of academic honesty and rigour.

List of Abbreviations

2SLS: Two-stage least squares

AUC: Area Under the Curve

BMI: Body Mass Index

CFS: Conditional $F-$Statistic

CRP: C-reactive Protein

DAG: Directed Acyclic Graph

GWAS: Genome-wide Association Study

GxE: Gene-by-Environment Interaction

InSIDE: Instrument Strength Independent of Direct Effect

IV: Instrumental Variable

IVW: Inverse Variance-Weighted Meta-analysis

MDD: Major Depressive Disorder

MR: Mendelian randomisation

MVMR: Multivariable Mendelian randomisation

PCA: Principal Component Analysis

SNP: Single Nucleotide Polymorphism

T1E: Type I Error

TRD: Treatment-Resistant Depression

UKB: UK Biobank

Chapter 1

Introduction to Epidemiology and Evidence from Observational Data

In this PhD thesis, I explore new methods in Mendelian randomisation (MR), an approach in the analysis of observational data that uses genetic variants as instrumental variables and allows for causal inference. The focus is the development of new ways to analyse large-scale genetic data, and linked health exposures and outcomes, with the major focus of the applied analyses being the relationship of body weight, inflammatory status, and depression. In the field of psychiatric research, uncovering such causal links could inform treatment strategies for depression and other psychiatric disorders.

1.1 A Brief history of evidence-based medicine in Epidemiology

Epidemiology, as it is currently practiced, can be defined as the scientific study of the distribution and determinants of health-related states and events in populations, with the aim of identifying and understanding the causes as well as the consequences of

diseases, and developing effective strategies for their prevention and control [1]. This approach to health research is rooted in the methodological traditions of statistics, and reflects a particular worldview that values empirical observation, experimentation, and the use of statistical methods to analyse data. A range of practices resulting from this empiricist turn are codified in the wide term evidence-based medicine ('EBM'). This approach relies on the systematic collection and critical appraisal of evidence through data to inform clinical decision-making, rather than relying solely on clinical experience or intuition [2].

EBM represents a significant turning point in the history of medicine. It was driven by the need to counteract the adoption of questionable medical practices that had gained acceptance. A classic example is lobotomy and the field of 'psychosurgery', which involved a range of techniques based on severing connections between the prefrontal cortex and the rest of the brain to treat mental disorders. This procedure was first performed by Italian surgeon Fiamberti and later developed by American surgeon Walter Freeman II in Washington D.C., and was performed without adequate evidence to support its efficacy and often led to serious side effects and permanent damage [3]. What led to its abolition was in part the social movements against it and the gradual emergence of effective medications, the first of whom being chlorpromazine, an anaesthetic found to have sedating effects.

Likewise, in the 19th and early 20th centuries, people with mental illness were frequently institutionalised in large asylums, where they were subjected to neglect and abuse. However, this approach was later found to be ineffective and harmful, and

many of these institutions were eventually shut down.

1.2 Hierarchy of Evidence

A puzzling recognition is that the seemingly inhuman practices mentioned above were not completely devoid of any evidence. The key is that the level of evidence, which predominantly constituted personal experience seen through the lens of political and religious doctrine, was poor within today's EBM framework. Indeed, the interrelated fields of statistics and epidemiology arguably grew out of a need to address the inadequacies of this traditional approach [4].

EBM emphasises a reproducible and systematic testing of medical practices, rather than relying on the personality traits, eminence or political influence of individual medical practitioners or institutions where they are practiced. This approach helps ensure that medical practices are based on reliable evidence and are consistently effective and safe, rather than being based on outdated or dubious practices. Among the popular tools of EBM is a hierarchy of evidence, visualised in Figure 1.1, which solidifies the notation that not all evidence is equal. It has become familiar to healthcare practitioners when appraising, applying, or teaching medicine [5]. At the bottom of the pyramid, we find case series, which consist of descriptions of individual disease cases from a single or small number of study centres. If the group of cases share a common characteristic that is viewed to be rare in the general population, then this provides some evidence that the characteristic could play a role in the disease. They are limited by lack of comparison groups and selection bias. A notorious example is the publication of a case series in the Lancet suggesting a link between

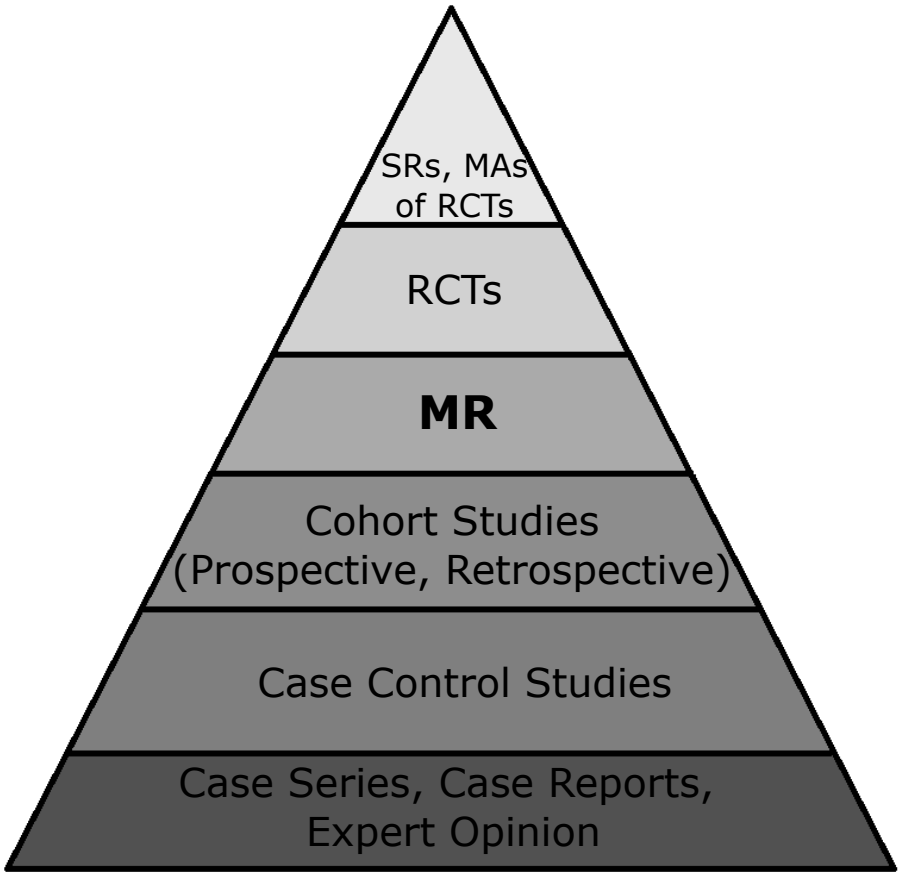


Figure 1.1: Hierarchy of evidence. We follow Davies et al. [6] in the introduction of MR studies in this common visualisation tool [5]. SRs: systematic reviews; MAs: meta-analyses; RCT: randomised controlled trial; MR: Mendelian Randomisation.

the MMR vaccine and autism [7]. This in turn caused a reluctance in vaccination. Subsequent epidemiological studies did not provide evidence for this, and the case series was retracted, with the Lancet admitting that the article contained false and fabricated elements.

Moving up the pyramid, we find case-control studies, which compare individuals with a particular condition (cases) to those without it (controls). They are seen as an improvement over the first level of the pyramid as, by adding a control group that is closely matched as possible to the cases in terms of measured characteristics (such as age, sex, and socioeconomic status), they provide an ideally close approximation

of what the outcome would have been if the individuals were not exposed (counterfactual). A famous example is a preliminary case-control study of smoking and lung cancer by Sir Richard Doll and Sir Austin Bradford Hill that reported a strong association between smoking status and lung cancer [8]. The authors recognised the limitations of this case-control design and they led research efforts that became the British Doctors Study, a meticulous longitudinal study that convincingly demonstrated that smoking precedes lung cancer development [9] (also see next paragraph on cohort studies). Case-control studies can suffer from bias due to the inherent difficulties in obtaining matched groups, and causality cannot be established. If the controls are not matched from the same or a reasonably similar reference population, then the comparison group may differ systematically from the cases in ways that affect the association [1]. However, for many research questions such as the link of smoking with lung cancer, well-performed case-control studies are a valuable tool.

Cohort studies, which follow a group of individuals over time, are considered a step up in quality from case-control studies. They follow individuals over a sustained period, either forward in time following initial recruitment, or retrospectively from a fixed point in the past to the present day. One representative example of a modern prospective cohort study is the UK Biobank (UKB) [10]. UKB was launched in 2006 and follows over 500,000 individuals over time, collecting extensive information on a wide range of biological and environmental factors. At baseline, participants completed a comprehensive questionnaire that inquired about their lifestyle, personal history, and medical conditions. This information, together with biological samples that are being added iteratively (e.g. increasing levels of genetic data are being added since 2015 [11]), has enabled researchers to investigate a broad range of

diseases and health outcomes. One major strength of cohort studies like UKB is the ability to query multiple risk factors in very well-powered samples. However, they are not immune to bias. A key criticism of UKB is that the study is not fully representative of the UK general population, by virtue of its participants being slightly older, wealthier, thinner, less likely to smoke or drink alcohol, and having had fewer diagnosed health conditions at baseline [12]. This makes disease associations uncovered in UKB hard to generalise to people of non-European ancestries and socioeconomic status. Another bias affecting many UKB analyses is its use of self-reported data on diet and lifestyle factors such as as smoking [13] or physical activity [14]. For these two examples, the correlation between self-reported measures and more objective measures such as cotinine levels or accelerometer readings is often modest, which inevitably leads to bias.

Close to the top of evidence pyramid are randomised controlled trials (RCTs), which are the predominant example of an experimental study, and offer numerous advantages for minimizing bias [15]. The canonical RCT design involves assigning participants at random to receive either a putative active treatment for a medical condition, or a control treatment. Where possible, the allocation is blinded to the study coordinator (e.g. a doctor) and the participant. Randomisation ensures that other factors that may impact patient prognosis are balanced between the treatment and control groups. This means that any differences in the outcomes of patients at the end of the study can be confidently attributed to the treatment itself. The same certainty cannot be automatically extended to case control and cohort studies, where such prognostic factors would have to be directly identified and accurately measured,

before being appropriately adjusted for in a complex statistical model. The concepts of confounding and causation will be covered in detail in the remaining parts of this chapter.

Meta-analyses of RCTs are considered the most robust type of evidence [16], as they combine results from multiple trials that attempt to answer the same scientific question. This generally leads to a large increase in the precision of the overall estimate. Furthermore, by virtue of summarising evidence from multiple study teams with subtle but unavoidable variations in their study population or trial implementation, their findings are more generalisable and representative than those of a single study.

Despite sitting at the top of the evidence hierarchy, meta-analyses are not immune to bias caused by selective reporting of results by study authors or selective publication of 'significant' findings by journal editors. Fortunately, by virtue of the fact that meta-analyses contain multiple independent study results, these biases can be detected through graphical methods, such as Egger regression [17].

1.2.1 Summary remarks

While the evidence pyramid was a turning point for summarising evidence and for improving research conduct, it can be oversimplifying if used as a sole marker of quality. For example, the pyramid places RCTs at the top of the hierarchy. However, their quality varies greatly depending on factors such as allocation concealment (ensuring participants are randomly allocated to groups without influence from researchers),

blinding (keeping participants and researchers unaware of which group they are in), and attrition bias (loss of participants during the study due to withdrawal or loss to follow-up) [18]. In addition, the pyramid does not always account for nuances in research design, such as sample size and the analytical choices of the study analysts.

Modern cohort studies such as UKB have very extensive baseline characterisation of their target populations, linkage with many other data sources, and much larger sample sizes than RCTs. Due to ethical and logistical reasons, sometimes the best available evidence is a non-randomised study, such as in the case of smoking and lung cancer where there have been no randomised experiments however the evidence has been convincing and influential. Therefore, it is important to consider the limitations of the pyramid and to approach evidence evaluation with a critical eye, taking into account the nature of the research question, pragmatic limitations and external available evidence.

1.3 Why are observational studies unreliable for learning about causality?

We have so far touched on several types of bias that can affect the validity of interpreting observational study findings that link a risk factor or exposure to a health outcome in a causal manner, when there the exposure has not been fixed by design. One of the most significant sources of bias is *confounding*, where a potentially unmeasured quantity exists that influences both the exposure and the outcome. A classic example is the association of coffee consumption and various health outcomes. Confounding arises since coffee drinkers are also more likely to smoke, and smoking

is an established risk factor for cardiovascular, metabolic and neoplastic disease. In a recent wide-scoping meta-analysis, adequate adjustment for smoking in primary studies largely nullified previously identified harmful associations of coffee drinking [19]. *Reverse causation* is another important concern if the 'outcome' could also be a cause of the exposure. For example, in studies assessing the association between depression and the risk of chronic diseases, depression may be both a cause as well as a consequence of chronic diseases such as stroke [20], arthritis [21, 22], and chronic obstructive pulmonary disease [23]. The burden on health and well-being is so substantial that depression can and does often ensue, but the inverse is also plausible. *Information bias* may occur when there are inaccuracies in measuring either the exposure or the outcome. This can happen for several reasons, such as the use of self-reported measures, which can lead to inaccuracies if participants do not report their symptoms truthfully or if the questions do not accurately quantify the true severity of depression.

1.4 Using Directed Acyclic Graphs (DAGs) to understand bias

In light of the increasing importance of using observational data to answer causal research questions, it is crucial to understand its limitations from a theoretical perspective, so that methods can be developed to overcome them. Perhaps the single most useful methodological framework to address this are directed acyclic graphs (DAGs) [24], which provide a graphical representation of the assumptions about the relationships between a set of variables.

In a DAG, variables are represented as nodes, and their connections are illustrated

by lines connecting two nodes known as edges/arcs in graph theory. The edges point from one node to another, that is they are 'directed' to represent the flow of causality from one variable to another. The absence of an arrow between two variables denotes no causal relationship between them.

For example, Figure 1.2 (a) represents the scenario where there is no direct causal effect between an exposure X and an outcome Y, but the variable U (a confounder) is a common cause of both. Such a variable could lead to X and Y being statistically significantly associated, despite there being no causal effect. Figure 1.2 (b) represents the related scenario where, in addition to U confounding the X-Y association, X does exert a causal effect on Y.

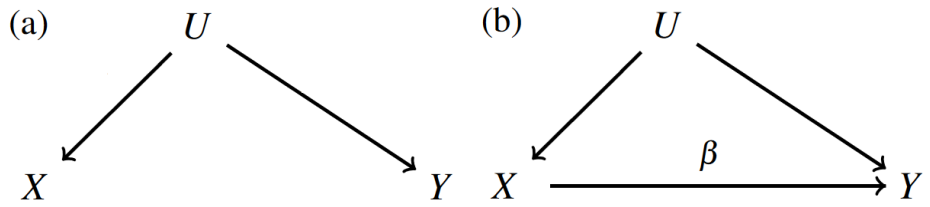


Figure 1.2: DAG representation of the relationship between risk factor X, disease Y in the presence of confounder U a) No causal effect of X on Y, b) a causal effect of X on Y

Figure 1.2b is consistent with the following linear model for Y given X and U for individuals i=1,...,n.

$$Y_i = \beta_i X_i + U_i. \tag{1.1}$$

The causal effect of X on Y is denoted by β_i, but what does this mean? One way to motivate this quantity is via the potential outcomes framework. Let $Y_i(X_i = x_i)$ be the potential outcome when X_i takes the observed value x_i. Let $Y_i(X_i = x_i - 1)$ be the potential outcome when X_i, contrary to fact, takes the value $x_i - 1$. That is, their exposure has been shifted down by one unit. The individual causal effect (ICE) for

person i can be defined as the difference between these two potential outcomes:

$$\begin{aligned} \text{ICE} &= Y(X_i = x_i) - Y(X_i = x_i - 1) \\ &= \beta_i X_i + U_i - (\beta_i(X_i - 1) + U_i) \\ &= \beta_i. \end{aligned} \tag{1.2}$$

The average causal effect (ACE) can be defined as the expected value of the same potential outcome contrast, taken over the entire population:

$$\begin{aligned} ACE &= E[Y(X_i = x_i) - Y(X_i = x_i - 1)] \\ &= \beta. \end{aligned} \tag{1.3}$$

In practice, we cannot estimate the ICE, only the ACE. Under the assumption of homogeneity (β_i = β), the two quantities are the same.

1.4.1 How does confounding bias causal estimates?

Consider an extended version of model 1.1:

$$X_i = \gamma_X U_i + \varepsilon_{Xi} \tag{1.4}$$

$$Y_i = \beta X_i + \gamma_Y U_i + \varepsilon_{Yi}, \tag{1.5}$$

where ε_{Xi} and ε_{Yi} are independent error terms and the parameters γ_X and γ_Y together govern the strength of confounding. Using a well known result, regressing Y on X will yield an observational association β_{obs} that will take the following value:

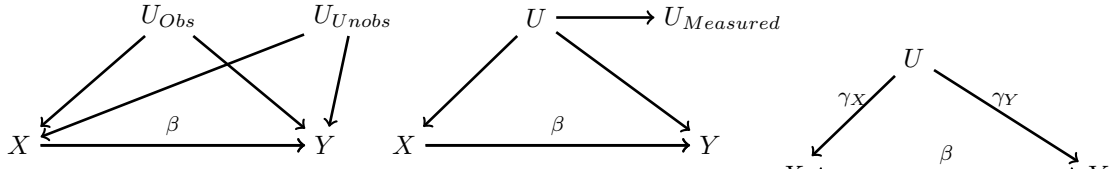
$$\begin{aligned}\beta_{\text{obs}} &= \frac{\text{cov}(X,Y)}{\text{var}(X)} \\ &= \frac{\beta\text{var}(X) + \gamma_Y\text{cov}(X,U)}{\text{var}(X)} \\ &= \beta + \frac{\gamma_X\gamma_Y\sigma_U^2}{\sigma_X^2},\end{aligned}$$

where σ_U^2 and σ_X^2 represent the variance of U and X respectively. We see that the obtained association is therefore not a reliable estimate of the causal effect β in the presence of unmeasured confounding. Specifically, a non-zero $\frac{\gamma_X\gamma_Y\sigma_Y^2}{\sigma_X^2}$ term could easily lead to a statistically significant observational association even if the causal effect β is zero. Clearly this bias increases as the strength of confounding increases.

One way to remove with bias would be to appropriately measure and adjust for the confounder U. For example, we may conduct an exhaustive search for potential confounders and adjust for them all in a complex statistical model. However, even if a number of important confounders have been found, we can never be certain that some have been missed. This is visualised in Figure 1.3a. Alternatively, even if all confounders have been found, it is possible that what is used in the analysis is an inaccurate proxy for U as shown in Figure 1.3b. This is especially pertinent for self-reported measures, such as questionnaires. In both cases, adjusting for U_{Obs} or $U_{Measured}$ only will not be sufficient to completely remove confounding bias, and the observational estimate would therefore be contaminated to allow a causal interpretation. This confounding structure is the simplest form of what is referred to as an open backdoor path in causal inference and DAGs. The term 'open' refers to it not being adjusted in the analysis, and 'backdoor' describes the existence of a path from

X to Y through U.

Confounders are a common issue in observational epidemiological studies, and while adjusting for known confounders helps reduce bias, it is difficult to explicitly model all potential confounders. This leads to inaccuracies in the estimation of a causal effect of an exposure on an outcome, even if all measured confounders are accounted for. In the next chapter, we will explore instrumental variables as a way to obtain estimates that are naturally robust to such biases. This is particularly important as accumulating evidence on multiple exposures affecting outcomes becomes available and it is not always possible to accurately specify all potential confounders.



(a) Observed and Unobserved Confounders

(b) Measurement Error in Confounder

(c) Reverse Causality

Figure 1.3: Common issues in interpreting observational associations: Unobserved confounding, measurement error and reverse causality

A third possibility exists to frustrate the ability of an observational association to reflect a true causal effect, namely reverse causation, as represented by the DAG in Figure 1.3c. In this case, even if all confounders could be appropriately measured and adjusted for, the observational association from a regression of Y on X would not reflect β. To see this we assume the following model in place of (1.5) linking X to Y, where X is now the dependent variable

$$X_i = \beta Y + \gamma_X U + \varepsilon_{Xi}, \tag{1.6}$$

Incorrectly regressing Y on X would be equivalent to fitting the model

$$Y_i = \frac{X - \gamma_X U - \varepsilon_{Xi}}{\beta}, \tag{1.7}$$

so that β_{obs} would incorrectly estimate $1/\beta$ instead. Therefore, in the absence of any certainty as to the temporal order of X and Y, as is the case with a sizeable majority of exposure-outcome pairs collected in observational data, reverse causality cannot be ruled out.

1.5 Instrumental Variables: A solution to confounding and reverse causation

A major focus of quantitative economics is determining the causal effect of individual decisions or governmental policies on economic output. However, proper randomised experiments are rarely performed to guide this and so there is a strong reliance on observational data, which opens up the possibility of confounding. For instance, in studies of enrolment in voluntary training for furthering job opportunities, those that undertake this 'treatment' may be actively seeking higher earnings, hence their enrolment. This additional factor can contribute to behaviours other than job training that are financially beneficial for them. Regressing their subsequent earnings on treatment status would therefore be agnostic to this latter fact. As a result, interpreting this estimate as an unbiased causal effect of voluntary training on earnings could be misleading.

As outlined in Section 1.1, unobserved variables that affect an exposure and an

outcome can bias the estimate of the causal effect. This can be understood from a statistical perspective as being a consequence of the fact that the unmeasured 'error' terms in equations (1.4) and (1.5) are *correlated* due to the presence of the confounder variable U in both. This is sometimes referred to in economics as an errors-in-variables problem. A straightforward approach would be to use multiple regression, that is to estimate $E(Y|X, \mathbf{U})$, in order to remove the contribution of U. Even if U was measured, it is not necessary that only U is a confounder (unmeasured confounders) or that U is measured accurately (see Chapter 1.1). It is also likely in practice that missing data on U would reduce the effective sample size.

The nature of this problem motivated the use of instrumental variables (IV). The early contributors to the theory and applications of the methods are the economists Philip G. Wright, Ragnar Frisch and Olav Reiersol [25]. The approach involves finding a variable that is strongly predictive of a treatment variable of interest but independent of the error term in the outcome equation. This way, the variance explained by the IV in the treatment can be used in place of the confounder-affected treatment status and an estimate of the causal effect that is naturally robust to errors-in-variables can be obtained. Reiersol used this approach to investigate the relationship between education and income, where education is the treatment variable and income is the outcome variable. Reiersol recognised that education is likely to be *endogenous*, meaning that there are unobserved confounders that affect both education and income. For example, individuals with particular personality traits such as conscientiousness may be more likely to obtain higher levels of education and higher incomes [26]. This can lead to biased estimates of the causal effect of education on income.

To address this problem, Reiersol used the distance between an individual's home and the nearest school as an IV for education.

1.5.1 A formal definition of an IV

An instrumental variable is an 'exogenous' quantity, Z, that is strongly associated with a treatment or exposure X. The crucial distinction is that Z should temporally precede X. It then targets that specific part of the variation of the endogenous X that is distinct from U. In the example of the first paragraph, if the participants in the job training program live near the program centre, then they may be more likely to enrol independently of other reasons. Three core conditions have to be met in order for the estimates to be valid. These are given below and represented using an extended DAG in Figure 1.4:

1. Z is strongly associated with X (*relevance assumption*).
2. Z affects Y only through X (*exclusion restriction*).
3. Z does not share causes with X or Y (*exchangeability assumption*).

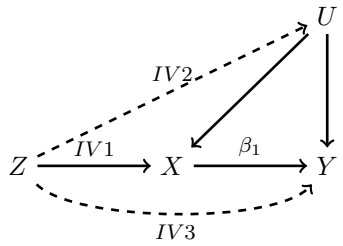


Figure 1.4: The IV core conditions.

If the three IV core conditions are satisfied, then Z should be independent of Y in the absence of a causal effect from X to Y. This means the Z-Y association can be used as a valid test for causality. In order to estimate the average causal effect of

assigning all participants to receive a college education ($X = 1$) versus no college education ($X = 0$):

$$\text{ACE} = E[Y(X_i = 1) - Y(X_i = 0)], \tag{1.8}$$

we required an additional condition of *Homogeneity*, that the causal effect is independent of Z [27]. The IV estimate for the ACE is then

$$\hat{\beta} \;=\; \frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,X)} = \frac{\frac{1}{n-1}\sum_{j=1}^{N}(z_j - \bar{z})(y_j - \bar{y})}{\frac{1}{n-1}\sum_{j=1}^{N}(z_j - \bar{z})(x_j - \bar{x})}. \tag{1.9}$$

The link between IV analyses and randomised experiments is clear if we consider random assignment (R) as the ultimate IV. In an idealised trial, R perfectly predicts which treatment an individual receives. It is independent of any confounders that could in theory influence whether (outside the confines of the trial) an individual takes a treatment or not, and can only affect the outcome through the treatment. It is for this reason that an analysis of trial data by randomised group, which is also known as an intention to treat (ITT) analysis, is a valid test for the causal effect of the treatment on the outcome. Furthermore, this test remains valid even if there is a degree of non-compliance, meaning that some patients in fact do not take the treatment they were randomised to receive. [28].

In econometrics, researchers typically use one or a small number of variables as instruments, which are chosen carefully to meet the IV core conditions. Examples of commonly used instruments include distance/ location [29], parental education level [30], and weather patterns [31].

The IV approach has been used in medical research. An example is in studies comparing the effectiveness of different drugs for a particular condition. Some clini-

cians may be more likely to prescribe one drug over others not just as a universally accepted and only therapeutic option, as many alternatives exist in many stages of diagnosis or treatment, but due to their personal preference or experience [32]. The crucial connection with IVs as discussed above is that this preference is *external* to the patient's condition and outcome, as well as external to any confounders of their association. Hence, clinicians' preference has been used as an IV.

In recent years, there has been increasing interest in the use of genetic variants as IVs in observational studies. The development of genome-wide association studies (GWAS) has led to the identification of thousands of genetic variants that are robustly associated with various traits and outcomes. In the next chapter, we will discuss in more detail the use of genetic variants as instruments in Mendelian Randomisation analyses and the challenges and limitations associated with this approach.

Chapter 2

Exploiting Genes as Instrumental Variables

2.1 Exploiting Genetic Variation for Causal Inference

2.1.1 Genes as the Basis of Inheritance

In 1865, Mendel published a report that summarised a series of experiments he performed on garden peas . He studied the pattern of inheritance of certain ancestral characteristics (e.g. colour, shape) to the progeny and proposed two that explained his experimental findings. He observed that differences in characteristics of the parents, specifically green and yellow colour hybridisation, led to a progeny with colours with a predictable distribution [33]. One of the colours of the parents was more frequently observed in the progeny (75%) and the other less so. The underlying factors that controlled this were then named dominant and recessive alleles.

The investigation of how multiple characteristics are passed on led to the formulation of the law of independent assortment (LIA). It was assumed that, if two genes

influence different characteristics, then they will follow an independent way of transmission. In other words, information for the heredity of one phenotype does not predict the presence of another phenotype.

With later works on cytology and microscopic observation of cell events that lead to the production of gametes (e.g. sperm cells, ova), the argument that chromosomes were the bearer of heredity was formulated and started getting adopted [34]. The genetic factors that Mendel described were physically anchored to chromosomes [35]. During the production of mature haploid cells (secondary oocyte, secondary spermatocyte), each chromosome is the result of a crossover between the two chromosomes in the premature diploid cell (primary oocyte, primary spermatocyte). With this knowledge of the chromosomes as the unit of heredity, it was reasonable to assume that LIA holds for genes that are positioned far apart in a chromosome and chiasmatic events would not affect their independent assortment. However, genes that are in close proximity and on the same chromosome are less likely to separate during this crossover. They are therefore more likely to be passed on jointly. For example, Bateson (1904) studied the colour in crossbred sweet peas and found that specific combinations of characteristics were more likely to be observed. This statistical dependence has come to be known as *linkage* [36].

From this there is a natural connection with RCTs as described in Chapter 1.2. Just as randomisation guarantees that the received treatment is a direct result of a random process (e.g. pseudorandom number generator), the independent assortment of sufficiently distant genes can be well described as a random process since each parent's genotype is carried forward in the offspring randomly [37]. This is commonly visualised as in Figure 2.1, where the common processes in both settings are ap-

parent [38]. As an intuitive example, we consider the case of *ALDH2* and alcohol consumption [39]. ALDH2 is an enzyme involved in the breakdown of alcohol. It catalyses the oxidation of acetaldehyde to acetic acid, a product that can be more readily eliminated from the urine. The gene that codes for ALDH2 is located on chromosome 12. There are two common isoforms of the gene: ALDH2*1 (wild-type) and ALDH2*2. ALDH2*2 contains a single nucleotide polymorphism (SNP) in position 487 that codes for a glutamate instead of lysine, with the functional results being a drastic reduction of enzymatic activity. As a result, acetaldehyde accumulates more easily when alcohol is consumed and a host of symptoms ensue (e.g. vasodilation leading to flushing and a drop in blood pressure, histamine release causing nausea and vomiting). As the parental alleles for ALDH2 are *randomly allocated* to offspring when gametes fertilise, an individual's genotype for ALDH2 is determined at random and that proportion of alcohol metabolism that is attributable to ALDH2 enzymatic activity is also determined at random. The practical consequence of this is that individuals with the ALDH2*2 version of the gene consume far less alcohol on average compared to ALDH2*1 carriers. We can therefore view the population as being randomised into a RCT in which they receive a lower or higher alcohol dose depending on their genotype. From this realisation, only a small conceptual leap is required in order to view the ALDH2 gene as an Instrumental Variable.

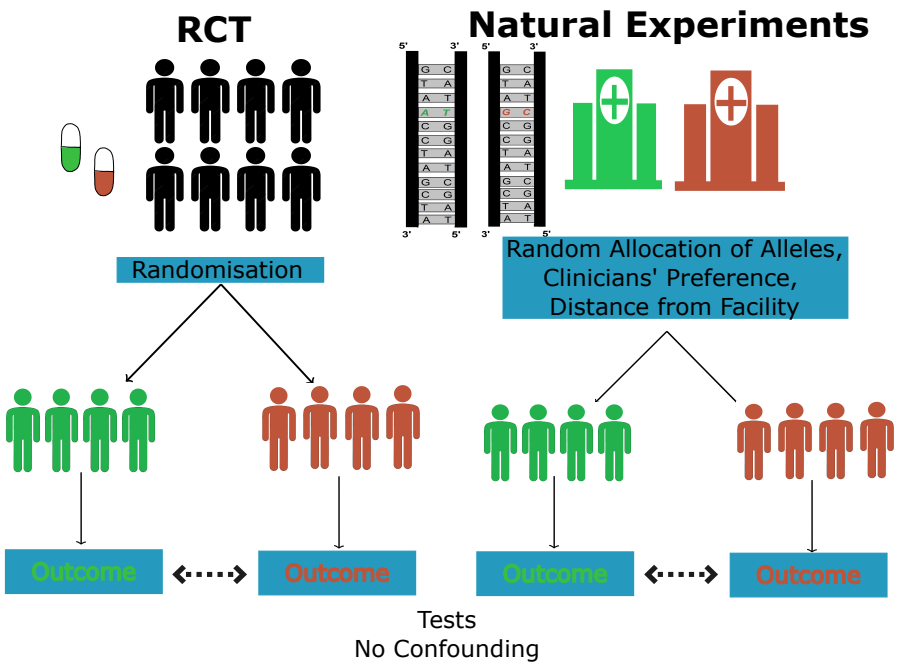


Figure 2.1: The concept of natural experiments can be likened to that of randomised controlled trials (RCTs), which are enabled by instrumental variable (IV) analyses. Various IVs have been used in health research, such as genetic variants that are randomly distributed, proximity to medical facilities, central differences in practices and policies, or preferences of clinicians. In both RCTs and natural experiments facilitated by IV analyses, the goal is to create intervention and control groups with balanced covariates [38].

2.2 Performing an MR analysis with individual level data

The first conceptualisation of the MR approach in epidemiology was motivated by discrepant findings in the role of cholesterol in cancer [40]. In a short letter to the editor of the Lancet in 1986 [41], Katan summarised the findings from the Seven Countries cohort study that looked into cancer rates in countries with different nutrition cultures and, as a result, different serum cholesterol population levels. Although the pooled estimate did not suggest a strong association, within-country estimates showed higher rates of cancer in those with lower cholesterol levels. Katan knew that a determinant of serum cholesterol apolipoprotein E (APOE) has three common isoforms: ApoE2, ApoE3, and ApoE4. Each of these isoforms is coded by a different

gene and, depending on the genetic profile, an individual can carry all possible pairwise combinations of them. Katan reasoned that, as long as different isoforms of the *ApoE* protein predicted varying serum cholesterol levels, the carriers of the genotype that is linked with lower cholesterol are expected to have lower odds of cancer if there is a true causal effect. He also observed that this approach could capture an inherent, lifelong predisposition to lower levels of cholesterol, rather than a spurious association due to cancer affecting cholesterol or due to the action of other confounding variables. One early application of the method published in 1991 involved the within-sibling comparisons of outcomes in acute myeloid leukemia [42]. The rediscovery of the method by Davey Smith and Ebrahim [37] and its generalisation to other epidemiological questions contributed to its popularisation.

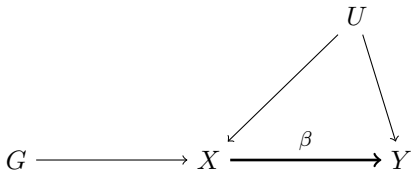


Figure 2.2: A directed acyclic graph (DAG) for Mendelian randomisation. The genetic variant G causes the exposure X, X causes the outcome Y, and the confounder U affects both X and Y.

Figure 2.2 formally defines a genetic variant G as an Instrumental Variable for use in estimating the causal effect (β) of modifying exposure X by a single unit on a health outcome Y. Assuming data for N individuals on G, X and Y, a practical way to obtain an estimate for the causal effect ($\hat{\beta}$) is by using two-stage least squares regression (TSLS) [43]. This simple method involves a first stage of regressing X on G in order to obtain the genetically predicted value of X; as can be seen in Fig. 2.2, these are not correlated with any confounders of X and Y (U), thus removing the contribution of external factors in the X-Y relationship. The first-stage model and the extraction of the genetically predicted exposure ($\hat{X}$) can be expressed as:

$$X = \gamma_0 + \gamma_G G + \varepsilon_X \tag{2.1}$$

$$\hat{X} = \hat{\gamma_0} + \hat{\gamma_G} G. \tag{2.2}$$

IV estimates are also consistent for a binary outcome [44]. In the first stage, the IV is regressed on the exposure, and the predicted values ($\hat{X}$) are then used in place of the exposure in the second stage. The only difference is that a logistic regression model [45] is typically used to estimate the effect of the exposure on the binary outcome. As the estimated parameters will be in the logit scale , careful interpretation of the effects is necessary. To improve clarity, we can calculate the average marginal effect [46]. This represents the change in the probability of the binary outcome for a unit increase in the exposure, holding all other variables constant and can be interpreted as a risk difference.

2.3 Mendelian randomisation with summary data

2.3.1 What is a genome-wide association study?

The MR approach was initially proposed for models that rely on individual-level data. To obtain the necessary data, we would need information on allele dosage for all k SNPs across all N participants, along with exposure X and outcome Y data for these participants, as well as a set of default covariates (such as genetic principal components that account for ancestry, and genotype chip that causes differences in measurement) and problem-specific covariates (such as age, sex, and area of residence). By using this data, we can estimate the causal effect of X on Y, as shown in

Eq. 1.9. However, individual level data is not always available and alternative methods must be used.

With the advent of more affordable technologies and large-scale cross-institution collaborations, a standard practice of identifying associations of genetic variants with phenotypes was developed, the genome-wide association study (GWAS). An early GWAS study of myocardial infarction identified candidate genes that are involved in the inflammatory cascade [47]. Despite the many methodological efforts, the core of the approach remains that of an association analysis of the genotype status with the corresponding phenotype, sequentially for each variant. Therefore, given the availability of genetic data for a population with a known phenotype X, each allele is regressed on the phenotype one at a time as shown in Equation 2.3. In the above example of a binary phenotype of case or control status for myocardial infarction, a generalised linear model can be used, where the log-odds (logit) of the event are modeled, as shown in Equation 2.4.

$$X_c = \gamma_0 + \gamma_{C,G_i} G_i \tag{2.3}$$

$$\text{logit}(Pr(X_b = 1)) = \gamma_0 + \gamma_{B,G_i} G_i. \tag{2.4}$$

A common tool to visualise such statistically significant associations arising from these models is a Manhattan plot (Figure 2.3). In this plot, p-values are displayed for each tested genetic variant against their physical location on the chromosomes, commonly by increasing position. The position of each genetic variant is estimated by comparing its observed physical location on the reference genome, a complete, high-quality, and well-annotated representation of human genome. The vertical axis

of the plot shows the negative log10 of the p-value, which succinctly summarises the strength of the association in one number; a horizontal line to visualise the genome-wide significance threshold (5×10^{-8}) is commonly included. This visualisation quickly identifies common variants that are most strongly associated with the phenotype. This particular threshold is arrived at by bearing in mind the concept of type I error inflation. Each variant-exposure association represents a test, and performing multiple tests without some degree of correction is bound to lead to false positive results by chance [1]. Therefore, a standard practice is to divide the nominal value of $p = 0.05$ by 10^{-6} based on the assumption that there are approximately 1 million independent tests in the human genome given its length and the frequency and mapping of recombination events.

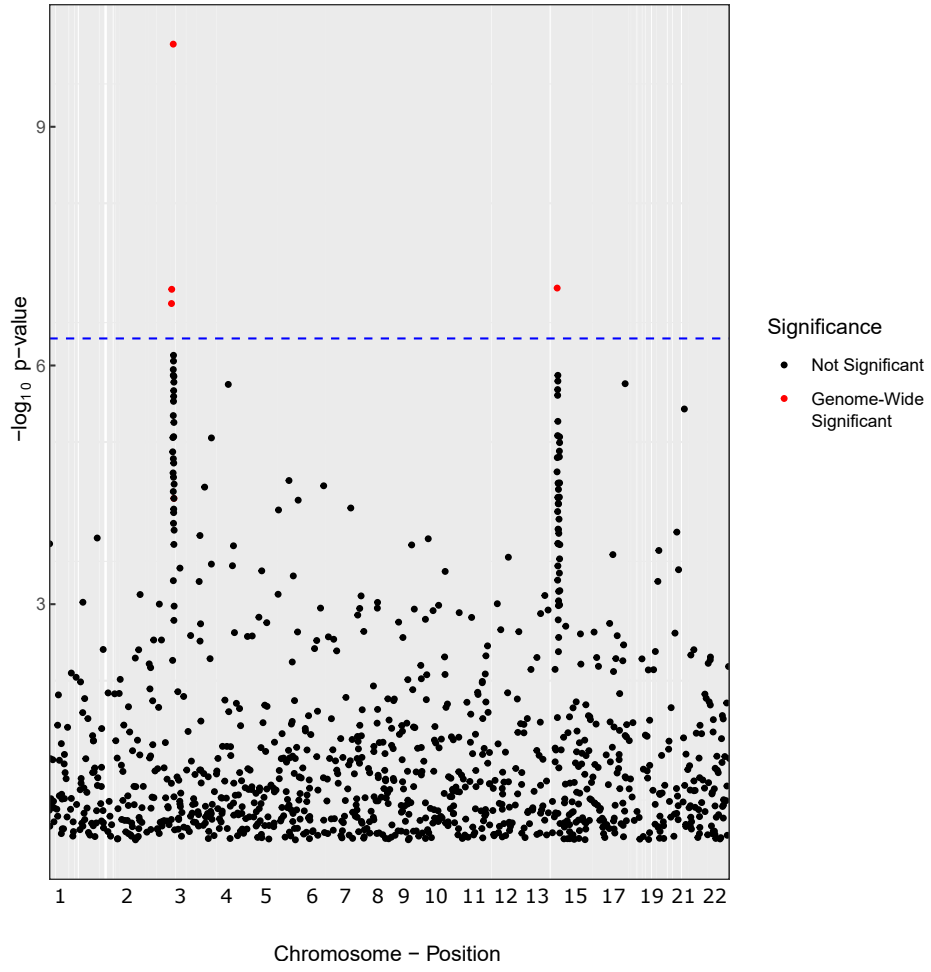


Figure 2.3: An example of a Manhattan plot to visualise GWAS results. The x-axis represents chromosome and position of the particular variant. Each dot represents a genetic variant, and the y-axis shows the negative log10 of the p-value for each variant's association with the phenotype of interest. Variants above the horizontal line (highlighted in red) represent statistically significant associations at the genome-wide level. This plot illustrates simulated data for $p-$values for illustrative purposes.

In GWAS studies, G realistically represents a selected sample of the genome because of two important considerations. The first point is the recognition that not the entire genome is covered. This stems from a historical technological obstacle. There are now technologies available to accurately characterise the entirety of the genetic profile of an individual (whole-genome sequencing); however, these are expensive and have just started becoming available in large biobanks (the first batch of whole

genome sequencing (WGS) of 100,000 individuals enrolled in UK Biobank was released in November 2021). What is more, for the particular issue of leveraging those rare variants as instruments in MR (Chapter 2.2), these may not explain a substantial proportion of variance in the *population* and common variants from existing GWAS studies are arguably better candidates (see Instrument Strength in Chapter 1.5). To counteract this, a standard in the area is array genotyping, a method that directly types a subset of all genetic variants, coupled with genetic imputation. Genetic imputation is a computational approach that infers an individual's genotype for a particular position that has not been directly characterised, through a comparison with a large reference panel of fully genotyped individuals, so that the statistical dependence (or linkage) of the variants is understood. This allows researchers to indirectly infer an individual's genotype with a high degree of certainty for a much larger set of genetic variants than simply those that are directly observed from the genotyping array [48]. Secondly, even in the covered area of the genome, the models described in Eq. 2.3 and Eq. 2.4 are stable if the variants targeted are commonly observed in the population. It has been observed that imputation methods perform suboptimally when rare variants are not observed. Therefore, a cutoff is commonly applied to exclude rare variants based on minor allele frequency (MAF), with a popular threshold choice being MAF> 0.01.

Two phenomena that started becoming apparent with the evolution of GWAS studies of common 'complex' traits (such as obesity or height) and have informed the development of dedicated MR methods are a) the modest magnitude of individual SNP-trait associations, [49], and simultaneously b) the large number of genetic variants robustly associated with a trait [50]. This motivated a displacement from simple

'Mendelian' traits towards a model with small contributions from many variants. The issues that arise when IV analyses are applied in genetics are discussed in Chapter 2.2.

Practical challenges in data availability, however, obviously limit the ability of individual researchers to perform such analyses because the release of genetic data and baseline variables is uncommon due to privacy concerns and data protection regulations [51]. However, the community is increasingly publishing GWAS results as summary statistics for the strength of association of variants with phenotypes, estimated as shown in Eq. 2.3 and Eq. 2.4. This practice enables collaborative projects (such as GWAS meta-analyses) and facilitates a range of downstream analyses, including an extension of the basic MR approach called *two-sample summary data MR*.[52].

In two-sample MR, the type of data that is required is thus simplified as follows. First, summary statistics for a set of K independent SNPs ($\gamma_{Xk}\ SE_{\gamma_{Xk}}, k = 1 \dots K$) that ideally have prior evidence supporting their causal contribution to X are retrieved from a publicly available GWAS of X. The association of these same variants with the outcome Y are also retrieved in a separate GWAS yielding ($\Gamma_{Yk}\ SE_{\Gamma_{Yk}}, k = 1 \dots K$). As each variant is an independent IV (Chapter 1.5), an individual causal effect for the kth SNP can be estimated as the 'Wald ratio' $\hat{\beta}_{XY,k} = \frac{\Gamma_{Y,k}}{\gamma_{X,k}}$ [53]. In words, the Wald ratio is the ratio of the SNP-Y and SNP-X association. If the causal parameter of interest is a ratio parameter, then the null is one and if it is a difference then it is zero. A combined inverse variance weighted causal effect estimate across all SNPs

can also be calculated as:

$$\hat{\beta}_{IVW} = \frac{\sum_{k=1}^{K} \hat{\beta}_{XY,k} w_{XY,k}}{\sum_{k=1}^{K} w_{XY,k}} \quad \text{where} \quad w_{XY,k} = \text{Var}^{-1}(\hat{\beta}_{XY,k}). \tag{2.5}$$

Typically, uncertainty in the SNP-X association is ignored, which is a reasonable assumption when all SNPs are genome-wide significant for the exposure. This means that

$$\text{Var}(\hat{\beta}_{XY,k}) \approx SE^2_{\Gamma_{Yk}} / \gamma^2_{X,k}, \tag{2.6}$$

and this variance estimate is therefore typically used when calculating the IVW estimate.

The IVW estimation strategy is borrowed from the general meta-analysis literature, where the aim is to quantitatively combine the results of many independent studies that aim to estimate the same, or reasonably similar, quantity (Chapter 1.1). As many concepts in summary-data MR are borrowed from this technique, we will introduce them briefly in this paragraph. The most efficient approach in synthesising such individual estimates is the inverse-variance weighted (IVW) shown in equation (2.7) [54]. Therefore, larger studies with more precise estimates contribute more to the final pooled IVW estimate than smaller studies. For clarity, I present an example of a hypothetical meta-analysis of the effects of physical activity on depression.

Table 2.1: Summary of three hypothetical studies investigating the effect of physical activity on depression. RCT: randomised controlled trial.

Study	Sample Size	Effect Size (β)	Variance (SE^2)
RCT 1	1200	-0.25	0.04
RCT 2	1500	-0.30	0.03
RCT 3	1000	-0.20	0.06

We first calculate the inverse-variance weight that we assign for each study. This

is as follows

$$w_i = \frac{1}{SE_i^2},$$

where $i = 1, ..., 3$ is a study index, and SE_i the corresponding standard error. The weights then are:

$$w_1 = \frac{1}{0.04} = 25, w_2 = \frac{1}{0.03} = 33.33, w_3 = \frac{1}{0.06} = 16.67.$$

We see that the smaller study with a consequently more variable effect size will be less represented in the pooled estimate. A weighted effect size estimate can be calculated as the sum of each study's effect size multiplied by its weight, over the sum of the weights:

$$\hat{\beta}_{IVW} = \frac{\sum i = 1^k w_i \hat{\beta}_i}{\sum_{i=1}^k w_i}.$$

The pooled IVW estimate is then calculated as:

$$\hat{\beta}_w = \frac{(25)(-0.25) + (33.33)(-0.30) + (16.67)(-0.20)}{25 + 33.33 + 16.67} = -0.26.$$

In figure 2.4, a Forest plot with of the summary data estimates is presented.

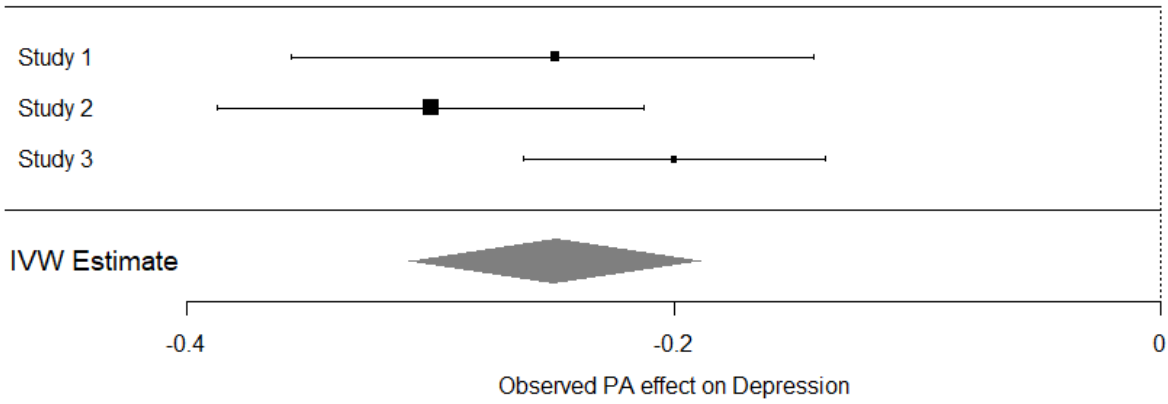


Figure 2.4: IVW meta-analysis of the effect of physical activity on depression. Size of the weight assigned to each study is inversely proportional to the variance of the estimate.

So the weighted effect size estimate for the relationship between physical activity and depression is -0.26, indicating a negative association between PA and depression. In the MR setting, we are therefore effectively treating the Wald ratio causal estimate for each SNP as an estimate from an independent study.

Returning to the general meta-analysis context, in order to perform statistical inference, we need a measure of how variable the pooled IVW estimate is. Its SE is then calculated as:

$$SE_{IVW} = \sqrt{\frac{1}{\sum w_i}}.$$

We see that this SE will be smaller than each of the contributors.

When performing a meta-analysis of studies, we usually expect some of the difference in their estimates to be explained by differences in their population, specific characteristics of the intervention, the exact comparison conducted, or the definition of the outcome. [55]. This can be quantified by estimating between-study variance (*heterogeneity*), a statistical test that characterizes this variability in effect sizes across studies, or in the context of MR how individual SNP causal effects across vary. The following formula is used to estimate the between-study heterogeneity, $\hat{\tau}^2$ and

Cochran's Q statistic as:

$$\hat{\tau}^2 = \frac{Q-(k-1)}{\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}}, \quad \text{where} \quad Q = \sum_{i=1}^{k} w_i(\hat{\beta}_i - \hat{\beta}_{IVW})^2.$$

Cochran's Q statistic tests for heterogeneity among the study effect sizes, which follows a Chi squared distribution on $K-1$ degrees of freedom if all studies are estimating precisely the same quantity. If Q is large with respect to this null distribution, the variance of the IVW estimate can be scaled up to account for the additional uncertainty due to heterogeneity. This can be done under an additive random effects model or a multiplicative random effects model, where in each case w_i is replaced with

$$\begin{aligned} \text{Additive Random Effects} : w_{Ai} &= \frac{1}{\text{Var}(\hat{\beta}_k) + \hat{\tau}^2}. \\ \text{Multiplicative Random Effects} : w_{Mi} &= \frac{\phi^2}{\text{Var}(\hat{\beta}_k)} \quad \text{where} \quad \phi = \frac{Q}{K-1}. \end{aligned}$$

The same theory transfers to summary data MR analyses, where differences in the causal effect estimates across SNPs could indicate possible violations of the IV core conditions, which we will cover in more detail in the next subsection.

As the uncertainty in the variant-exposure association is very low due to the selection process that satisfies IV1, the uncertainty of the Wald ratio is predominantly driven by the variance of the $\Gamma_{Y,1-K}$ summary statistics.

An important property of the IVW estimate of equation (2.5) is that it is asymptotically equivalent to the causal estimate obtained from individual level data using TSLS, as shown in Eq. 1.9. Summary data MR estimates remain valid when the SNP-exposure and SNP-outcome association estimates are obtained from separate

samples (i.e. a two-sample MR design), as long as the samples are sufficiently similar with respect to genetic architecture [56]. For example, if the genetic variants used as IVs have subsantially different allele frequencies in the two samples, this can imply a different population structure and may lead to inaccuracies in the effect estimates. Additionally, if only some of the IV variants on the exposure are located in the outcome sample, this can lead to inconsistencies in the effect estimates and invalid results. These potential concerns are the reason why practitioners use datasets from populations of similar ancestry and by performing appropriate quality control measures to ensure that the genetic variants used as IVs are consistent across the data sets.

2.3.2 Why do we use independent SNPs in summary data MR?

In GWAS studies, particularly those involving complex traits, many SNPs will be genome-wide significant for the exposure. Therefore, one might naively assume that all of them can be used as IV. However, the vast majority of these SNPs will only be associated because they are closely located (or in LD) with a single 'causal' SNP (Chapter 2.1.1). To circumvent this, only the top 'hit' from a distinct genomic region is used as an IV. This means that the SNPs used in the MR analysis are all independent and justify the simple meta-analytic form of the IVW estimate. Methodological advances that enable summary data MR with correlated SNPs have been proposed [57, 58, 59], but are beyond the scope of this thesis.

One popular way of picking the best set of independent SNP IVs is called *clumping* [60]. Let's assume that we have K SNPs that surpass the threshold of 5×10^{-8}. The LD matrix is a $K \times K$ matrix of the frequencies of common pairwise inheritances of

all variants in a given population. If the observed co-occurrence deviates from the one expected, then the two alleles are likely passed on together. Specifically, for two loci i (alleles I, i) and j (alleles J, j), the element ij in the LD matrix can be computed as $r_{ij} = \frac{p_i p_j - p_{ij}}{\sqrt{p_I p_i p_J p_j}}$ (p: observed probability). For our purpose of identifying a set of independent genetic variants, we obtain externally observed estimates of the linkage matrix, we specify a sufficiently low r^2 cutoff value, and only retain for further analyses variants that have the smallest $p-$value and are pairwise correlated to a degree lower than the threshold. A similar approach that does not preferentially retain the largest association but rather one based on MAF among the variants in LD is called *pruning* but is not well suited for MR as strength of association is important for the IV1 core condition; picking independent variants based on MAF might not preferentially choose those that are strongest [61].

Informing instrument selection from GWAS studies offers the advantage of expanding the search scope and increasing the number of SNPs. This increase in turn has been shown to enhance the power of MR analyses, thus making it easier to detect smaller effect sizes or retain power in limited sample sizes [62]. This, however, comes at the cost of possibly including some invalid instruments.

2.4 Why might genetic variants violate the IV core conditions?

2.4.1 Weak Instrument Bias

The validity of MR depends on how confidently we can argue that the IV core conditions are met. The first condition (Chapter 1.5) states that the SNP/IV must be strongly associated with the exposure; if it is weakly associated, a phenomenon known as weak instrument bias will emerge, leading to biased effect estimates [63].

This type of bias has originally been described in traditional instrumental variable analyses [64]. Instrument strength is commonly quantified by the F-statistic, a single measure that captures the strength of the G-X association.

$$\hat{F} = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{\gamma}_j^2}{SE_{\gamma_{Xk}}^2}. \tag{2.7}$$

As a rule of thumb, values lower than 10 indicate that such bias is expected. The nature of the bias is different in one-sample and two-sample MR. In one-sample MR, where both the instrumental variable and the outcome are measured in the same sample, weak instrument bias can result in estimates of the causal effect that are biased towards the observational association. This happens because the residual correlation of X and Y that is not removed by the IV is from the action of the same confounder U in the same sample. On the other hand, in two-sample MR, where X and Y are measured in independent samples, weak instrument bias leads to bias towards the *null*. The magnitude of the dilution of the expected $\hat{\beta}$ is then quantified as

$$\hat{\beta} = \beta \times \frac{F-1}{F}. \tag{2.8}$$

The use of the straightforward F-statistic is an easy means to diagnose its presence, and therefore it is one of the ways in which between-SNP causal effect heterogeneity can emerge.

2.4.2 Horizontal Pleiotropy and pleiotropy-robust MR

As discussed in Chapter 1.5, a crucial assumption in IV analyses is that the IV affects Y only through X; in other words, Y and the instrument are independent conditionally on X and all confounders of X and Y. Particularly for the purposes of applications

of IV in genetic epidemiology as is performed in MR, this is unlikely to hold. A more realistic view of variants is that some of them affect the expression of one or a few genes. If the genes code for or regulate genes that code for proteins, the proteins in turn affect a wide array of phenotypes; multiple proteins are critical actors in many biochemical cascades [65, 66]. The phenomenon whereby a SNP affects the outcome through pathways other than the exposure of interest is termed 'pleiotropy'. It is highly likely that pleiotropy will be present in many analyses, especially those involving complex traits. It is thus unlikely that all retrieved associations of a GWAS for a phenotype (see paragraph 2.3.2) represent true causal relationships. The obtained associations will contain a mix of true causal effects and indirect associations.

The first widely used pleiotropy-robust MR method was published in 2015 by Bowden and co-authors [67]. The authors underline the close relationship of MR with multiple instruments and meta-analysis of RCTs (also see Chapter 2.3). In both cases, individual causal effects are reported and a pooled estimate seeks to synthesise the available evidence to increase precision. Their method repurposed a sensitivity analysis, Egger regression [68], from the meta-analysis methodological literature with the following rationale: In meta-analyses of RCTs, smaller studies that report larger effects of the treatment under investigation are treated preferentially throughout the submission and publication process. This results in their over-representation in the literature and at the same time a lower representation of studies that report null or opposite findings. As a result, the pooled estimate of these studies together with larger studies may not accurately capture the underlying treatment effect. In a similar manner, each SNP provides one causal estimate and the final estimate of the effect is the meta-analysis of all individual effects (see also Chapter 2.3); there will

be deviation from the expected fitted line if there is pleiotropy, akin to publication bias in meta-analyses.

To make the concepts concrete, assume the following data-generating mechanism for X and Y

$$\begin{aligned} X &= \gamma G + U + \varepsilon_X. \\ Y &= \beta X + \alpha G + U + \varepsilon_Y. \end{aligned} \tag{2.9}$$

This differs in an important way to the previous canonical model because the genetic variant G affects Y through X by a magnitude of γ but also violates IV3 as it *directly* affects Y by a factor of α (the pleiotropic effect). This pleiotropic pathway render a TSLS/IVW analysis to be biased. To see why this is the case, we note that under this more general model the true SNP-outcome association for SNP k can be written as

$$\text{SNP k -}Y\text{ assoc}^n : \Gamma_{Yk} = \alpha_k + \beta\gamma_{Xk}. \tag{2.10}$$

The Wald ratio estimand for SNP k is then equal to

$$\begin{aligned} \text{Wald ratio} \quad \frac{\Gamma_{Yk}}{\gamma_{Xk}} &= \beta + \frac{\alpha_k}{\gamma_{Xk}} \\ &= \beta + \text{bias}. \end{aligned}$$

MR-Egger extends a basic IVW analysis by the addition of an additional intercept term in the IVW model (Chapter 2.3). This can directly model this pleiotropic pathway. The weaker assumption that MR Egger makes is that, across all SNPs, the pleiotropic effects α_k are independent of the G-X association (Instrument Strength independent of Direct Effect, InSIDE). MR Egger is shown to provide robust esti-

mates of β even if all SNPs have a pleiotropic effect and their mean value is non-zero, in which case the intercept reflects the average pleiotropic effect.

Although the InSIDE assumption is not as restrictive as assuming all SNPs are valid IVs, there still is a possibility that in many scenarios InSIDE is violated; for example if a SNP affects the outcome through a confounder, then γ and α will be correlated. This motivated the development of the median-based approach [69] which can tolerate more general violations of IV3. Specifically, Bowden et al. show that, if the majority ($> 50\%$) of SNPs are valid (so that their α values are zero), then the median estimator can retrieve the true causal effect even if the InSIDE assumption is violated. The uptake of the methods has been significant, with many applied MR analyses performing them as sensitivity analyses. In Chapter 3, we introduce a new method that is pleiotropy-robust and its validity is based on a more lenient assumption than the InSIDE assumption.

2.5 Multivariable Mendelian Randomisation

An alternative way to address the issue of pleiotropy is through multivariable MR (MVMR). MVMR is an extension of the MR framework that allows for the inclusion of multiple exposures in a single model. This approach was popularised in the MR field in 2015 [70]. The rationale is that prior knowledge of pleiotropic pathways can be used to select multiple exposures that are affected by the same genetic variants, which can then be included in a single multivariable MR model. In such models, the direct effects of each included exposure can be estimated, and if there are variants that impact many of the included exposures, MVMR will appropriately separate these effects and yield an accurate estimate. This can alleviate the issue of IV3 violation.

For example, in the context of the relationship between depression, adiposity and inflammation as proxied by C-reactive protein (CRP), CRP and BMI are known to be strongly correlated. This interplay between CRP and BMI is important to consider in analyses that aim to assess the direct causal effect of CRP. Some genetic variants of C-reactive protein (CRP) have been found to be strongly associated with BMI. A complete assessment of this motivated a detailed applied analysis and is reported in Chapter 5. If we were to use all the GWAS hits as instruments for CRP in an MR analysis, we would likely violate the IV3 core condition if BMI directly affects depression. However, in a MVMR model, we could include both CRP and BMI as individual exposures taking into account their pleiotropic relationship.

More generally, we assume a data generating model for two exposures:

$$X_1 = \gamma_1 G_1 + U + \varepsilon_{X2}.$$

$$X_2 = \beta_{X_1X_2} X_1 + \gamma_2 G_2 + U + \varepsilon_{X1}.$$

Assume that $G*$ contains G_1 and G_2 in Figure 2.5) and the outcome of interest is influenced by both exposures and the confounder:

$$Y = \beta_1 X_1 + \beta_2 X_2 + U + \varepsilon_Y.$$

The directed acyclic graph (DAG) for this model is presented in Figure 2.5. If a genome-wide association study (GWAS) of X_2 is used to guide the choice of the instrument for MR studies, it is likely that some of the identified SNPs, denoted as G^*, would reflect *indirect* associations with X_1. This is because X_2 is influenced by both X_1 and G_2 in the data generating model, and some of the SNPs associated with X_2

may be associated with X_1 indirectly through the $\gamma_2 \times \beta_{X_1 X_2}$ pathway. This pathway also affects the outcome Y through the β_2 coefficient.

Therefore, if we were to use the G^* SNPs as instruments for MR analysis, we would not be able to estimate the direct causal effect of X_2 on Y, as an additional pathway would be inadvertently included in the analysis. To overcome this problem, we can use a multivariable MR (MVMR) approach. Specifically, we can include genetic information on both X_1 and X_2 and simultaneously proxy both exposures with a common set of genetic instruments G^*. This approach allows us, in theory, to retrieve the direct and indirect effect of X_1 on Y through the pathway involving X_2, as well as the direct effect of X_2 on Y through the β_2 coefficient.

In order for MVMR to be valid, the genetic variants used must meet certain assumptions, similar to those required for the previously described IV analyses (Chapter 1.5). Specifically, there are three conditions that the set of variants have to fulfill [70]:

1. G^* is jointly associated with at least one of the exposures,
2. G^* is not associated with any confounding factors of the exposure-outcome associations, and
3. G^* is independent of the outcome, conditionally on the exposures and confounding factors.

The first MVMR condition requires that we can use G^* to genetically predict all of the exposures in the model (in this case X_1 and X_2) whilst additionally guaranteeing that the genetically predicted values are not co-linear. In Chapter 4, we demonstrate a particularly challenging scenario of performing MVMR analysis with multiple, highly correlated exposures using high-resolution metabolite data. We explore how dimensionality reduction techniques, such as principal component analysis, can transform

the exposures to meaningful independent components that suffer less from severe IV condition violations.

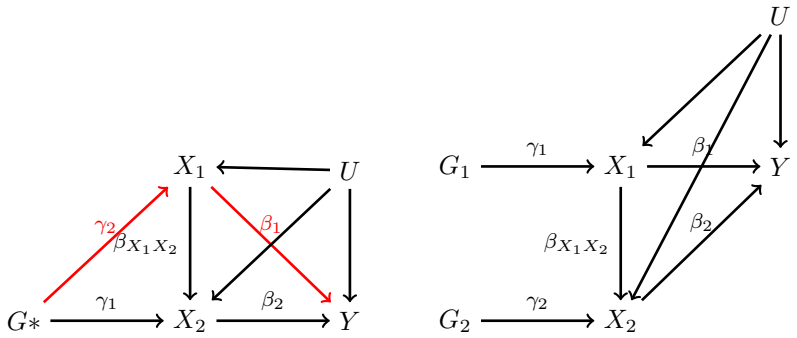


Figure 2.5: Assumed data-generating mechanisms that motivate MVMR. In the left panel, a pleiotropic effect is highlighted in red. In the right panel, two causal effects are explicitly defined and estimated. The instrument selection procedure is generalised to include associations with X_1 *or* X_2.

2.5.1 Mediation Analysis

One active area of research that is a natural continuation of MVMR models is mediation analysis, which looks to further investigate the mechanisms of an exposure's effect on an outcome. For instance, height may have a range of downstream effects that can be measured with MR (especially given the strong genetic component of height), however intervening on height to prevent negative consequences is unrealistic. Hence, a more interesting and actionable question would be an examination of pathways through which height affects health. A hypothetical mediation analysis would estimate the independent effect of height on a given outcome and track how this changes when other, potentially modifiable exposures are also jointly modelled. In observational epidemiology and especially psychology, traditional mediation approaches are based on fitting multiple models based on the inferred causal mechanism. Assuming one exposure and one mediator for simplicity, this approach would include fitting models assuming a data structure consistent with the causal diagrams in 2.6a. First, an estimate for the effect of exposure X on mediator M is obtained

($\hat{\beta}_{XM}$) by regressing X on M. Then, both X and M are used as predictors in a multivariable model. The mediated effect is then defined as the product of $\hat{\beta}_{XM}$ and $\hat{\beta}_M$. A recent extension with many methodologically attractive properties is the combination of causal thinking and mediation analysis [71]. Here, we present the ideas in the context of MR, that is with genetic variants as instruments, as has been previously discussed and implemented by Carter and co-authors [72]. The process is similar to the one described above. A model is first fitted to estimate the effect of genetically proxied X on M (β_{XM}. Then, a multivariable MR model is fitted where both X and M are jointly predicted by a common set of variants to estimate the direct effects β_X and β_M. Finally an MR model is fitted to estimate the total causal effect of X on Y, which equals $\beta_{XM}\beta_M + \beta_X$. From these separate analyses it is then possible to decompose the total causal effect into its direct and indirect (mediated) components. In Chapter 5.3, we present investigations on methods for estimating mediated effects.

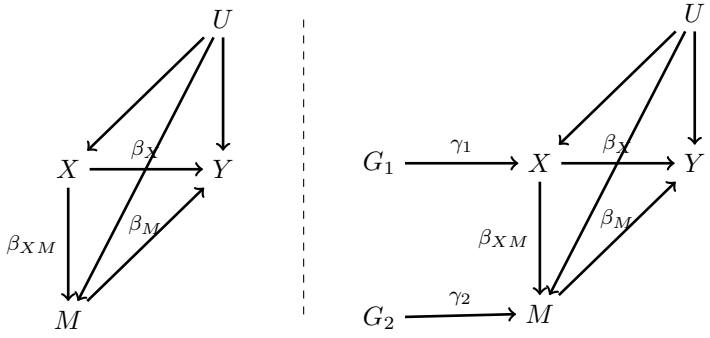


Figure 2.6: Causal mediation analysis in MR.

2.6 Summary & Aims of Dissertation

In summary, MR provides a useful way of leveraging genetic data to infer the causal influence of exposures on health outcomes subject, which can avoid bias due to confounding, but relies instead on a variety of strong assumptions. In Chapter 3, we

present a novel approach that uses gene-by-sex interactions to accommodate extensive pleiotropy that violates the InSIDE assumption. In Chapter 4, we investigate a series of dimensionality reduction approaches to simultaneously assess a large number of highly correlated exposures in a joint MVMR model. In Chapter 5, we report an applied analysis on how inflammation and body mass independently affect a range of mood-related outcomes. We conclude with a discussion of the main contributions of our work to the MR literature and point to further work that remains to be done.

Chapter 3

Sex Stratification and Pleiotropy

3.1 Introduction

The present Chapter describes a novel approach of pleiotropy-robust estimation of a causal effect in MR (Chapter 2.4). Through an investigation of gene-sex interactions, an alternative instrument that targets sex interactions can cancel out the pleiotropic contribution, subject to certain more lenient assumptions regarding the distribution of the pleiotropic effects. Parts of this work have been published in *Genetic Epidemiology* [73].

As presented in Section 1, pleiotropy, a direct effect of the SNPs used as instruments on the outcome of interest, is one of the factors that may hinder the validity of the MR estimates. While pleiotropy is increasingly recognised as an inherent characteristic of the genome, IV analyses and their applications in MR have strong assumptions that require the variables used as instruments to be much simpler in their function and ideally affecting only the targeted exposure. Therefore, additional care has to be taken to translate the results of GWAS studies to reliable indicators of SNPs as instruments. In Chapter 2.5, we describe how prior information on how a given set

of variants affects multiple exposures (*measured pleiotropy*) can guide the construction of a joint MVMR model. In Chapter 2.4, we present the different models that are pleiotropy-robust without any requirements on specifying the phenotype through which the pleiotropic effect is exerted (*unmeasured pleiotropy*). These models have desirable properties as it is not always feasible to point exactly to the physiological pathway (Chapter 2.5), more so in less well studied variants or variants in non-coding regions.

In this chapter, we introduce an approach in this latter category of robust methods. Robustness to pleiotropy is conceptualised differently and is cancelled exactly, without the strong distributional assumptions of other methods. We follow the approach of previous works on gene-by-environment interaction (GxE) and focus on a binary environmental interaction that conveniently allows for the cancellation. We generalise the approach to accomodate more realistic scenarios of targetting weaker interactions.

3.2 Gene-Environment Interactions & MR

Recent observations suggest that the data generating mechanisms of gene-phenotype associations are not optimally described by the simple single-phenotype regressions performed in GWAS studies (Equations 2.3, 2.4) but are rather parts of larger networks with multiple gene-by-environment interactions. In line with these findings, pleiotropy-robust MR methods that leverage these interactions have been proposed. The characteristic property of such an interacting environmental trait is that it modulates the magnitude of the association of the variant with the phenotype. In DAG notation, this can be visualised as an additional interacting node between G and S

(Figure 3.1). This assumed data generating mechanism can then serve as a source of alternative instruments for MR. The canonical example in MR is alcohol consumption, aldehyde dehydrogenase-2 (ALDH2) and sex ([39], Chapter 3.1). In the existing body of literature that makes use of GxE interactions to provide pleiotropy-robust MR estimates, a prime example is the investigation of the effect of alcohol consumption on blood pressure in an Asian population [39]. In this study, the authors use homozygosity status for a common polymorphism in alcohol dehydrogenase 2 (ALDH2) as an IV. Biological reasoning justifies this choice as ALDH2 is directly involved in the enzymatic breakdown of ethanol. It is located in the mitochondria and, after the reduction of alcohol to acetaldehyde by ALDH1, converts acetaldehyde to acetate, a nontoxic product. A point mutation causes a drastic reduction in this enzymatic activity and is highly prevalent in Asians. What is of particular interest to the scope of this chapter is the sensitivity analysis that the authors use to test the validity of the ALDH2 instrument. Given the culturally determined low consumption of alcohol among women, they hypothesised that any association with ALDH2 homozygosity status with blood pressure would reflect a pleiotropic effect. They interpret the null association as supportive of the IV3 core condition, that is that ALDH2 exerts its effect on blood pressure *only through* alcohol consumption. The validity of the MR estimates would depend on the extent to which the interaction is strong.

Spiller et al. [74, 75] have developed the MR-GxE method, a formal framework that *explicitly* models the interaction. They use the example of a varying G-BMI association across strata of TDI [75] and a range of anthropometric, lifestyle, and disease status covariates [74].

Another related method that *implicitly* uses interactions is the MR G-Estimation un-

der No Interaction with Unmeasured Selection (MR-GENIUS) [76]. Tchetgen Tchetgen et al. use the Lewbel's estimator for endogenous regressors with heteroskedasticity [77] and extend it to allow for violations of IV2 and IV3 (here discussed in Chapter 1.5). A crucial advantage of this approach is the appropriate performance without a need for specification of the interacting covariate.

With the present work, we wish to address some issues in the practical use of interactions in MR. First, within the MR GxE framework, the authors acknowledge that complications can arise when the true data generating mechanism includes an interaction variable that is in reality *downstream* of both G and X. It is particularly difficult to completely exclude this possibility for complex traits that occur in later life stages, such as education, income, or behavioural traits. Secondly, it may be more difficult to obtain interactions that are as strongly predictive of X as the associations are. Assuming that such very strong interactions are rare, which is generally observed in GxE studies [78], leveraging weaker interactions could induce a dilution of the estimated effect; thus, a weak-instrument robust approach may be warranted.

In Section 3.3, we describe the pleiotropy cancellation process and the simulation designs that will be investigated.

3.3 Data Generating Mechanisms

We consider the following models for a continuous exposure X and a continuous outcome Y of a set of i.i.d individuals. The causal mechanism is visualised in Figure 3.1.

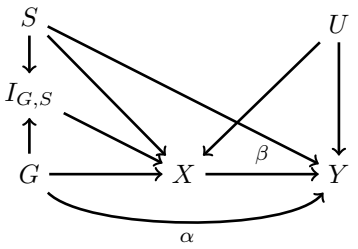


Figure 3.1: Causal diagram representing the assumptions of an MR GxE analysis with a binary covariate of interaction (Sex). G: Genetic variant used as instrument; X: Exposure; Y: Outcome; U: Confounder; S: Sex; $I_{G,S}$: Gene-Sex Interaction Variable.

$$X|G_{\cdot}, S, U = \sum_{j=1}^{k} \gamma_j G_j + \beta_{XS} S + \sum_{j=1}^{k} \Delta_j S G_j + \beta_{UX} U + \varepsilon_X. \quad (3.1)$$

$$Y|X, S, G_{\cdot}, U = \beta X + \sum_{j=1}^{k} \alpha_j G_j + \beta_{UY} U + \beta_{SY} S + \varepsilon_Y. \quad (3.2)$$

A set of k independent genetic variants influence the exposure X (magnitude of effect γ) and the outcome Y (α). This pleiotropic effect α invalidates IV3 and its distribution determines whether IVW or the pleiotropy-robust methods can retrieve consistent estimates for β (Chapter 2.4). If this effect is normally distributed and zero-centred ($\alpha \sim N(0, \sigma_\alpha^2)$) and is orthogonal to the strength of the $G - X$ association ($\alpha \perp\!\!\!\perp \gamma$), then IVW is asymptotically unbiased. If α satisfies the InSIDE assumption, then MR Egger may be useful; alternatively, if the majority of the α elements are zero, that is if most of the SNPs are valid, then the median- and mode-based estimators can be accurate (Chapter 2.4). We will be focusing on challenging cases where the InSIDE assumption is violated. At this stage of presentation, we keep Δ and α independent but we elaborate on this in Section 3.4 where we describe when and how it can be relaxed. The binary covariate S and G both contribute to a multiplicative interaction $I_{G,S}$ which affects the exposure (Δ). We also assume that there is a direct effect of S on X and Y.

The target of the estimation is the average causal effect of an intervention on X by a single unit, while all other factors in model (2) are held fixed. In potential outcomes notation, this causal effect can be written as the population average:

$$\beta = E[Y(X) - Y(X-1)].$$

where β is the coefficient of X in outcome model (2). When the data are generated according to models (3.1) and (3.2), applying standard Two Stage Least Squares (2SLS) to estimate the association between Y and the genetically predicted exposure will not yield a consistent estimate of the causal effect, because each genetic variant exerts a direct pleiotropic effect on Y not through X. This bias can be seen when we rewrite model (3.2) and express Y as a function only of G, S and U in (3.3),

$$Y|G_{\cdot}, S, U = \sum_{j=1}^{k}\{\alpha_j + \beta(\gamma_j + \Delta_j S)\}G_j + (\beta\beta_{SX} + \beta_{SY})S + (\beta\beta_{UX} + \beta_{UY})U + \beta\varepsilon_X + \varepsilon_Y \quad (3.3)$$

Thus, this α term in Eq. 3.3 is carried through when we predict X with G and inaccuracies can ensue. If S is unobserved, but follows a distribution with probability mass function $P(S)$ that is independent of G and U, we can then marginalise over S. Here we assume that $E(S|G,U) = E(S) = \frac{1}{2}$, which gives the reduced form model in (3.4):

$$\begin{aligned} Y|G_{\cdot}, U &= \sum_{j=1}^{k}\{\alpha_j + \beta(\gamma_j + \frac{1}{2}\Delta_j\}G_j + \frac{1}{2}\beta_{SY} + (\beta\beta_{UX} + \beta_{UY})U + \beta\varepsilon_X + \varepsilon_Y \\ &= \sum_{j=1}^{k}(\alpha_j + \beta\gamma_j^*)G_j + \varepsilon_Y^* \\ &= \sum_{j=1}^{k}\Gamma_j^* G_j + \varepsilon_Y^*. \end{aligned} \quad (3.4)$$

The Wald ratio causal estimand for a single SNP j is the ratio of the G_j-Y association, Γ_j^*, and the G_j-X association averaged over S, γ_j^* :

$$\beta_j = \beta + \frac{\alpha_j}{\gamma_j^*}. \tag{3.5}$$

When all SNPs are chosen to be mutually independent (not in LD, Chapter 2.1.1), the 2SLS estimate is asymptotically equivalent to an inverse variance-weighted average of the SNP-specific causal effect estimates (Chapter 2.3). The IVW estimate is generally used in MR studies because of this asymptotic equivalence, but also because it can be calculated with only summary data (Chapter 2.3, [79]). MR with summary data also facilitates the inspection of heterogeneity in causal estimates across SNPs (for example due to pleiotropy).

3.3.1 Robustness to Pleiotropy and Weak Instrument Bias

As described in Chapter 2.4, the IVW estimate is able to consistently estimate the causal effect as long as the pleiotropic effects α_j follow a zero-centered distribution and are independent of the SNP-exposure associations. This holds if the sample covariance $\widehat{\text{Cov}}(\alpha_j, \gamma_j^*)$ = 0, which is referred to as the InSIDE assumption [80]. The InSIDE assumption is automatically satisfied if α_j = 0 for all SNPs. If some SNPs have a non-zero pleiotropic contribution to Y, then IVW can still be valid if the distribution of the effects suggests that effectively pleitropic contributions are cancelled out within this set ($\alpha \sim N(0, \sigma_\alpha^2)$). Still in that case, there will be heterogeneity among the individual causal effect of each SNP and this additional heterogeneity can be modelled by performing a random effects (RE) meta-analysis. Under an additive RE

formulation, this is equivalent to fitting

$$\hat{\Gamma}_j^* = \beta\hat{\gamma}_j^* + \alpha_j + \sigma_{Yj}\varepsilon_j, \quad \varepsilon_j \sim N(0,1), \quad \alpha \sim N(0,\tau^2). \tag{3.6}$$

This is most commonly done using least squares to estimate β and the DerSimonian and Laird moment-based estimate for τ^2 [81]. In meta-analyses of trials, this modelling approach assumes that the true study effect in each individual study is different and the RE model this difference by estimating a variance component for the total effect.

The InSIDE assumption could be implausible in many settings and further robust estimation strategies that rely on alternative identifying assumptions have been developed (median and weighted median estimators [69], mode-based estimator [82], Chapter 2.4).

Another issue affecting MR studies is weak instrument bias. For example, in the two sample context, the IVW estimate is known to be diluted towards zero by a factor of

$$x = \frac{\bar{F} - 1}{\bar{F}},$$

where $\bar{F} = \frac{1}{K}\sum_{j=1}^{K} \frac{\gamma_j^2}{SE_{\gamma_j}^2}$ is the mean F-statistic. Weak instrument bias can be corrected for with Simulation-Extrapolation [83]. A benefit of the Robust-Adjusted Profile Score (MR-RAPS) approach [84] is that it offers an exact solution for dealing with weak instrument bias as well as one form of pleiotropy. It estimates the value of β and τ^2 that maximises the profile log-likelihood:

$$l(\beta,\tau^2) = -\frac{1}{2}Q(\beta,\tau^2) + \log(\sigma_{Yj}^2 + \tau^2),$$

where

$$Q(\beta,\tau^2)=\sum_{j=1}^{k} w_j(\beta,\tau^2)(\hat{\beta}_j-\beta)^p, \quad \text{and} \quad w_j(\beta)=\frac{\hat{\gamma}_j^2}{\sigma_{Yj}^2+\tau^2+\beta\sigma_{Xj}^2}, \quad \sigma_{Xj}^2=\text{Var}(\hat{\gamma}_j). \tag{3.7}$$

Here τ^2 is the pleiotropy variance and p denotes a user-specified loss function. Some examples include the standard L_2 loss (p=2), or a customised function that enforces robustness to (pleiotropic) outliers such as Huber or Tukey loss functions, as described in [56]. With these two modifications, MR-RAPS can then be heuristically viewed as a weak instrument and pleiotropy robust combination of the IVW and median-based methods.

3.3.2 SNP-Level Cancellation of Pleiotropy

Model (3.1) includes an interaction term between each genetic instrument and the binary covariate S; individuals with different values of S thus have different strengths of SNP-exposure associations. When performing an MR analysis, we would calculate SNP-exposure associations by *marginalising* over S to estimate the summary quantities $\hat{\gamma}_j^*$ and $\hat{\Gamma}_j^*$. We will explicitly make use of this covariate when constructing our pleiotropy-robust analysis. Under models (3.1) and (3.2), we can express the true SNP-exposure and SNP-outcome associations within each stratum of S as

$$
\begin{aligned}
\gamma_{j1} &= E[X|G=1,S=1] - E[X|G=0,S=1] = \gamma_j + \Delta_j. \\
\gamma_{j0} &= E[X|G=1,S=0] - E[X|G=0,S=0] = \gamma_j. \\
\Gamma_{j1} &= E[Y|G=1,S=1] - E[Y|G=0,S=1] = \beta(\gamma_j + \Delta_j) + \alpha_j. \\
\Gamma_{j0} &= E[Y|G=1,S=0] - E[Y|G=0,S=0] = \beta\gamma_j + \alpha_j = \Gamma_j.
\end{aligned}
$$

From these equations, we observe that the difference in SNP-outcome associations between each strata of S divided by the difference in SNP-exposure associations between each strata of S cancels out exactly the pleiotropic contributions and identifies the causal effect:

$$
\frac{\Gamma_{j1} - \Gamma_{j0}}{\gamma_{j1} - \gamma_{j0}} = \frac{\beta(\gamma_j + \Delta_j) + \alpha_j - (\beta\gamma_j + \alpha_j)}{\gamma_j + \Delta_j - \gamma_j} = \beta\frac{\Delta_j}{\Delta_j} = \beta. \tag{3.8}
$$

This estimand has some attractive properties. Rather than explicitly accounting for additional heterogeneity due to pleiotropy under InSIDE, the pleiotropic effect of each SNP is cancelled out exactly. This cancellation is not affected by whether the pleiotropy violates InSIDE. This in turn can allow a fixed effect analysis, which may make it more efficient than methods which explicitly model residual heterogeneity due to pleiotropy.

3.3.3 Estimation

Two-Sample Setting

We first consider the case of data availability for two independent samples summary data (one for SNP-exposure and one for SNP-outcome associations). In both sam-

ples, the data generating models 3.1 and 3.2 hold. In the first sample, we obtain estimates of the association between G_j ($j = 1, .., K$ SNPs) and X at each level of the interacting variable S ($\hat{\gamma}_{0j}$ (σ^2_{X0j}) and $\hat{\gamma}_{1j}$ (σ^2_{X1j}) . In the second sample, we obtain estimates of the association between G_j and Y at each level of S ($\hat{\Gamma}_{0j}$ (σ^2_{Y0j}) and $\hat{\Gamma}_{1j}$ (σ^2_{Y1j})). We can then readily calculate *for each SNP* the ratio of the difference in sexes in SNP-Outcome associations over the difference in sexes in SNP-Exposure associations, as described in Eq. 3.8. A pooling of the individually estimated SNP effects can then be performed with the IVW meta-analysis method (Chapter 2.3). We refer to this as the 'sex-stratified IVW' estimate. As the target of this instrument is the interaction and it is possible that many interactions will be weaker in magnitude, the causal effect estimate may in turn suffer from weak-instrument bias. The dedicated formula that quantifies the degree of this anticipated regression dilution (due to weak instruments) is as follows

$$\frac{\bar{F}_{\text{Strat}} - 1}{\bar{F}_{\text{Strat}}}, \quad \text{where } \bar{F}_{\text{Strat}} = \frac{1}{k}\sum_{j=1}^{k} \frac{(\hat{\gamma}_{j1} - \hat{\gamma}_{j0})^2}{\sigma^2_{X0j} + \sigma^2_{X1j}}. \tag{3.9}$$

To corroborate this, we present a simulation study that shows how the observed values of regression dilution agree with the expected ones that F_{strat} suggests (Figure A.1). Generally, we would expect F_{strat} to be lower than F, as $\hat{\gamma}_{j1}$-$\hat{\gamma}_{j0}$ would need to have opposite signs or drastically different values for this not to occur. To protect the analyses from this anticipated weak-instrument bias, we propose to use sex-stratification within the MR-RAPS framework [84], by adapting its inputs to fit our setting. Specifically, we will estimate the values of β (and potentially τ^2) which maximises the profile log-likelihood but with our previous definition $Q(\beta, \tau^2)$ replaced

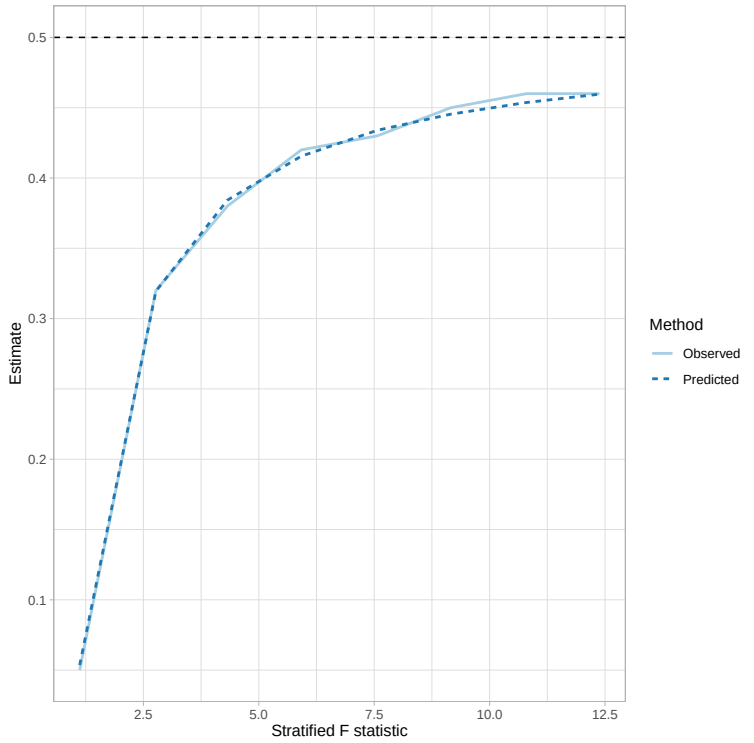


Figure 3.2: Agreement of the observed regression dilution and the formula-predicted one (Eq. 3.9). dashed black line: causal effect of X on Y ($\beta = 0.5$); Observed: mean estimate of the causal effect across all simulations $\overline{\hat{\beta}}_{IVWstrat}$; Predicted: $\beta \times \frac{\hat{F}_{strat}-1}{\hat{F}_{strat}}$.

with

$$Q(\beta, \tau^2) = \sum_{j=1}^{k} w_j(\beta, \tau^2) \left(\frac{\hat{\Gamma}_{j1} - \hat{\Gamma}_{j0}}{\hat{\gamma}_{j1} - \hat{\gamma}_{j0}} - \beta \right)^p, \quad w_j(\beta, \tau^2) = \frac{(\hat{\gamma}_{j1} - \hat{\gamma}_{j0})^2}{\sigma^2_{Y0j} + \sigma^2_{Y1j} + \tau^2 + \beta^2(\sigma^2_{X0j} + \sigma^2_{X1j})}. \tag{3.10}$$

Fixing τ^2 to zero and estimating β under a fixed effect analysis is justified under our proposed data generating models because the pleiotropy terms cancel out. However, over-dispersion may still be present in the causal estimates across SNPs, which could be due to a variety of factors including (but not limited to) pleiotropy. For example, a SNP's pleiotropic effect may differ across sexes because it is differentially expressed in men and women (scenario 3, Section 3.4.1). In this case, we would want to calculate and report heterogeneity adjusted estimates under a random effects model. We therefore propose a three-step strategy:

1. Use MR-RAPs to find $\hat{\beta}$ which maximises profile log-likelihood for $Q(\beta, 0)$ in (3.10) under a fixed effect model;

2. If the value of the Q statistic $Q(\hat{\beta}, 0)$ (evaluated by setting $p = 2$) is $\leq$ the (1-ψ)th percentile of a χ^2_{k-1} density using a pre-specified Type I error threshold of ψ then use the fixed effect estimate;

3. If $Q(\hat{\beta}, 0)$ is $>$ (1-ψ)th percentile of a χ^2_{k-1}, replace $Q(\beta, 0)$ with $Q(\beta, \tau)$ and maximise the profile log likelihood under a random effects model.

We will subsequently explore the performance of this estimation strategy that attempts to use the more statistically efficient fixed effect analysis where possible and only the random effects analysis if necessary.

One-Sample Setting

When implementing the procedure in the one sample setting, uncertainty in the SNP-exposure and SNP-outcome associations will be correlated, which violates the key condition of the two-sample MR approach and MR-RAPS. To account for this, we employ a novel extension of the general Collider-Correction approach [85] in the sex-stratified setting, to enable the analysis to proceed using the standard MR-RAPS software. Borrowing the original terminology in [85], the algorithm is as follows:

1. Regress the exposure X on G_j within each level of S separately, to give $\hat{\gamma}_{1j}$ and $\hat{\gamma}_{0j}$ for each SNP;

2. Regress Y on X, G_j and S and extract:

    - The collider-biased estimated coefficient of X, $\hat{\beta}^*$ with variance $\sigma^2_{\hat{\beta}^*}$;

- The collider-biased estimated coefficients for $G_j|S=0$ ($\hat{\alpha}^*_{j0}$) and $G_j|S=1$, ($\hat{\alpha}^*_{j1}$), for each SNP, with variances $\sigma^2_{\alpha^*0j}$ and $\sigma^2_{\alpha^*1j}$ respectively.

3. Under models (3.1) and (3.2), the parameter estimates in step 1 and 2 are linked via

$$\begin{aligned}\hat{\alpha^*}_{1j} &= \alpha_j + (\beta - \beta^*)\hat{\gamma}_{1j} + \varepsilon_{1j}.\\ \hat{\alpha^*}_{0j} &= \alpha_j + (\beta - \beta^*)\hat{\gamma}_{0j} + \varepsilon_{0j}.\end{aligned}$$

We therefore use MR-RAPS to estimate the single parameter $(\beta - \beta^*)$ which maximises the profile log likelihood where $Q(\beta, \tau)$ is replaced with

$$\begin{aligned}Q_{CC}(\beta - \beta^*, \tau^2) &= \sum_{j=1}^{k} w_j(\beta - \beta^*, \tau^2)\left(\frac{\hat{\alpha^*}_{j1} - \hat{\alpha^*}_{j0}}{\hat{\gamma}_{j1} - \hat{\gamma}_{j0}} - (\beta - \beta^*)\right)^p, \quad \text{where}\\ w_j(\beta - \beta^*, \tau^2) &= \frac{(\hat{\gamma}^2_{j1} - \hat{\gamma}^2_{j0})^2}{\sigma^2_{\alpha^*0j} + \tau^2 + \sigma^2_{\alpha^*1j} + (\beta - \beta^*)^2(\sigma^2_{X0j} + \sigma^2_{X1j})},\end{aligned}$$

where τ^2 is at first set to zero but is estimated under a random effects model if $Q_{CC}(\widehat{\beta - \beta^*}, 0)$ is adequately large. The variance of $\widehat{\beta - \beta^*}$ is $\sigma^2_{\beta - \beta^*}$

4. Estimate the causal effect $\hat{\beta}$ as $\hat{\beta}^* + \widehat{\beta - \beta^*}$, with variance $\sigma^2_{\beta^*} + \sigma^2_{\beta - \beta^*}$

3.4 Simulation Studies

In this section, we compare the sex-stratified MR estimator against alternative approaches in various simulation scenarios of both individual level data (one-sample MR) and summary data (two-sample MR). In all scenarios, model (3.1) was used to generate continuous exposure data. Continuous outcome data were generated from a broader version of model (3.2) presented below:

$$Y|X,S,G_.,U = \beta X + \sum_{j=1}^{k} \alpha_j G_j + \beta_{UY} U + \beta_{SY} S + \sum_{j=1}^{k} \theta_j S G_j + \varepsilon_Y. \tag{3.11}$$

The difference between model (3.2) and model (3.11) has an additional interaction term between sex and the genetic instruments, whenever the parameters $\theta_1, \ldots, \theta_j$ are non-zero. Intuitively, this represents a setting of a direct and sex-specific effect of G on Y. We note that, at a minimum, this additional interaction would prohibit the perfect cancellation of the pleiotropy terms in the sex-stratified analysis as described in 3.3.2. As a result, a fixed effect analysis would not be the optimal modelling choice. In Figure 3.3, an overview of the data generating mechanism is presented. The data generating models for all of the data, the precise parameter choices and accompanying code are available at `github.com/vaskarageorg/stratMR`, but important summary details of the simulation are now given:

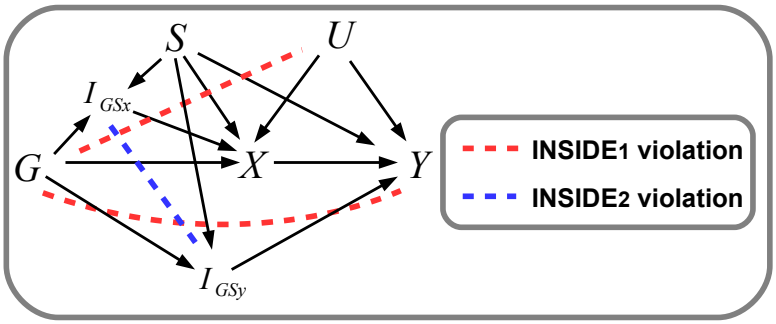


Figure 3.3: Left: Causal diagram representing the underlying data generating model in equation 3.11

- In all scenarios, the number of SNPs, denoted by k, was maintained at 25 and accounted for roughly 5% of the variance in X. In each iteration, the parameters γ_j, Δ_j, α_j, and θ_j were generated randomly from separate distributions, and were connected by the following equations.

$$\alpha_j = \mu_{\alpha,j} + \beta_\alpha \gamma_j. \quad (3.12)$$

$$\theta_j = \mu_{\theta,j} + \beta_\theta \Delta_j. \quad (3.13)$$

- The expectation of the sample covariance, $\widehat{\text{Cov}}(\alpha_j, \gamma_j)$, is equal to zero when β_α is set to zero, which we describe as the fulfillment of the "InSIDE(1)" assumption. In general, we quantify the magnitude of this violation using $\rho_1 = E[\widehat{\text{Cor}}(\alpha_j, \gamma_j)]$.

- When β_θ is set to zero, the expectation of the sample covariance $\widehat{\text{Cor}}(\theta_j, \Delta_j)$ is zero, which we denote as the "InSIDE(2)" assumption being fulfilled. We examine situations where the InSIDE(2) assumption is both fulfilled and violated, using $\rho_2 = E[\widehat{\text{Cor}}(\theta_j, \Delta_j)]$ to quantify the degree of InSIDE(2) violation.

- For individual-level one-sample analyses, data were extracted from the models (3.1) and (3.11) for the same n individuals, resulting in correlated errors. Conversely, for two-sample analyses, data were obtained from the models (3.1) and (3.11) for two separate sets of n individuals, resulting in independent errors. The value of n was varied to produce data with a range of F_{Strat}-statistics between approximately 1 and 15, as determined by equation (3.9).

- In separate simulations, we fixed the causal effect either to β=1 or to β=0 in order to evaluate power and Type I error (T1E) respectively. Other metrics of performance included bias and coverage.

3.4.1 Two-Sample Summary Data Setting

In Scenario 1, we generated two-sample summary data under moderate to severe violations of the InSIDE(1) assumption (ρ_1=0.6 and 0.9), with the InSIDE(2) assumption being trivially fulfilled by setting all θ_j=0. We compared the results of four methods: the standard IVW estimate, the standard MR-RAPS estimate, a sex-stratified IVW estimate, and a sex-stratified MR-RAPS estimate, all of which were performed using a fixed-effects analysis. The operating characteristics are shown in Figure 3.4 for a range of sample sizes that produced $\bar{F}$ statistics between 1 and 180 and $\bar{F}_{strat}$ statistics between 1 and 10.

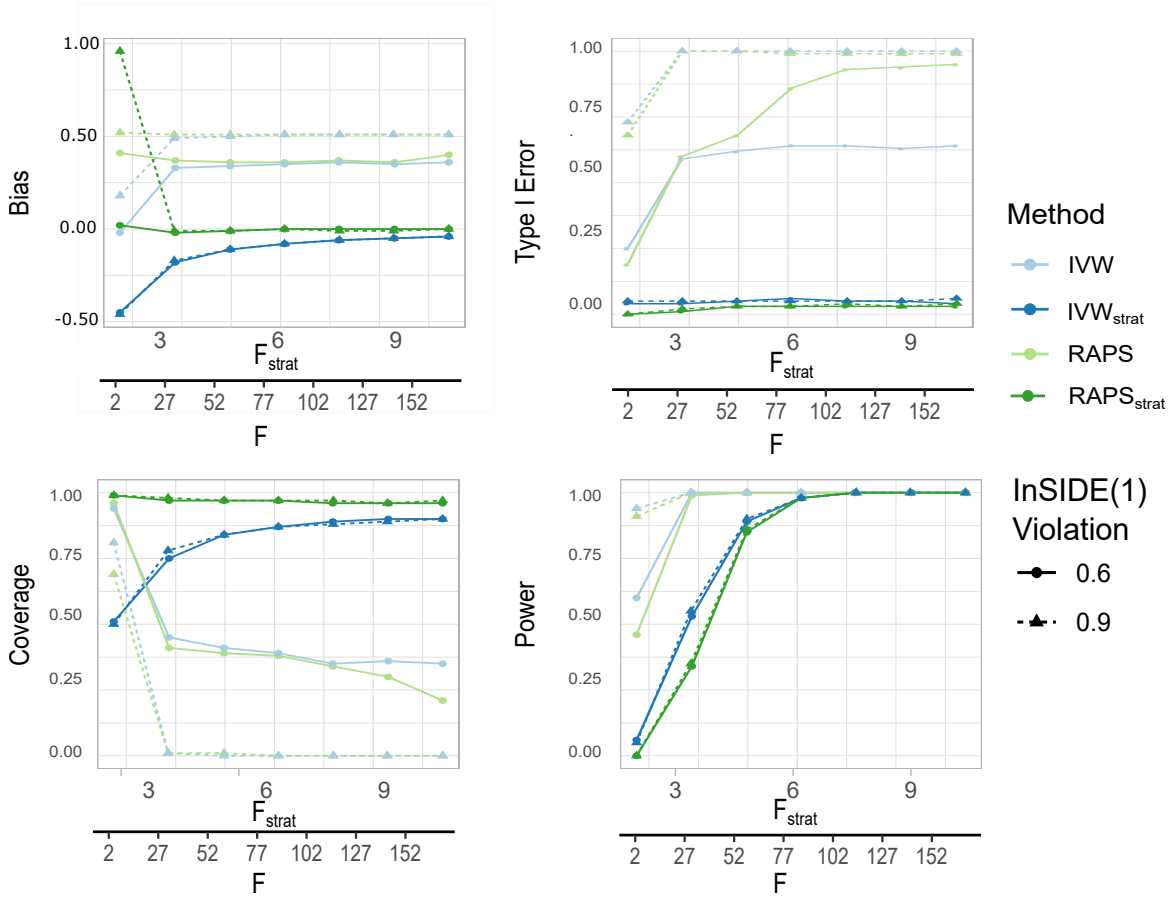


Figure 3.4: Bias, Type I error (TIE), Coverage and Power for different degrees of InSIDE(1) violation as a function of increasing sample size (and hence increasing F and F_{strat} statistics). TIE assessment is based on a 5% level test under the null ($\beta = 0$).

According to Figure 3.4, standard IVW and MR-RAPS estimates yield biased outcomes regardless of the strength of the instrument, with a growing bias for increasing ρ_1 values. This lack of accuracy is reflected in inadequate coverage and high TIE inflation. The stratified IVW and MR-RAPS estimates, on the other hand, perform much better. The IVW estimate stratified by sex displays some weak instrument bias when $\bar{F}_{strat}$ values are small, but this decreases as instrument strength increases, as theorised in formula (3.9). Further evidence of this is presented in the *Appendix* (A.1.1). This bias appears unaffected by the magnitude of ρ_1. Coverage and TIE of the stratified IVW estimate approach expected levels as $\bar{F}_{strat}$ increases. For moderate InSIDE(1) violation (ρ_1=0.6), the MR-RAPS estimate stratified by sex is unbiased across all $\bar{F}strat$ values, with coverage and TIE at their nominal levels. For strong InSIDE(1) violation (ρ_1=0.9), some bias is present in the estimate for $\bar{F}_{strat}$ values below 4.

In Scenario 2, our attention was focused exclusively on the sex-stratified IVW and MR-RAPS estimators and we explored how the performance of these estimators was influenced by selecting SNPs based on interaction strength (F_{strat}). Utilising the same data generating mechanism as previously, we varied the proportion of SNPs without any sex-specific effect (i.e. Δ_j = 0) from 0% to 70%, leading to a set of truly sex-dimorphic instruments ranging from 30% to 100% of the available SNPs. We then applied two approaches to the data, (a) using all SNPs and (b) using only SNPs with an individual $F_{strat} \geq 3$. The outcomes are depicted in Figure 3.5. Without the selection, the stratified IVW estimate is significantly biased due to weak instruments. This bias is partially reduced by selecting SNPs based on instrument strength, resulting in a modest increase in coverage and precision. The MR-RAPS estimator stratified

by sex is biased when instruments are very weak ($\bar{F}strat \leq 1$), but this bias vanishes for $\bar{F}strat \geq 3$. With SNP selection, precision of the sex-stratified MR-RAPS estimate increases; however, this comes at the cost of a rise in downward bias, a resulting decrease in coverage and T1E inflation. Our conclusion is that unless the F_{strat} estimate is extremely weak, SNP selection does not enhance the performance of the sex-stratified MR-RAPS estimate and may introduce winner's curse bias.

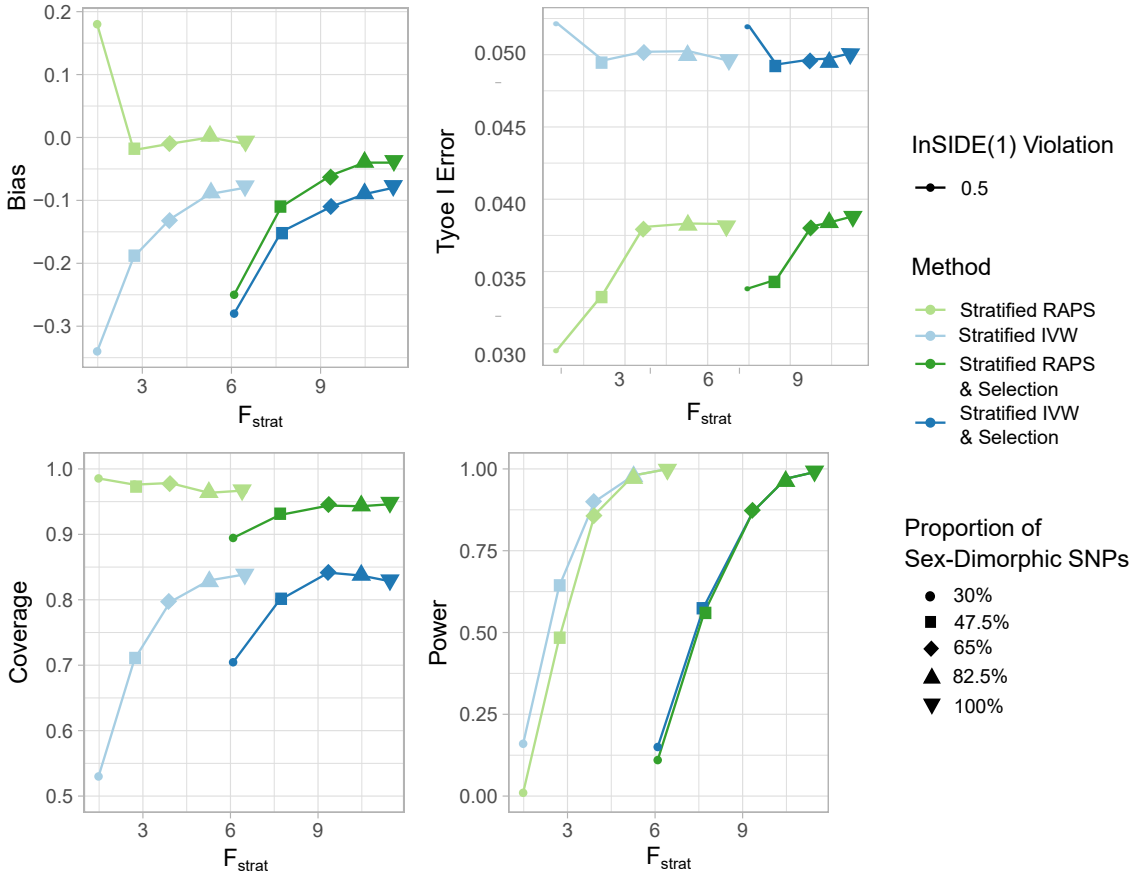


Figure 3.5: Impact of different proportions of SNPs without true interaction ($\Delta_j = 0$) on bias, empirical standard error, Type I error, coverage, and power. The sample size was kept constant and the variation in F_{strat} was solely due to differences in the proportion of SNPs that had sex-differential associations with the exposure X. Selection: only those SNPs that were differentially associated with X were considered. Assessment of the TIE was based on a 5% significance level test under the null hypothesis ($\beta = 0$).

In scenario 3, we examine settings where the pleiotropic effects of SNPs are also sex-specific by making the θ_j term in (3.11) non-zero. We further specify this

pleiotropy as either conforming to the InSIDE(2) assumption ($\rho_2 = 0$) or not (ρ_2=0.5). The degree of InSIDE(1) violation was set to 0.5 in both scenarios. Our objective is to detect this sex-specific pleiotropy as heterogeneity with the Q-statistic and determine whether a fixed or random effects analysis is appropriate, as outlined in Section 3.3.3. This method, referred to as "Choice," is then compared to the constant use of either a fixed effect or a random effect model. The results are displayed in Figure 3.6.

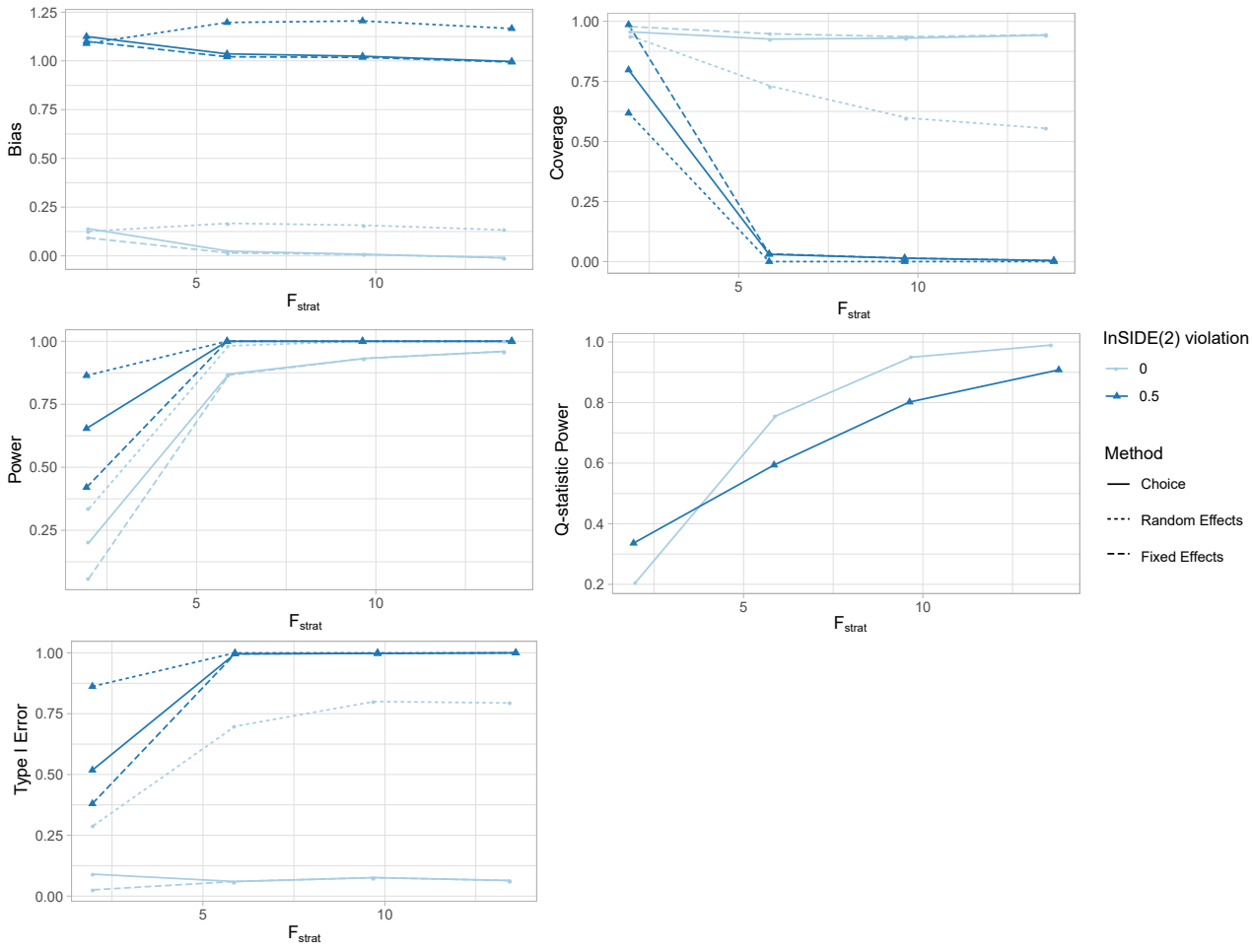


Figure 3.6: Performance measures of the MR-RAPS stratified approach under varying degrees of InSIDE(2) violation and increasing sample sizes. The measures shown are bias, coverage, power, Q-statistic power, and Type I error (T1E) assessed under a 5% level test under the null hypothesis ($\beta = 0$). The sample sizes correspond to mean (standard deviation) F_{strat} statistics of $2.23(0.29), 6.03(0.72), 9.43(1.11)$, and $13.41(1.24)$. The figure displays results for three implementations of the MR-RAPS stratified approach: fixed effect, random effects, and choice.

The results showed that when the InSIDE(2) assumption was satisfied ($\rho_2 = 0$), the random effects sex-stratified MR-RAPS estimator delivered approximately unbiased estimates of the causal effect and achieved close to nominal coverage and Type I Error (TIE). This is reassuring, as the convergence of the over-dispersed MR-RAPS estimator to the single correct parameter value is not guaranteed without further assumptions and constraints on the form of the loss function used as outlined in Section 5 of [86].

On the other hand, the fixed effect implementation was positively biased, leading to a loss of coverage and inflation of TIE. However, using the Q-statistic to guide the choice between a fixed or random effects analysis resulted in estimates that were only minimally biased. These estimates were also more precise than the blanket random effects estimate, although there was a small loss in coverage. The power of the Q-statistic to detect heterogeneity increased with sample size, as shown in Figure 3.6 (bottom-right).

In the challenging scenario of violation of InSIDE(2) (ρ_2=0.5), the findings indicate that both fixed and random effects estimates are subject to bias (deep blue, Figure 3.6). As a result, neither method can provide reliable estimates in this scenario. Utilising the Q statistic for model selection leads to an estimate that has operating characteristics that lie between those of the fixed and random effects models.

In a further simulation, the magnitude of heterogeneity was varied and the performance of the three methods was evaluated (Figure 3.7). This examination allowed for an assessment of the $Q-$statistic test across the full range of possible results. This is of particular relevance as it enables anticipation of varying degrees of sex-specific pleiotropy with different magnitudes, and hence varying levels of diagnostic power

with the $Q-$statistic. The fixed effect (FE) method was found to be biased as heterogeneity increased. The random effects (RE) approach exhibited some overcoverage initially but converged to 95% for higher values of the parameter. The performance of the Choice method was appropriate, although it was noted that there was a slight inflation of T1E ($7-8\%$) for a statistically significant $Q-$statistic in $50-75\%$ of the simulations.

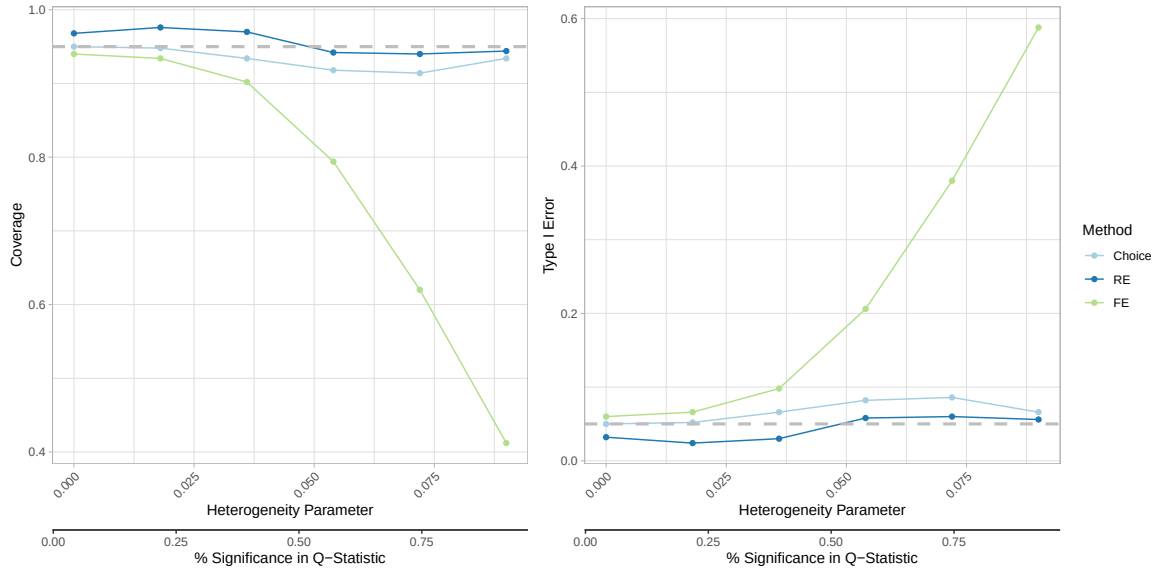


Figure 3.7: Performance of FE, RE and choice between the two based on $Q-$statistic (Choice). Heterogeneity Parameter: θ_j in Equation 12, sex specificity of pleiotropy.

3.4.2 One-Sample Setting

We now illustrate the effectiveness of the sex-stratified MR-RAPS estimator when applied in a one-sample individual-level data scenario using the Collider Correction technique [85]. Similar to the two-sample setting, we first evaluate the performance of the estimator when there is no sex-specific pleiotropy (θ_j=0) using a fixed-effects implementation (Scenario 1). The results are presented in Figure 3.8. For comparison, we also show the results of a Two-Stage Least Squares (2SLS) model that does not take advantage of the sex interaction and of MR GxE [75, 74] which leverages

the interaction but lacks a weak-instrument robust implementation. The stratified MR approach with Collider Correction is generally unbiased, except in simulations with very low F_{strat} values, and attains approximately nominal coverage. The 2SLS estimate is biased and, while it appears to have higher power, this bias results in poor coverage. The MR GxE method is unbiased when the sex interaction is strong, but is susceptible to severe weak-instrument bias when it is not.

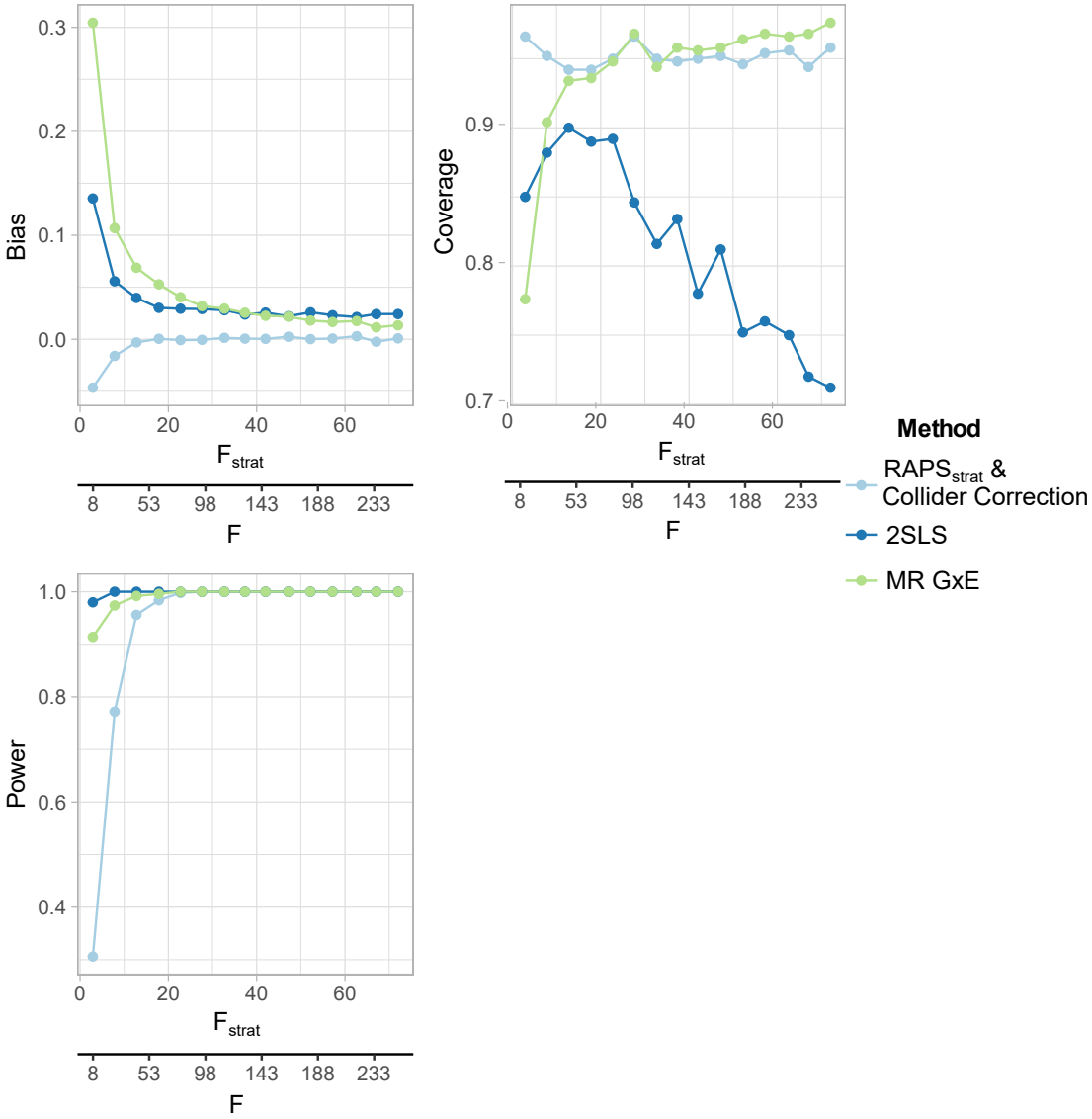


Figure 3.8: Bias, power and coverage of the Collider-Correction sex-stratified MR-RAPS estimator, 2SLS and MR-GxE in the one sample setting.

We now extend Scenario 3 from the two-sample setting to the one-sample setting

to demonstrate the capability of Collider Correction and the Q_{cc} statistic in guiding the selection between a fixed or a random-effects sex-stratified MR-RAPS model. Similar to Scenario 3, data is generated with sex-specific pleiotropy that either satisfies or violates the InSIDE(2) assumption. The results are presented in Figure 3.10. Consistent with the two-sample setting, when InSIDE(2) is satisfied, the results reveal that the fixed-effect estimate is biased and has poor coverage, while the random-effects estimate and the Q_{cc}-driven estimate perform well and are nearly indistinguishable. When InSIDE(2) is violated, the fixed-effect, random-effect, and Q_{cc}-driven estimates perform poorly.

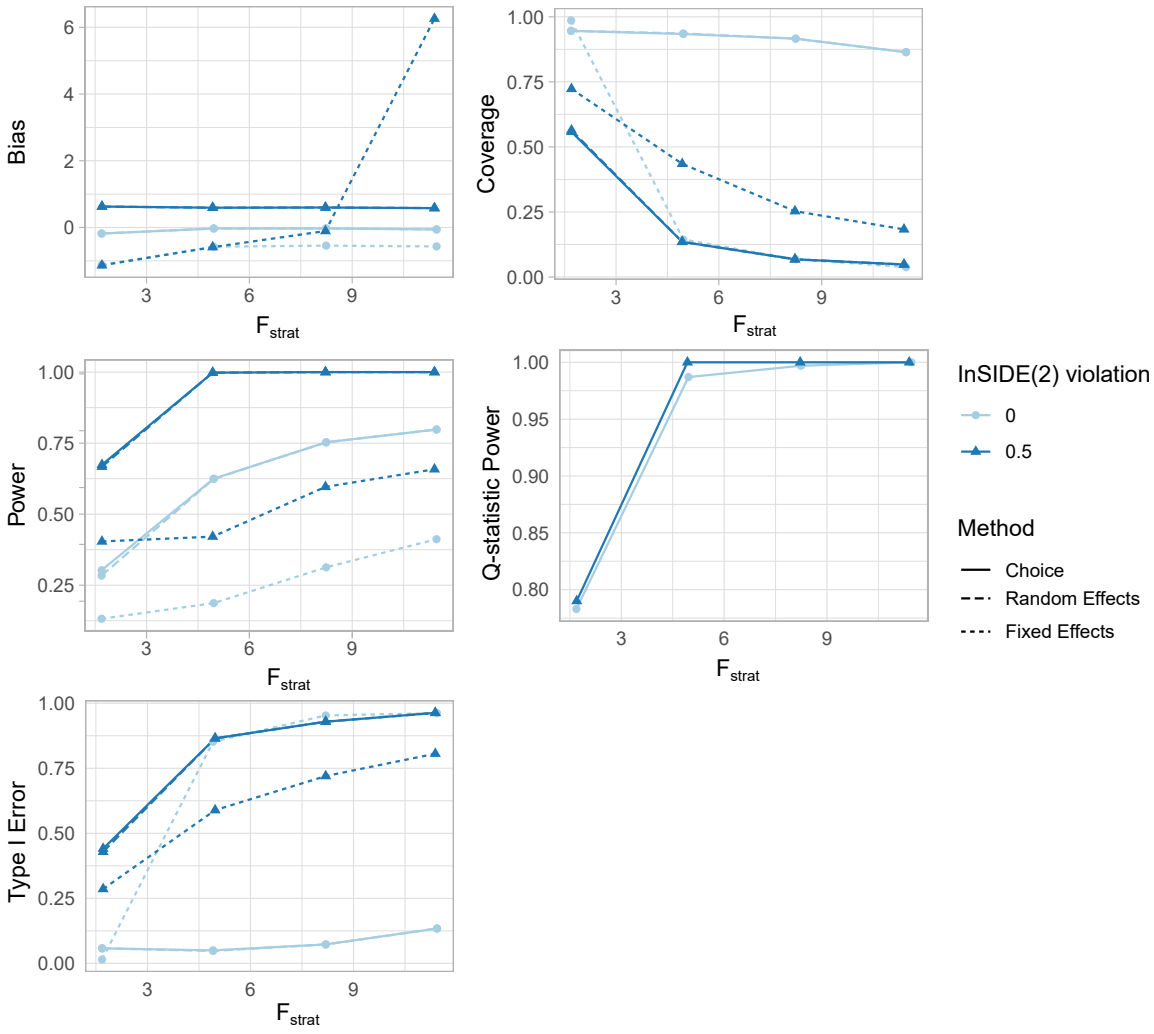


Figure 3.9: The figure presents the Bias, Power, Coverage, TIE, and Q-Statistic Power of the Collider-Corrected sex-stratified MR-RAPs estimator in the one sample setting. The performance measures are plotted against increasing sample sizes that correspond to mean (SD) F_{strat} statistics of 1.54 (0.28), 5.02 (0.69), 8.24 (0.97), and 11.13 (1.17). The data is generated with pleiotropy either satisfying the InSIDE(2) assumption ($\rho = 0$) or violating it ($\rho = 0.5$), and three different implementations of the approach are reported (Random Effects, fixed effect, and Choice). The TIE assessment is based on a 5% level test under the null ($\beta = 0$).

One-Sample MR or Two-Sample MR?

A practical question raised by a reviewer was to determine whether to implement a one-sample or two-sample sex-stratified analysis when individual-level data is available for N participants. The former approach is the one described in Section 3.3.3 and the latter splits the data into two independent subsets of size $\frac{N}{2}$. Figure 3.10

compares these two strategies. The results show that both approaches are approximately unbiased as F_{Strat} increases. However, the two-sample approach demonstrates lower power due to its less precise estimates.

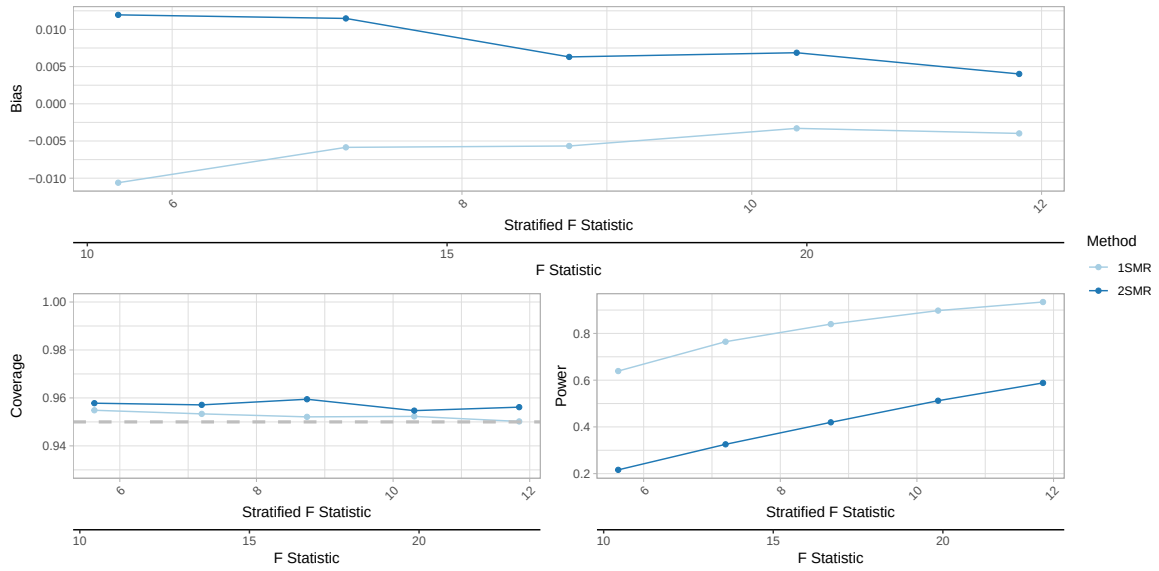


Figure 3.10: Bias, power and coverage, TIE in the one-sample (Sample size N) and two-sample ($\frac{N}{2}$ for exposure and outcome) settings. The performance measures are plotted against the increasing sample sizes that correspond to mean (SD) F_{strat} statistics of $5.62(0.61)$, $7.19(0.84)$, $8.73(0.98)$, $10.30(1.07)$, $11.84(1.22)$. The data are generated under the model of Figure 3.3, with pleiotropy satisfying InSIDE(2) assumption ($\rho = 0$).

3.4.3 Many Weak IVs or Few Strong and Weak Instrument Correction

Keeping the sample size and the proportion of variance in the exposure explained by genetics constant, we examined how the behaviour of the sex stratified MR RAPS estimator varied when many weakly interacting SNPs were available or if only a few strong ones were. The results are shown in Figures 3.11 and 3.12. In summary we show our method achieves appropriate coverage that is stable across the range of individual IV strengths. However, when more SNPs are used, power is in fact seen to increase.

To investigate the impact of a small number of arbitrarily weak IVs on the performance of the weak-instrument robust estimator, and to examine if an increasing

number of SNPs improves performance, we conducted a series of simulations by varying the number of SNPs to be 5, 12, 30, 50, or 100. Additionally, we varied the strength of the individual SNP-X associations while keeping the total variance in the exposure explained by the SNPs constant, in order to simulate a realistic scenario of either choosing few stronger IVs or many weaker ones. The sample size was kept fixed at 20,000.

Our results show that methods that leverage the interaction but are sensitive to weak instruments (MR GxE in one-sample MR, IVW$_{strat}$ in two-sample MR) suffer from regression dilution. This results in a bias towards the observational association for one-sample MR (as shown in Fig.3.12) and towards the null in two-sample MR (as shown in Fig.3.11), leading to a drop in coverage as the number of IVs increases, despite identical interaction strength ($F_{strat} \approx 16$ in all sets of IVs).

However, for the methods that leverage the interaction (RAPS$_{strat}$ in two-sample MR, RAPS$_{strat}$ + collider correction in one-sample MR), bias and coverage remained constant and well-controlled throughout while power actually improved as the number of IVs increased.

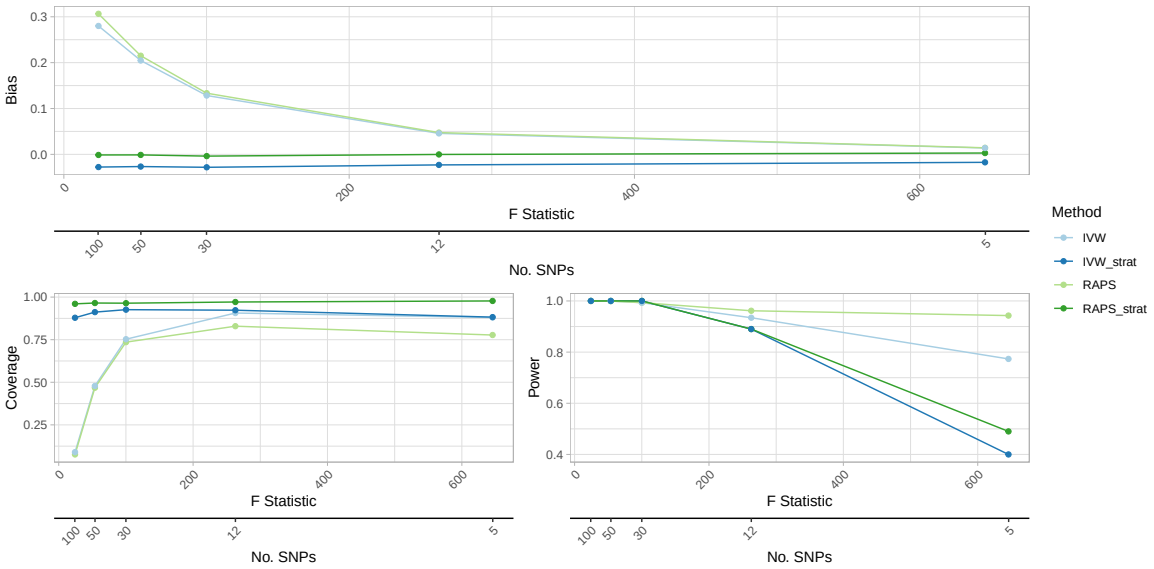


Figure 3.11: Comparison of performance in contexts of many weak IVs or few strong, Two-Sample MR

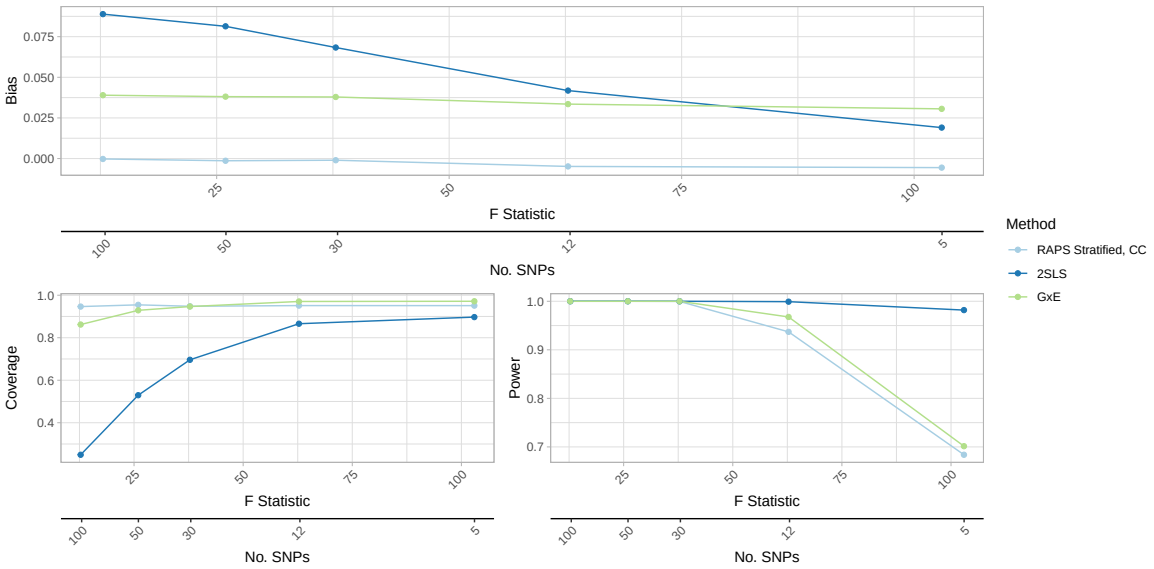


Figure 3.12: Comparison of performance in contexts of many weak IVs or few strong, One-Sample MR

3.5 Applied Examples

To illustrate the effectiveness of the sex-stratified MR-RAPS method, we conduct a series of single-sample MR studies using the UK Biobank data [87]. The goal is to

determine the causal effect of waist-to-hip ratio (WHR) on various binary and continuous outcomes linked to body composition or fat distribution. WHR is a measure of body shape and fat distribution, and it is particularly sensitive to central obesity, which is known to increase the risk of cardio-metabolic diseases. A 2015 study from the GIANT consortium revealed that 20 out of the 49 SNPs with genome-wide significance for WHR showed different effects in men and women [88]. A more recent meta-analysis, including data from UK Biobank, increased this number to 105 [89]. In accordance with the simulations, we employ three distinct MR methods for the data analysis.

1. Applying a conventional 2SLS approach by utilising all 49 SNPs as instruments;
2. Conducting a sex-stratified MR-RAPS analysis with all 49 SNPs, incorporating Collider Correction and determining the use of fixed or random effects analysis based on the Q_{cc} statistic;
3. As described in item 2, but limiting the analysis to only the 20 SNPs that have been found to be sexually dimorphic in the external GIANT cohort.

In the analysis, 17 outcomes (6 binary and 11 continuous) were evaluated. The models were adjusted for the following factors: age at baseline, first five genetic principal components to account for ancestry, center of assessment, and genotyping array. For continuous outcomes, the causal estimates represent the population average standardised effect of changing WHR by 1 standard deviation. For binary outcomes, the causal estimates show the average risk difference in the outcome for a 1 standard deviation increase in WHR, obtained by fitting logistic regression models and converting the model fitted values to average marginal effects using the 'margins()' package (refer to [90, 91] for more information). This approach differs from the more

commonly used logistic regression modeling. Our model depends on calculating precise differences in association estimates between men and women, and as the prevalence of some conditions can be very different across sexes (refer to Appendix, Table A3.1), the non-collapsibility of odds ratio could potentially introduce unwanted bias into the analysis (as discussed in [92] and [93]).

Scatter plots of the summary statistics used in our analyses can be found in Appendix A.1.2.

3.5.1 Strength of sex-differential associations with WHR

In the UK Biobank, the 49 genome-wide significant WHR SNPs from the GIANT consortium were analysed for all participants. The mean F-statistic was found to be 8.92 ($R^2 = 3.86\%$) in males and 146.23 ($R^2 = 5.37\%$) in females, indicating that the SNPs exhibiting sex-dimorphic differences in the GIANT cohort also display differential associations across genders in UK Biobank. The mean F_{strat}-statistic for the 49 SNPs that were jointly genome-wide significant in both males and females was 37.41, while the mean F_{strat}-statistic for the 20 SNPs that were genome-wide significant and sex-specific was higher at 69.63. The F and F_{strat} statistics for these 20 SNPs are depicted in Figure 3.13. As mentioned in Section 3.1, it was observed that for 10 SNPs, the F_{strat} was greater than F.

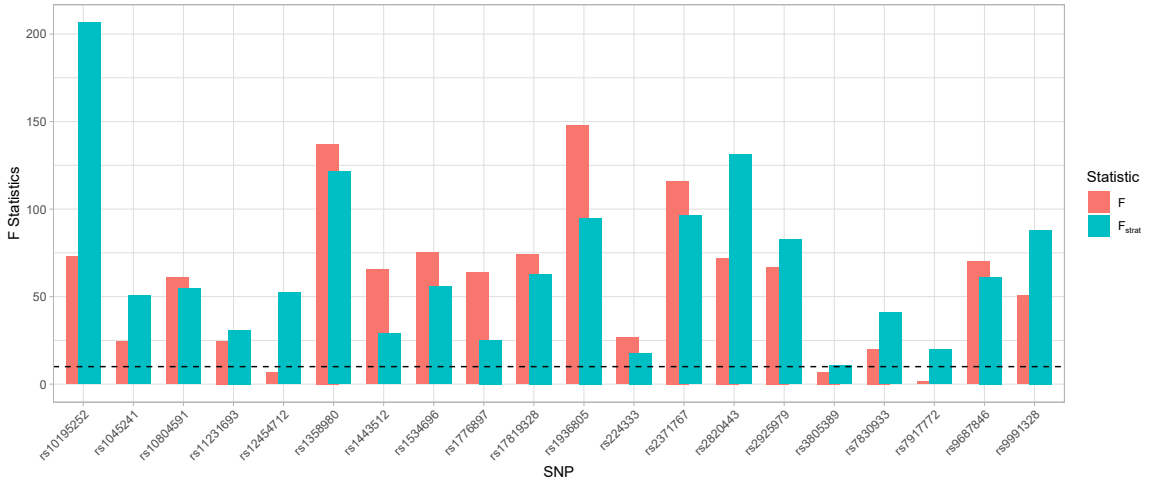


Figure 3.13: F-Statistics for SNP-WHR Associations in Women and Men within UK Biobank. The selection of sex-dimorphic SNPs was based on an external sample (Shungin et al. 2015). The dashed line represents the conventional threshold of 10, while F_{Strat} represents the F-statistic for the sex interaction term.

3.5.2 Results

In Figure 3.14, the results for the six binary outcomes are presented. These were: type 2 diabetes, myocardial infarction, osteoarthritis of any location, stroke, atrial fibrillation and any heart valve disease. Outcome prevalences for males and females are reported in Table 3.1. There seems to be some sex difference for the outcomes, with AF, T2D, MI and stroke being more common in males and osteoarthritis in females.

	Number & Prevalence in Males	**Number & Prevalence in Females**
Atrial Fibrillation	2587, 1.25%	1158, 0.47%
T2D	9462, 4.7%	4907, 2.04%
MI	8512, 4.13%	2077, 0.85%
Osteoarthritis	32913, 15.96%	48201, 19.69 %
Stroke	3828, 1.86%	2660, 1.09 %

Table 3.1: Cases & Proportion of diagnoses in males and females.

In terms of the effects of waist-hip ratio (WHR) on atrial fibrillation, osteoarthritis, and stroke, no causal relationship was detected across the three methods (1-3). For type 2 diabetes and myocardial infarction, the two-stage least squares (2SLS)

method indicated a robust causal effect with effect sizes of 0.038 (95% CI: 0.031, 0.044) and 0.016 (0.010, 0.022), respectively. This association was not found through the stratified approaches. Conversely, for heart valve disease, the results from the sex-stratified approaches indicated a protective effect of WHR on the outcome, whereas the 2SLS estimate did not show the same relationship.

It was found that for the method 2 in atrial fibrillation, a random effects model was deemed more appropriate based on the Q_{CC} statistics. A fixed effect model was preferred for all other estimates.

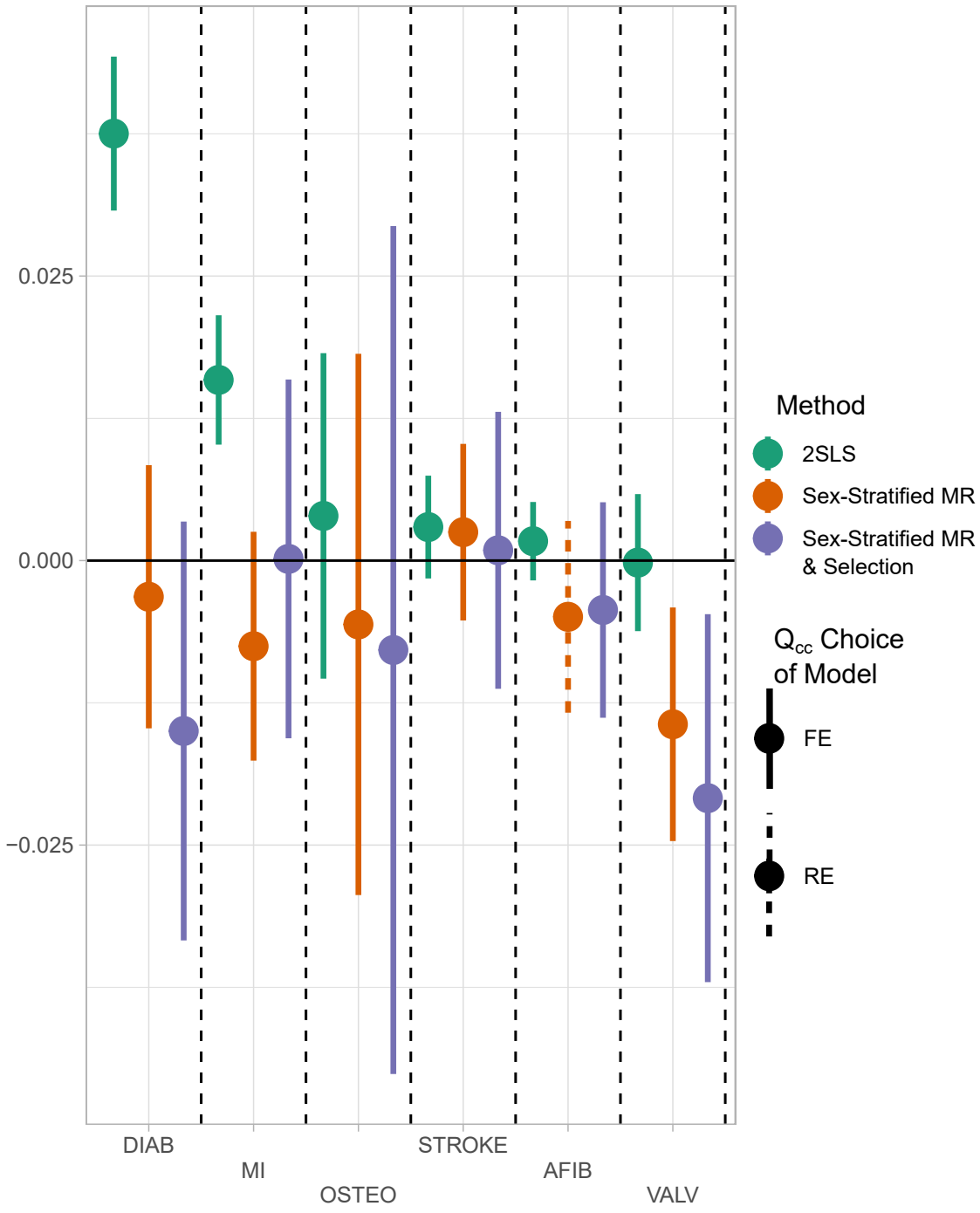


Figure 3.14: One-Sample binary outcome MR results for 2SLS estimate (analysis 1) and the sex stratified MR-RAPS approaches (analyses 2 and 3). Estimates are ordered by the magnitude of the 2SLS estimates. In the horizontal axis, different outcomes are represented; in the vertical axis, the effect size is shown as a point with the lines denoting the 95% confidence intervals (CIs). AFIB: atrial fibrillation; DIAB: type 2 diabetes; MI: myocardial infarction; OSTEO: osteoarthritis.

The results section presents the findings of the analysis for 11 continuous outcomes, namely height, high density lipoprotein (HDL), body mass index (BMI), C-

reactive protein (CRP), alcohol consumption (ALC), patient health questionnaire-9 (PHQ9), composite international diagnostic interview for depression (CIDI-MDD), low density lipoprotein (LDL), glucose (GLC), systolic blood pressure (SBP) and diastolic blood pressure (DBP), as shown in Figure 3.15. Among these outcomes, estimates from all methods fail to reject the null hypothesis at the 5% significance level for four outcomes (CRP, ALC, CIDI, PHQ-9). For height, BMI, and GLC, the 2SLS approach suggests an association of WHR with these phenotypes, but this finding is not supported by the stratified analyses.

A causal effect of WHR is implied by all three methods for SBP, DBP, HDL and LDL, although the magnitude of the effect is lower in the stratified approaches. Statistically significant heterogeneity was detected in the Q_{CC}-statistic at the 5% level for height, HDL, glucose, SBP and DBP for methods 2 and 3, while for CRP, heterogeneity was detected for method 2 but not for method 3. No heterogeneity was identified in all other outcomes, and a fixed effect model was used for these cases. Overall, the results suggest varying effects of WHR on different outcomes and demonstrate the importance of considering such stratified analyses that are naturally robust to pleiotropy to understand the underlying mechanisms.

Depression Outcomes: In keeping with the applied scope of the thesis, I investigated the effects of WHR on two depression questionnaire outcomes, the Patient Health Questionnaire-9 (PHQ-9) and the Composite International Diagnostic Interview (CIDI) (Figure 3.15). These are described in detail in Chapter 5. As above, the novel method and 2SLS were applied. There was no evidence of heterogeneity (Q) in the two models, suggesting consistent effects across individual SNPs. Both the conventional 2SLS analysis and the one-sample stratified method failed to reject the

null, with slightly less precise estimates in the latter.

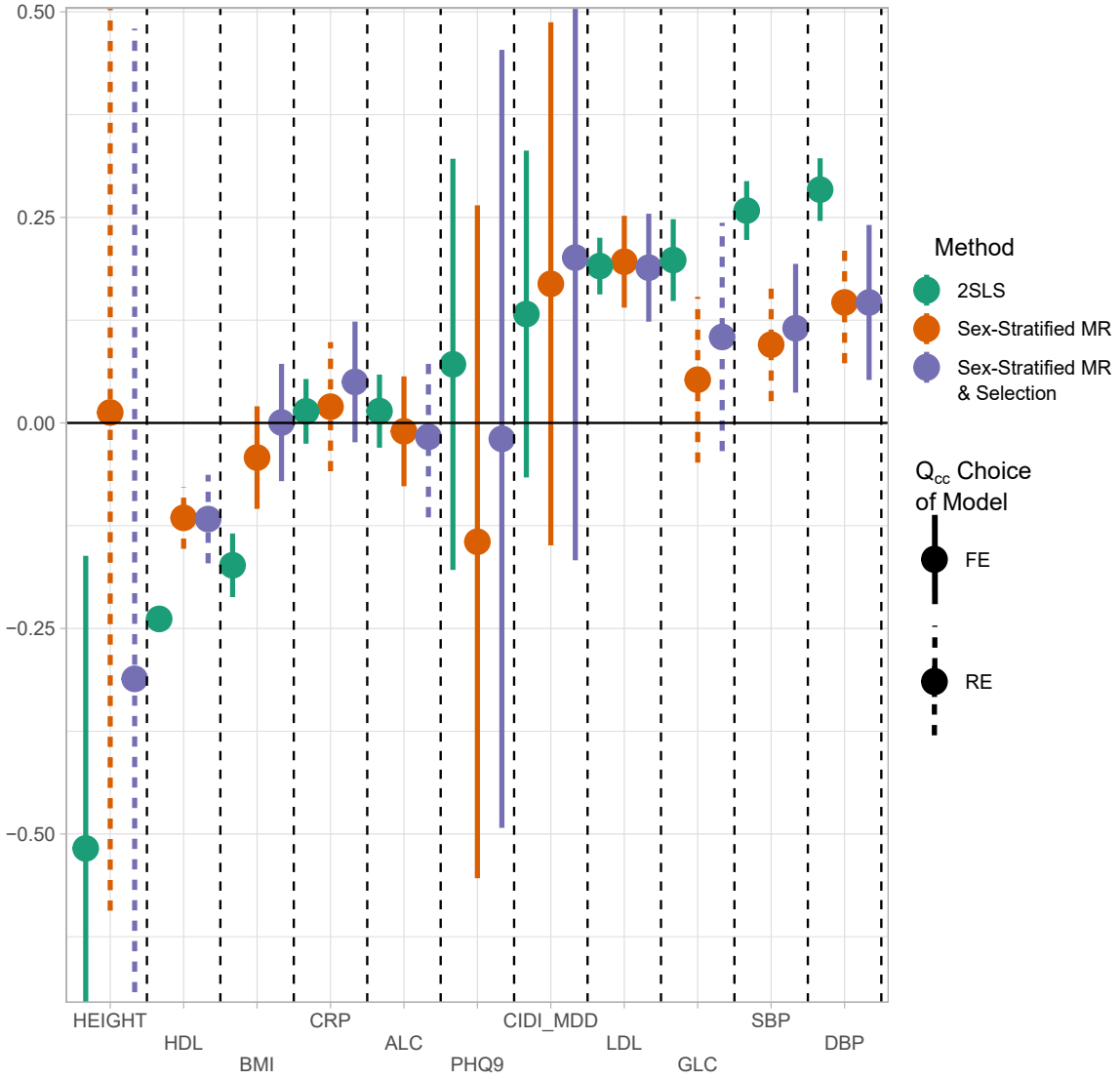


Figure 3.15: One-Sample continuous outcome MR results for 2SLS estimate (analysis 1) and the sex stratified MR-RAPS approaches (analyses 2 and 3) ALC: weekly alcohol consumption; BMI: body-mass index, inverse normalised; CIDI_MDD: Composite International Diagnostic Interview for depression; CRP: C-reactive protein; DBP: diastolic blood pressure; GLC: glucose; HDL: high-density lipoprotein; LDL: low-density lipoprotein; PHQ9: patient health questionnaire-9 (depression module); VALV: valvular disease.

3.6 Discussion

The current study demonstrates that horizontal pleiotropy, a common challenge in MR studies, can significantly bias the analysis and result in hopelessly biased esti-

mates. This is consistent with the findings of previous studies [94]. To address this issue, we have proposed a simple yet effective approach, building on the work of Spiller et al [75, 74], which utilises multiple gene-sex interactions to cancel out pleiotropy at the level of each individual instrument. This approach is robust to pleiotropy that violates the traditional InSIDE assumption [80].

However, we also note that a type of sex-specific pleiotropy, referred to as InSIDE 2 violation, could bias the results. This can occur when a polymorphism is differentially associated with both the exposure and outcome in men and women, and the two interactions are not independent. We propose ways to test for such sex-specific pleiotropy, which we argue is a more lenient assumption than those of other MR approaches.

Our approach can be viewed as a special case of MR-GxE [75] due to its use of a single binary covariate. However, by integrating the approach into the framework of MR-RAPS [86] and utilising the technique of Collider-Correction [85], our implementation is statistically efficient within a special class of estimators, weak instrument robust, and applicable to both one-sample and two-sample data. The approach can be implemented using standard existing software, and the software code is available at `https://github.com/vaskarageorg/stratMR`.

Our method provides a practical solution for handling pleiotropy in MR studies that investigate exposures exhibiting genetic heterogeneity between sexes. However, additional research is required to evaluate the generalisability and reliability of our approach, especially when faced with more intricate types of pleiotropy.

In our study, we compared the performance of our proposed sex-stratified MR approach to standard MR approaches in the analysis of the link between waist-to-hip

ratio (WHR) and various disease traits. Our results indicate that the standard MR approach leads to a larger number of tentative causal effects than the sex-stratified MR approach, which guards against pleiotropic bias. We believe that the lack of persistence of these effects in the sex-stratified MR analysis is mainly driven by pleiotropy involving body mass index (BMI). Variants associated with WHR also affect BMI [89], which in turn predisposes individuals to dysregulations of glucose metabolism [95] and cardiovascular risk [96]. We acknowledge that our approach has limitations, particularly in relation to the availability of publicly available sex-stratified genetic associations and the need for instrument selection in an external dataset to avoid Winner's curse. In addition, our approach assumes the homogeneity of causal effects of the exposure on the outcome across sexes, although we plan to adapt the approach to relax this assumption in future work. We recommend the use of a heterogeneity statistic to choose between two different implementations of the model, which is more efficient than a blanket use of the random effects model. Finally, we plan to extend our approach to allow for in-sample SNP selection, through implementing an explicit Winner's curse adjustment.

Alternative modelling frameworks

While we maintain that our estimation method based on differences is statistically efficient because it eliminates pleiotropy and supports a fixed effect model, we acknowledge that there exist various other options for modeling sex-specific summary statistics when pleiotropy is present, as noted by a reviewer. One such alternative approach, as described using the same notation as before, would be to employ two distinct linear models for SNP-outcome associations in males and females sepa-

rately, as shown below:

$$\Gamma_{j0} = \beta\gamma_{j0} + \alpha_{j0}, \quad \Gamma_{j1} = \beta\gamma_{j1} + \alpha_{j1}, \quad \alpha_{j0}, \alpha_{j1} \sim N(0, \tau^2). \tag{3.14}$$

It is important to note that the two models share the same causal effect β, but their pleiotropic effects, which originate from the same normal distribution parameterised by τ^2, are different. Estimating β and τ^2 together in this random effects model would likely result in higher efficiency compared to our proposed fixed effect difference model. However, the random effects model requires the InSIDE assumption that the sex-specific SNP-exposure associations and pleiotropic effects are independent, whereas our difference-based model does not depend on this assumption. Further exploration of the advantages and disadvantages of each modeling approach in terms of efficiency and robustness is a topic for future research.

3.7 Summary

In this Chapter, we show how pleiotropy can render MR estimates biased, and how our method solves this issue when gene-sex interactions are available. The key technique is that we cancel out pleiotropy at the level of each individual instrument. We point to cases where sex-specific pleiotropy can still not salvage the analysis. Our method is statistically efficient, weak instrument robust, and applicable to both one-sample and two-sample data.

Chapter 4

Dimensionality Reduction Approaches in MVMR

In Chapter 2.5, I describe how MVMR, an extension of the basic univariable approach, can be used to disentangle complex causal mechanisms and illuminate mediating pathways when multiple, possibly correlated phenotypes affect a health outcome [97, 98]. This is especially relevant if we consider the pleiotropic nature of genetic variants (Chapter 2.4) and hence the possibility that most MR designs suffer from some form of IV3 violation (Chapter 1.5). By incorporating many exposures simultaneously, MVMR can accommodate some degree of measured pleiotropy through the other exposures, such as in the investigation into the effect of various lipid traits on coronary heart disease (CHD) risk [98].

A logical conclusion would be to drastically increase the number of exposures that are to be included in such a MVMR model in order to try to accomodate many possible pleiotropic pathways with an extended instrument. However, it is important to note that this may not always be the most effective approach due to the potential for

correlations between exposures. MVMR can model correlated exposures, but it may not perform optimally when there are many highly correlated exposures, and in turn their genetically proxied values aren't clearly separated. This is best understood as a problem of conditionally very weak instruments [99], and hence can only be avoided if the genetic instruments are strongly associated with each exposure *conditionally* on all the other included exposures. To evaluate if the assumption is met, the conditional F-statistic can be used, with a minimum value of 10 for all exposures being deemed sufficiently strong as a rule of thumb [99].

When analysing multiple highly correlated exposures such as metabolite data or imaging data, genetic instruments are more likely to become conditionally weak, which can lead to extreme bias and unreliable causal estimates. Weak-instrument robust MVMR methods can address this to some extent, but at the cost of reduced precision [100, 101]. Additionally, MVMR models assume the ability to intervene and change each individual exposure while holding the others fixed, which may not be practically achievable in high dimensional and highly correlated exposure settings. Correlated exposures in physiology are a result of functional connectedness, such as involvement in the same biochemical cascades. For instance, low-density lipoprotein, a protein that carries cholesterol from the liver to other organs and tissues, can be further subdivided in many subclasses according to size and density [102]. Due to their partially shared function, strong genetic determinants that clearly separate these traits are not likely to be found and other approaches may be useful.

The objective of this chapter is to assess dimensionality reduction techniques and how they summarise groups of correlated genetically predicted exposures into more compact sets of principal components (PCs). We subsequently investigate MR anal-

yses with these PCs in place of the original exposures. We propose the adoption of sparse methods to improve inference and interpretability in the resulting factors.

I initiated work on this project during my MSc in Health Data Analytics & Machine Learning at Imperial College London, which has since been substantially developed and elevated to a publishable standard during my PhD studies. Parts of this chapter have been published in *elife* [103] and presented as oral presentations in two conferences [104, 105].

4.1 Multicollinearity & Dimensionality Reduction

In statistical analysis, multicollinearity is a common issue that arises when a matrix of exposures **X** is high-dimensional and contains several blocks of correlated vectors. If we attempt to regress this entire matrix on a response variable **Y**, this can lead to precision issues in the estimates due to the unstable nature of the parameter estimates.

Let $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p$ be the columns of the $n \times p$ matrix **X**, where n is the sample size and p is the number of exposures. Multicollinearity occurs when there is a high correlation between two or more of the exposure variables. In other words, some of the vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p$ can be expressed as linear combinations of other vectors, making it difficult to uniquely identify the effects of each exposure on the response variable.

Dimensionality reduction techniques are popular for dealing with these types of issues. One such technique is the singular value decomposition (SVD) of a matrix **X**. SVD includes a decomposition into three matrices as such: $\mathbf{X} = \mathbf{UDV}^T$, where **U** and **V** are orthogonal matrices and **D** is a diagonal matrix with singular values [106].

The matrix **V**, also known as the loadings matrix, is a useful indicator of which ex-

posures contribute to each principal component. The elements of each column represent the particular linear transformation that creates the corresponding PC, with highly correlated variables receiving coefficients of similar magnitude. By examining the loadings, we can identify which exposures are primarily driving the values in the PC scores, with numerically larger loadings showing a larger contribution to the linear transformation. Using this decomposition, we can obtain a smaller set of independent principal components that can be used for regression analysis. Let $\mathbf{Z} = \mathbf{UD}$ denote the $n \times m$ matrix of the first m principal components, where $m < p$. We can perform regression analysis on **Z** rather than **X** to estimate the effect of the exposures on the response variable **Y**. This technique is called PC regression, allows for inference and is useful in exploratory data analyses, [106]. Given the focus of this chapter in estimating causal effects, we make use of this method in the proposed approach.

An important part of applied PCA is the determination of a sensible number of informative PCs that should be retained. Having addressed the issue of multicollinearity, there still is the issue of keeping only relevant PCs while at the same time substantially reducing the number compared with the original exposures. There are various such methods for selecting the number of components. The elbow method involves selecting this number at the point where the marginal gain in explained variance begins to flatten. Due to its reliance on subjective visual judgement, this method is not ideal and can be prone to error. Another approach is to use a simple cutoff value for the eigenvalue of the principal components. This is typically set to 1, with this value expressing the amount of variance explained by a single variable, and are therefore considered non-informative. The permutation method is another approach that in-

volves randomly permuting X and then conducting PCA on the permuted data. This process is repeated multiple times, and the components that explain more variance than those obtained in the permuted data are considered informative.

One attractive property of using principal components is that the scores **UD** are independent with each other. This is because the remaining principal components are constructed directly orthogonal to the previous ones. Therefore, in the second stage of PC regression, we can perform inference on groups of variables rather than individual elements of **X**, without the concerns of multicollinearity for the predictor variables.

An active area of research is how sparse methods, that is methods that favour a solution where the loadings matrix V is constrained to contain many zero values, can be combined with dimensionality reduction so as to enhance interpretability [107]. By reducing the number of exposures that contribute to each principal component, we can isolate the contributions of groups of correlated exposures to specific components. This approach is particularly useful for making statistical inference in such contexts, as it allows us to more readily identify which exposures drive observed associations. In contrast, in PCA without sparsity constraints, all elements of X contribute to some degree to all components, making it difficult to directly map PCs to groups of exposures.

The problem of multicollinearity can lead to unstable parameter estimates when regressing a high-dimensional matrix of exposures on a response variable. This also applied to multivariable MR [98, 100]. In the following sections, we present investigations on dimensionality reduction techniques, such as PCA and sparse methods, for addressing this issue.

4.2 Motivation, Data Generating Mechanisms

In the context of high-throughput experiments, we consider the case of a data generating mechanism that we believe reflects common scenarios found in real-world applications. Specifically, we consider a set of exposures X, which can be partitioned into blocks based on shared genetic background. Certain groups of genes may contribute exclusively to specific blocks of exposures, while having no effect on other blocks. This in turn leads to correlations among the exposures of a certain block but no substantial correlation of exposures across blocks, only that attributable to the common confounder U. For the interests of MR inference and the satisfaction of the IV assumptions (Chapter 1.5), we can observe the potential complications in instrument strength (IV1) in Figure 4.1. A precise instrument for X_1 conditioning only on X_2 is lacking and only the groups of exposures are separated by leveraging the genetic information, rather than the individual exposures. The data set consists of n participants, k exposures, p SNPs, with both k and p consisting of b discrete blocks, and a continuous outcome Y.

In notation, we assume the data generating model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$ and $\mathbf{X} = \mathbf{G}\boldsymbol{\gamma} + \mathbf{U}$. The structure of the β vector is sparse (that is, few of the exposures contribute to Y) and the structure of the γ matrix is such that certain groups of variants G_b give rise to the correlated exposures of X_b. For simplicity, we assume no pleiotropic contributions.

4.2.1 Dimension reduction via PCA

As described in Section 4.1, PCA can be used to decompose the above described matrix X. We have two options depending on data availability. If we have individual

level data on X and G, we can genetically proxy X and perform PCA on $\hat{X}$. If, on the other hand, we have only summary statistics of the G-X associations (Equations 2.3, 2.4), one option would be to decompose this $p \times K$ matrix of SNP-exposure associations $\hat{\gamma}$ instead as follows:

$$\hat{\boldsymbol{\gamma}} = UDV^T,$$

where U and V are orthogonal matrices and D is a square matrix whose diagonal values are the variances explained by each component and all off diagonal values are 0. Again, as in the individual level data instance described in 4.1, V is the loadings matrix and serves as an indicator of the contribution of each metabolite to the transformed space of the PCs. The matrix UD (PCs/ scores matrix) is used in the second-step IVW regression in place of $\hat{\gamma}$. As V estimation does not aim for sparsity, all exposures will contribute to some degree to all components, making the interpretation more complicated. Therefore, we assessed multiple sparse PCA methods that intentionally limit this.

$\mathbf{y} = \hat{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{u}}$ where $\hat{\mathbf{X}} = \mathbf{G}\hat{\gamma}$. PCA on $\hat{\mathbf{X}}$ is approximately equivalent to PCA on $\hat{\gamma}$ since $\hat{\mathbf{X}}^T\hat{\mathbf{X}} = \hat{\gamma}^T\hat{\gamma}$ if $\mathbf{G}$ is normalised so that $\hat{\gamma}$ represent standardised effects. In the appendix we provide further simulation results that show that the loadings matrix derived from a PCA on $\hat{\mathbf{X}}$ and $\hat{\gamma}$ are asymptotically equivalent.

Sparse PCA (sPCA Zou et al.): Sparse PCA by Zou et al. [107] estimates the loadings matrix through an iterative procedure that progressively penalises exposures so that they do not contribute to certain PCs. In principle, this leads to a more clear picture for the consistency of each PC. This is performed as follows

1. Setting a fixed matrix, the following elastic net problem is solved

$\xi_j = argmin_\xi(\alpha_j - \xi)^T\hat{\gamma}^T\hat{\gamma}(\alpha_j - \xi) + \lambda_{1j}\|\xi\| + \lambda\|\xi\|^2$, where j is the PC;

2. For a fixed Ξ, $\hat{\gamma^T}\hat{\gamma}\Xi = UDV^T$ is estimated and update $A = UV^T$;

3. Repeat steps 1 and 2 until convergence.

Here λ_1 is an $L1$ sparsity parameter that induces sparsity, λ_2 is an $L2$ parameter that offers numerical stability, and Ξ is a matrix with sparsity constraints for each exposure [108]. As a result of the additional $\lambda_1\|\xi\|$ norm, there is sparsity in the loadings matrix and only some of the SNP-exposure associations $\hat{\gamma}$ contribute to each PC, specifically a particular subset of highly correlated exposures in $\hat{\gamma}$.

RSPCA: This approach differs in that it employs a robust measure of dispersion that is not unduly influenced by large single values of $\hat{\gamma}$ that contribute a large amount to the total variance. [109, 110]. As above, an L_1 norm is used to induce sparsity. For optimisation, the Tradeoff Product Optimisation (TPO) is maximised. It does not impose a single λ value on all PCs, thus allowing different degrees of sparsity.

Sparse Fused Principal Component Analysis (SFPCA)[111]: A method that can in theory exploit distinct correlation structures. Its goal is to derive a loadings matrix in which highly positively correlated variables are similar in sign and highly negative ones are opposite. Similar magnitudes also tend to be obtained for those variables that are in the same blocks in the correlation matrix. Like the Sparce PCA optimisation in Zou et al. [107], SFPCA works by assigning highly correlated variables the exact same loadings as opposed to numerically similar ones (Figure 4.4d). This is achieved with two norms in the objective function: λ_1 which regulates the L_1 norm that induces sparsity and λ_2 for the $L2$ regularisation (squared magnitude of $\hat{\gamma}$) to

guard against singular solutions. A grid search is used to identify appropriate parameters for λ_1 and λ_2. The following criterion is used

$$min_{A,\Xi}\|\hat{\gamma} - \hat{\gamma}\Xi A^T\|_F + \lambda_1\|\xi\| + \lambda_2|\rho_{s,t}||\xi_{s,t} - sign(\rho_{s,t}\xi_{t,k})|,$$

such that $A^T A = I_K$. The 'fused' penalty (last term) purposely penalises discordant loadings for variables that are highly correlated. The choice of the sparsity parameters is based on a BIC criterion.

Sparse Component Analysis (SCA): SCA [112] is motivated by the relative inadequacy of the classic approaches in promoting significant sparsity. It addresses this by rotating the eigenvectors to achieve approximate sparsity whilst keeping the proportion of variance explained the same. Simulation studies show the technique works especially well in high dimensional settings such as gene expression data, among other examples [112].

4.2.2 Choice of Components

In all dimensionality reduction approaches applied to correlated variables, there is no upper limit to how many transformed factors can be estimated. However, only a proportion of them are likely to be informative in the sense of collectively explaining a meaningful amount of total variance in the original data set. To guide this choice, a permutation-based approach was implemented [113] as follows: Firstly, the $\hat{\gamma}$ matrix was randomly permuted and the sparse PCA method of interest was applied on the permuted set. The computed eigenvalues are assumed to come from a null distribution consistent with a non-informative component. This process is repeated multiple

times (e.g. $perm = 1000$) and the mean eigenvalues for all components stored. Finally, the sparse PCA method is performed in the original $\hat{\gamma}$ matrix and whichever component has an eigenvalue larger than the mean of the permuted sets is considered informative and kept. Due to the computational complexity of the permutation method, particularly for SFPCA, an alternate method - the Karlis-Saporta-Spinakis (KSS) criterion [114] - was also used. This is based on a simple correction on the minimum non-trivial eigenvalue (Cutoff$_{KSS} = 1+2\sqrt{\frac{K-1}{p-1}}$). The authors show that the method is robust to non-normal distributions [114]. Although KSS was not compared with the above described permutation approach, it performed better than simpler approaches, such as choosing those PCs whose eigenvalue is larger than 1 (Kaiser criterion), the broken stick method [115] and the Velicer method [116].

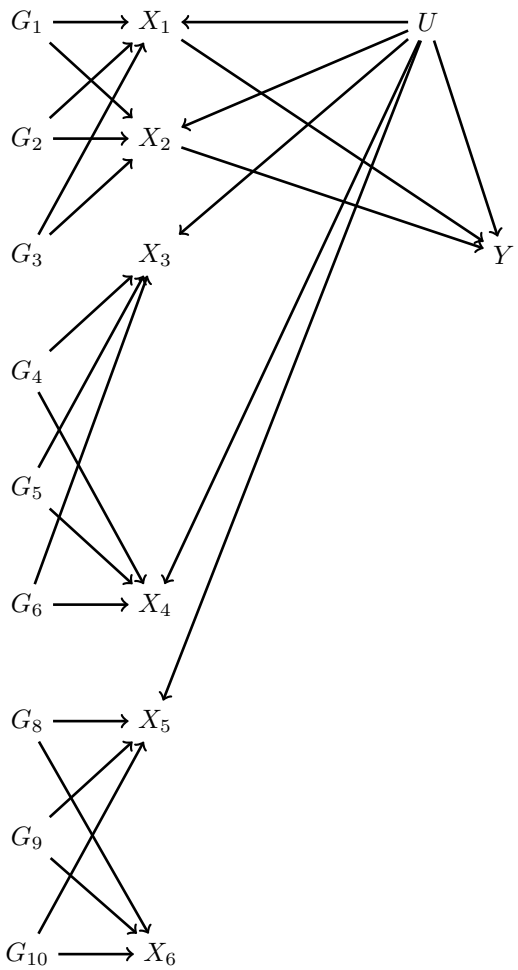


Figure 4.1: Directed acyclic graph (DAG) depicting a scenario with multiple variants (G_1-G_{10}) affecting multiple exposures X_1 to X_6, and X_1 and X_2 in turn affect the outcome variable Y. The shared genetic background of X_{1-2}, X_{3-4}, X_{5-6}. There are three distinct blocks of exposures. A confounding variable U affects the exposures and the outcome.

4.2.3 Instrument Strength of PCs

In MVMR, the $IV1$ assumption requires a set of genetic variants that robustly associate with at least one of the exposures X (Chapter 1.5). This is quantified by CFS [100]. With summary statistics of the SNP-X associations $\hat{\gamma}_{p,k}$ (p: SNP, k: exposure), the mean $F-$statistic for exposure k used in a standard UVMR analysis is the far simpler expression

$$F_k = \frac{\sum_{j=1}^{p} \left(\frac{\hat{\gamma}_{j,k}}{SE_{\hat{\gamma}_{j,k}}}\right)^2}{p}. \tag{4.1}$$

We provide a dedicated formula for estimating instrument strength measures for the F-statistic for the PCs that is closely related to Eq. 4.1 rather than the original expression. This simplification is due to the fact that an MVMR analysis of a set of PCs is essentially equivalent to a UMVR analysis of each exposure separately. The full derivation is reported in Section A.3.1 of the Appendix.

4.3 Results

Workflow Overview. Our proposed analysis strategy is presented in Figure 4.2. Using summary statistics for the single-nucleotide polymorphism (SNP)-exposure ($\hat{\gamma}$) and SNP-outcome ($\hat{\Gamma}$) association estimates, where $\hat{\gamma}$ (dimensionality 148 SNPs $\times 97$ exposures) exhibits strong correlation, we initially perform a principal component analysis (PCA) on $\hat{\gamma}$. Additionally, we perform multiple sparse PCA modalities that aim to provide sparse loadings that are more interpretable (block 3, Fig. 4.2). The choice of the number of principal components (PCs) is guided by permutation testing or an eigenvalue threshold. Finally, the PCs are used in place of $\hat{\gamma}$ in an IVW MVMR meta-analysis to obtain an estimate of the causal effect of the PC on the outcome. Similar to PC regression and in line with unsupervised methods, the outcome (SNP-outcome associations ($\hat{\Gamma}$) and corresponding standard error ($SE_{\hat{\Gamma}}$)) is not transformed by PCA and is used in the second-step MVMR in the original scale. In the real data application and in the simulation study, the best balance of sparsity and statistical power was observed for the method of sparse component analysis (SCA) [112]. This favoured method and the related steps are coded in an *R* function and are available at GitHub (https://github.com/vaskarageorg/SCA_MR/).

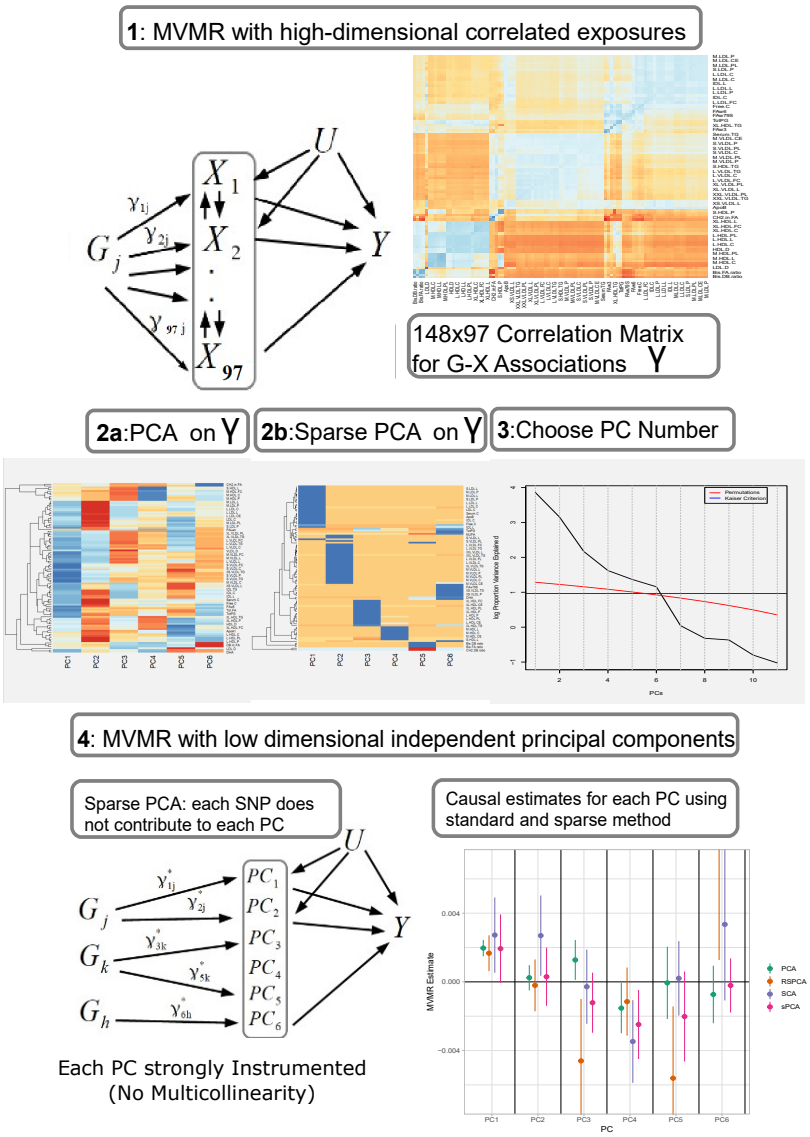


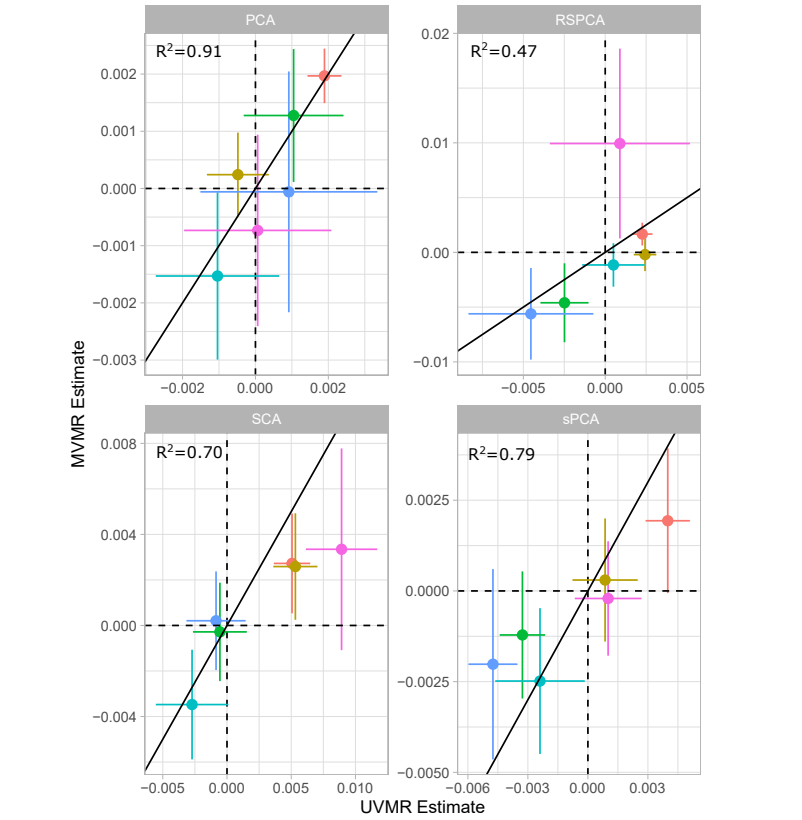
Figure 4.2: Proposed workflow. Step 1: MVMR on a set of highly correlated exposures. Each genetic variant contributes to each exposure. The high correlation is visualised in the similarity of the SNP-exposure associations in the correlation heatmap (top right). Step 2 and 3: PCA and sparse PCA on $\hat{\gamma}$. Step 4. MVMR analysis on a low dimensional set of PCs. X: exposures; Y: outcome; k: number of exposures; PCA: principal component analysis; MVMR: multivariable MR

Univariable MR (UVMR) & Multivariable MR (MVMR). A total of 66 traits were associated with CHD at or below the Bonferroni-corrected level ($p = 0.05/97$, Table 4.1). Two genetically-predicted lipid exposures (M.HDL.C, M.HDL.CE) were negatively associated with CHD and 64 were positively associated (Table 4.3). In a MVMR model including only the 66 Bonferroni-significant traits, fitted with the pur-

pose of illustrating the instability of IVW-MVMR in conditions of severe collinearity, conditional F-statistic (CFS) (*Methods*) was lower than 2.2 for all exposures (with a mean of 0.81), highlighting the severe weak instrument problem. In Figure A.17, the MVMR estimates are plotted against the corresponding UVMR estimates. We interpret the reduction in identified effects as a result of the drop in precision in the MVMR model (variance inflation). Only the independent causal estimate for ApoB reached our pre-defined significance threshold and was less precise (OR$_{MVMR}$ (95% CI): $1.031(1.012, 1.37)$, OR$_{UVMR}$ (95% CI): $1.013(1.01, 1.016)$ (Fig. A.18). We note that, for M.LDL.PL, the UVMR estimate ($1.52(1.35, 1.71)$, p $<10^{-10}$)) had an opposite sign to the MVMR estimate (OR$_{MVMR} = 0.905(0.818, 1.001)$).

	Positive	Negative
VLDL	M.VLDL.C, M.VLDL.CE, M.VLDL.FC, M.VLDL.L, M.VLDL.P, M.VLDL.PL, M.VLDL.TG, XL.VLDL.L, XL.VLDL.PL, XL.VLDL.TG, XS.VLDL.L, XS.VLDL.P, XS.VLDL.PL, XS.VLDL.TG, XXL.VLDL.L, XXL.VLDL.PL, L.VLDL.C, L.VLDL.CE, L.VLDL.FC, L.VLDL.L, L.VLDL.P, L.VLDL.PL, L.VLDL.TG, S.VLDL.C, S.VLDL.FC, S.VLDL.L, S.VLDL.P, S.VLDL.PL, S.VLDL.TG	None
LDL	LDL.C, L.LDL.C, L.LDL.CE, L.LDL.FC, L.LDL.L, L.LDL.P, L.LDL.PL, M.LDL.C, M.LDL.CE, M.LDL.L, M.LDL.P, M.LDL.PL, S.LDL.C, S.LDL.L, S.LDL.P	None
HDL	S.HDL.TG, XL.HDL.TG	M.HDL.C, M.HDL.CE

Table 4.1: Univariable MR results for the Kettunen dataset with CHD as the outcome. Positive: positive causal effect on CHD risk; Negative: negative causal effect on CHD risk.



	RSPCA	SFPCA	sPCA	SCA
PC1	LDL, VLDL	VLDL, LDL	LDL	LDL
PC2	VLDL	HDL, IDL, VLDL	Large HDL	VLDL
PC3	HDL	Small, Medium, Large HDL, IDL, VLDL	VLDL	Large HDL
PC4	Double bonds ratios	Medium HDL	Small, Medium HDL	Small, Medium HDL
PC5	Double bonds ratios	Double bonds ratios, Fatty acid ratios	Small, Medium VLDL	Double bonds ratios
PC6	Double bonds ratios	Double bonds ratios, Fatty acid ratios	Double bonds ratios	Small VLDL

Figure 4.3: Comparison of UVMR and MVMR estimates and presentation of the major group represented in each PC per method.

To see if the application of a weak-instrument robust MVMR method could improve the analysis, we applied MR GRAPPLE [101]. As the GRAPPLE pipeline suggests, the same three-sample MR design described above is employed. In the external selection GWAS study (GLGC), a total of 148 SNPs surpass the genome-wide significance level for the 97 exposures and were used as instruments. Although the method did not identify any of the exposures as statistically significant at nominal or Bonferroni-adjusted statistical significance level, the strongest association among all

exposures is ApoB.

PCA. Standard PCA with no sparsity constraints was used as a benchmark. PCA estimates a square loadings matrix of coefficients with dimension equal to the number of genetically proxied exposures K. The coefficients in the first column define the linear combination of exposures with the largest variability (PC1). Column 2 defines PC2, the linear combination of exposures with the largest variability that is also independent of PC1, and so on. This way, the resulting factors seek to reduce redundant information and project highly correlated SNP-exposure associations to the same PC. In PC1, VLDL- and LDL-related traits were the major contributors (Figure 4.4a). ApoB received the 8th largest loading (0.1371, maximum was 0.1403 for cholesterol content in small VLDL) and LDL.C received the 48th largest (0.1147). In PC2, HDL-related traits were predominant. The first 18 largest positive loadings are HDL-related and 12 describe either large or extra-large HDL traits. PC3 received its scores mainly from VLDL traits. Six components were deemed statistically significant through the permutation-based approach (Fig. 4.2, *Methods*).

In the second-step IVW regression (Step 4 in Fig. 4.2), MVMR results are presented. A modest yet precise (OR = $1.002(1.0015, 1.0024)$, $p < 10^{-10}$) association of PC1 with CHD was observed. Conversely, PC3 was marginally statistically significant for CHD at the 5% level (OR = 0.998 (0.998, 0.999), p = 0.049). Since $\hat{\gamma}$ has been transformed with linear coefficients (visualized in loadings matrix, Fig. 4.4), the underlying causal effects are also transformed and interpreting the magnitude of an effect estimate is not straightforward, since it reflects the effect of changing the PC by one unit on the outcome; however, significance and orientation of effects can be interpreted. When

positive loadings are applied to exposures that are positively associated with the outcome, the MR estimate is positive; conversely, if negative loadings are applied, the MR estimate is negative.

Sparse PCA methods. We next employed multiple sparse PCA methods (Table A.9) that each shrink a proportion of loadings to zero. The way this is achieved differs in each method. Their underlying assumptions and details on differences in optimisation are presented in Table A.9 and further described in Chapter 4.2.1.

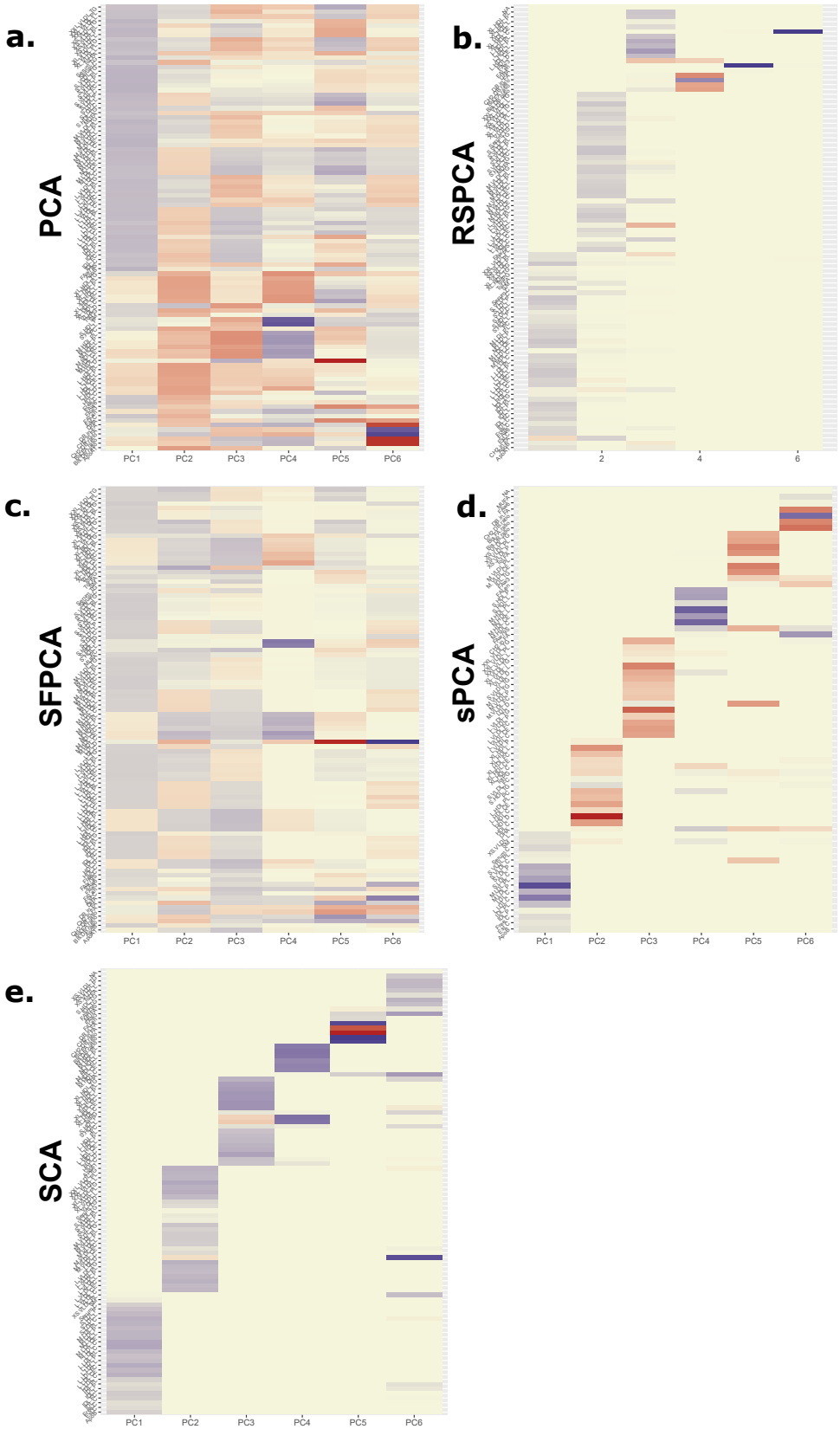


Figure 4.4: Heatmaps for the loadings matrices in the Kettunen dataset for all methods (one with no sparsity constraints (a), four with sparsity constraints under different assumptions (b-e)). The number of the exposures plotted on the vertical axis is smaller than $K = 97$ as the exposures that do not contribute to any of the sparse PCs have been left out. Blue: positive loading; Red: negative loading; Yellow: zero.

	PCA	RSPCA	SFPCA	sPCA	SCA
Overlap	1	0.938	1	0.187	0.196
Overlap in PC1,PC2	1	0.433	1	0.010	0
Sparse %	0	0.474	0.082	0.835	0.796
VLDL Significance in MR	Yes	No	Yes	No	Yes
LDL Significance in MR	No	Yes	No	No	Yes
HDL Significance in MR	Yes	Yes	Yes	No	No
Small, Medium HDL Significance in MR	Yes	No	Yes	Yes	Yes

Table 4.2: Results for PCA approaches. Overlap: Percentage of metabolites receiving non-zero loadings in ≥ 1 component. Overlap in PC1, PC2: overlap as above but exclusively for the first two components which by definition explain the largest proportion of variance. VLDL, LDL and HDL significance: results of the IVW regression model with CHD as the outcome for the respective sPC's (the sPC's that mostly received loadings from these groups). The terms VLDL and LDL refer to the respective transformed blocks of correlated exposures; for instance, VLDL refers to the weighted sum of the correlated VLDL-related $\hat{\gamma}$ associations, such as VLDL phospholipid content and VLDL triglyceride content. †: RSPCA projected VLDL- and LDL-related traits to the same PC (sPC1). ‡: SCA discriminated HDL molecules in 2 sPC's, one for traits of small- and medium-sized molecules and one for large- and extra-large-sized. †: significance is not directly reported in this model

RSPCA [117]. Optimisation and the KSS criterion pick six PCs to be informative [114]. The loadings in Figure 4.4 show a VLDL-, LDL-dominant PC1, with some small and medium HDL-related traits. LDL.C and ApoB received the 5th and 40th largest positive loadings. PCs 1 and 6 are positively associated with CHD and PCs 3 and 5 negatively so (Table A.10).

SFPCA [111]. The KSS criterion retains 6 PCs. The loadings matrix (Figure 4.4) shows the 'fused' loadings with the identical coloring. In the two first PCs, all groups are represented. Both ApoB and LDL.C received the seventh and tenth largest loadings, together with other metabolites (Figure 4.4). PC1 (all groups represented) was positively associated with CHD and PC4 (negative loadings from large HDL traits) negatively so (Table A.10).

Sparse PCA (sPCA [107]). The number of non-zero metabolites per PC was set at $\frac{148}{97} \sim 16$ (see Figure A.21). Under this level of sparsity, the permutation-based approach suggested that six sPC's should be retained. Seventy exposures received a zero loading across all components. PC1 is constructed predominantly from LDL

traits and is positively associated with CHD, but this does not retain statistical significance at the nominal level in MVMR analysis (Figure 4.3). Only PC4 that is comprised of small and medium HDL traits (Fig. 4.4b) appears to exert a negative causal effect on CHD (OR (95% CI): $0.9975(0.9955, 0.9995)$). The other PCs were not associated with CHD (all p values > 0.05, Table A.10).

Sparse Component Analysis (SCA) [112]. Six components were retained after a permutation test. In the final model, five metabolites were regularised to zero in all PCs (CH2.DB.ratio, CH2.in.FA, FAw6, S.VLDL.C, S.VLDL.FC, Figure 4.4). Little overlap is noted among the metabolites. PC1 receives loadings from LDL and IDL, and PC2 from VLDL. The contribution of HDL to PCs is split in two, with large and extra-large HDL traits contributing to PC3 and small and medium ones to PC4. PC1 and PC2 were positively associated with CHD (Table A.10, Figure 4.3). PC4 was negatively associated with CHD.

Comparison with Univariable MR. In principle, all PC methods derive independent components. This is strictly the case in standard PCA, where subsequent PCs are perfectly orthogonal, but is only approximately true in sparse implementations. We hypothesised that UVMR and MVMR could provide similar causal estimates of the associations of metabolite PCs with CHD. The results are presented in Figure 4.3 and concordance between UVMR and MVMR is quantified with the R^2 from a linear regression. The largest agreement of the causal estimates is observed in PCA. In the sparse methods, SCA [112] and sPCA [107] provide similarly consistent estimtates, whereas some disagreement is observed in the estimate of PC6 for RSPCA [117] on CHD.

A previous study implicated LDL.c and ApoB as causal for CHD [118]. In Figure

A.19, we present the loadings for these two exposures across the PCs for the various methods. Ideally, we would like to see metabolites contributing to a small number of components for the sparse methods. Using a visualisation technique proposed by Kim et al. [119], this is indeed observed (see Fig. A.19 in the Appendix). In PCA, LDL.c and ApoB contribute to multiple PCs, whereas the sparse PCA methods limit this to one PC. Only in RSPCA do these exposures contribute to two PCs. In the second-step IVW meta-analysis, it appears that the PCs comprising of predominantly VLDL/LDL and HDL traits robustly associate with CHD, with differences among methods (Table 4.2).

Instrument Strength. Instrument strength for the chosen PCs was assessed via an F-statistic, calculated using a bespoke formula that accounts for the PC process (Chapter A.3.1). The F-statistics for all transformed exposures cross the cutoff of 10. There was a trend for the first components being more strongly instrumented in all methods (see Figure A.20), which is to be expected. In the MVMR analyses, the CFS for all exposures was less than three. Thus the move to PC-based analysis substantially improved instrument strength and mitigated against weak instrument bias.

Simulation Studies. We consider the case of a data generating mechanism that reflects common scenarios found in real-world applications. Specifically, we consider a set of exposures X, which can be partitioned into blocks based on shared genetics. Certain groups of variants contribute exclusively to specific blocks of exposures, while having no effect on other blocks. This in turn leads to substantial correlation among the exposure blocks and a much reduced correlation of between exposure blocks, due only to shared confounding. This is visualised in Figure 4.5a. This data

structure acts to reduce the instruments' strength in jointly predicting all exposures. The data set consists of n participants, k exposures, p SNPs (with both k and p consisting of b discrete, equally sized blocks) and a continuous outcome, Y. We split the simulation results into one illustrative example (for didactic purposes) and one high-dimensional example.

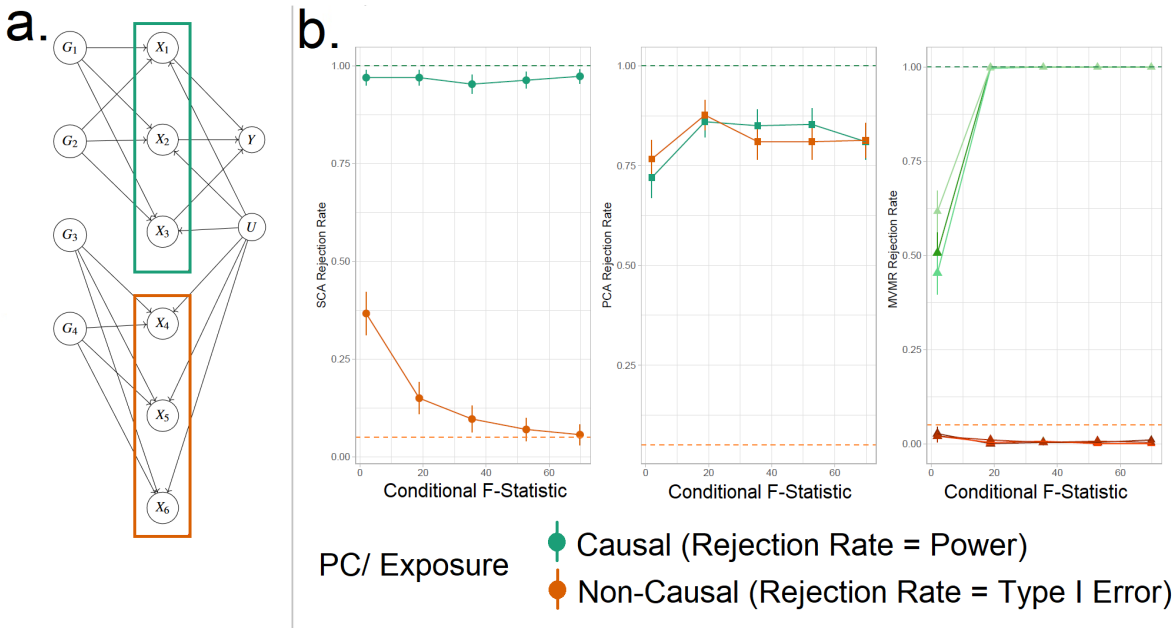


Figure 4.5: a. Data generating mechanism for the simulation study, illustrative scenario with six exposures and two blocks. In red boxes, the exposures that are correlated due to a shared genetic component are highlighted. b. Simulation results for six exposures and three methods (SCA [112], PCA, MVMR). The exposures that contribute to Y (X_{1-3}) are presented in shades of green colour and those that do not in shades of red (X_{4-6}). In the third panel, each exposure is a line. In the first and second panels, the PCs that correspond to these exposures are presented *as single lines* in green and red. Monte Carlo SEs are visualised as error bars. Rejection rate: proportion of simulations where the null is rejected.

Simple illustrative example: We generate data under the mechanism presented in Figure 4.5a. That is, with six individual exposures $X_1, ..., X_6$ split into two distinct blocks ($X_1 - X_3$ and $X_4 - X_6$). A continuous outcome Y is generated that is only causally affected by the exposures in block 1 ($X_1 - X_3$). A range of sample sizes were used in the simulation in order to give a range of conditional F statistic (CFS) values from approximately 2-80. We apply a) MVMR with the six individual expo-

sures separately , and b) PCA and SCA. The aim of approach b) is to demonstrate the impact of reducing the six-dimensional exposure into two PCs, so that the first PC has high loadings for block 1 ($X_1 - X_3$) and the second PC has high loadings for block 2 ($X_4 - X_6$). Although two PCs were chosen by both PCA methods using a KSS criterion in a large majority of cases, to simplify the simulation interpretation we fixed *a priori* the number of PCs at 2 across all simulations.

Our primary focus was to assess the rejection rates of MVMR versus PCA rather than estimation, as the two approaches are not comparable in this regard. To do this we treat each method as a test, which obtains true positive (TP), true negative (TN), false positive (FP), and false negative (FN) results. In MVMR, a TP is an exposure that is causal in the underlying model *and* whose causal estimate is deemed statistically significant. In the PCA and sparse PCA methods, this classification is determined with respect to a) which exposure(s) determine each PC and b) if the causal estimate of this PC is statistically significant. Exposures are considered to be *major contributors* to a PC if (and only if) their individual PC loading is larger than the average loading. If the causal effect estimate of a PC in the analysis deemed statistically significant, major contributors that are causal and non-causal are counted as TPs and FPs respectively. TNs and FNs are defined similarly. Type I error therefore corresponds to the false positive rate and power corresponds to the true positive rate. All statistical tests were conducted at the α/B= $\alpha/2$ = 0.025 level.

SCA, PCA and MVMR type I error and power are shown in the three panels (left to right) in Fig. 4.5b) respectively. These results suggest an improved power in iden-

tifying true causal effects both with PCA and SCA compared with MVMR when the CFS is weak, albeit at the cost of an inflated type I error rate. As sample size and CFS increase, MVMR performs better. For the PC of the second block's null exposures, PCA seems to have a sub-optimal type I Error control (red in Figure 4.5b). In this low dimensional setting, the benefit of PCA therefore appears to be limited.

Complex high-dimensional example: The aim of the high-dimensional simulation is to estimate the comparative performance of the methods in a wider setting that more closely resembles real data applications. We simulate genetic data and individual level exposure and outcome data for between $K = 30 - 60$ exposures, arranged in $B = 4 - 6$ blocks. The underlying data generating mechanism and the process of evaluating method performance is identical to the illustrative example, but the number of variants, exposures and the blocks are increased. We amalgamate rejection rate results across all simulations, by calculating sensitivity (SNS) and specificity (SPC) as:

$$SNS = \frac{TP}{TP + FN} \quad SPC = \frac{TN}{TN + FP}, \tag{4.2}$$

and then compare all methods by their area under the estimated receiver-operating characteristic (ROC) curve (AUC) using the meta-analytical approach of Reitsma et al [120]. Briefly, the Reitsma method performs a bivariate meta-analysis of multiple studies that report both sensitivity and specificity of a diagnostic test, in order to provide a summary ROC curve. A bivariate model is required because sensitivity and specificity estimates are correlated. In our setting the 'studies' represent the results of different simulation settings with distinct numbers of exposures and blocks.

Youden's index J ($J = SNS + SPC - 1$) was also calculated, with high values being indicative of good performance.

Two sparse PCA methods (SCA [112], sPCA [107]) consistently achieve the highest AUC (Fig. A.22). This advantage is mainly driven by an increase in sensitivity for both these methods compared with MVMR. A closer look at the individual simulation results corroborates the discriminatory ability of these two methods, as they consistently achieve high sensitivities (Figure A.24). Both standard and Bonferroni-corrected MVMR performed poorly in terms of AUC (AUC 0.712 and 0.660 respectively), due to poor sensitivity. PCA performed poorly, with almost equal true and false positive results (AUC 0.560). PCA and RSPCA did not accurately identify negative results (PCA and RSPCA median specificity 0 and 0.192 respectively). This extreme result can be understood by looking at the individual simulation results in Figure A.24; both PCA and RSPCA cluster to the upper right end of the plot, suggesting a consistently low performance in identifying true negative exposures. Specifically, the estimates with both these methods were very precise across simulations and this resulted in many false positive results and low specificity. We note a differing performance among the top ranking methods (SCA, sPCA); while both methods are on average similar, the results of SCA are more variable in both sensitivity and specificity (Table A.11). The Youden indexes for these methods are also the highest (Fig. A.22a). Varying the sample sizes (mean instrument strength in $\hat{\gamma}$ from $\bar{F} = 221$ to 1109 and mean conditional F statistic $C\bar{F}S = 0.34 - 12.81$ (Figure A.23) suggests a similar benefit for sparse methods.

Even with large sample sizes ($\bar{F} = 1109.78$, $C\bar{F}S = 12.82$), MVMR can still not dis-

criminate between positive and negative exposures as robustly as the sparse PCA methods. A major determinant of the accuracy of these methods appears to be the number of truly causal exposures, as in a repeat simulation with only four of the exposures being causal, there was a drop in sensitivity and specificity across all methods. Sparse PCA methods still outperformed other methods in this case, however (Table A.12).

What determines PCA performance?: In the hypothetical example of Figure 4.5 and indeed any other example, if two PCs are constructed, PCA cannot differentiate between causal and non-causal exposures. The only information used in this stage of the workflow (Steps 2 and 3 in Figure 4.2) is the SNP-X association matrix. Thus, the determinant of projection to common PCs is genetic correlation and correlation due to confounding, rather than how these blocks affect Y. Then, if only a few of the exposures truly influence Y, it is likely that, PCA will falsely identify the entire block as truly causal. This means the proportion of non-causal exposures within blocks of exposures that truly influence Y is a key determinant of specificity. To test this, we varied the proportion of non-causal exposures by varying the sparsity of the causal effect vector β vector and repeated the simulations, keeping the other simulation parameters fixed. As fewer exposures within blocks are truly causal, the performance in identifying true negative results drops for SCA (Figure 4.6). However, our simulation still provides a means of making comparisons across methods for a given family of simulated data.

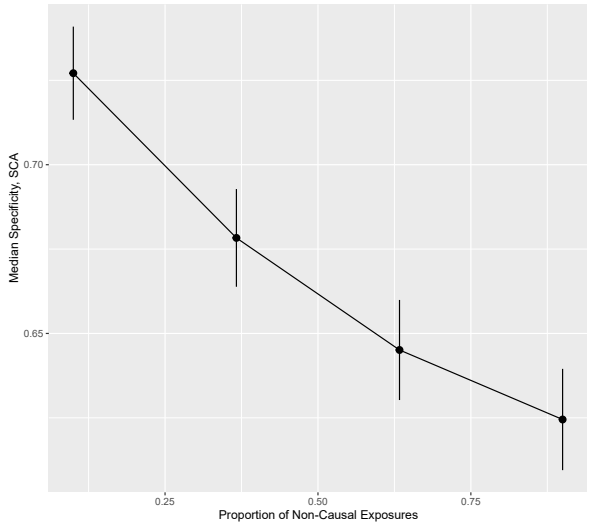


Figure 4.6: Specificity $\pm 1.96 SE_{MC}$ (ability to accurately identify true negative exposures) of SCA as a different proportion of exposures in each block are causal for Y. SE_{MC}: Monte Carlo SE.

4.4 Discussion

We propose the use of sparse PCA methods in MVMR in order to reduce high-dimensional exposure data to a lower number of PCs and infer the latter's causal contribution. As the dimensionality of available data sets for MR investigations increases (e.g. in NMR experiments [121] and imaging studies), such approaches are becoming ever more useful. Our results support the notion that sparse PCA methods retain the information of the initial exposures. Although there is no single optimal method that correctly factorises the SNP-exposure matrix, the goal is to find some grouping of the multiple, correlated exposures such that it may resemble a latent biological variable that generates the data. The SCA [112] and sPCA [107] methods, performed best in simulation studies and the SCA approach performed best in the positive control example of lipids and CHD. While conventional MR approaches did not identify any protective exposures for CHD, SCA identified a cluster of small and medium HDL exposures that appeared to independently reduce the risk of CHD.

This particular subset of HDL particles has previously been implicated in coronary calcification [122] and shown to be associated with coronary plaque stability [123].

By employing sparse PCA methods in a real dataset [102], we show that the resulting PCs group VLDL, LDL and HDL traits together, whilst metabolites acting via alternative pathways receive zero loadings. This is a desirable property and indicates that the second-step MR enacted on the PCs obtains causal estimates for intervening on biologically meaningful pathways. [124]. This is in contrast with unconstrained PCA, in which all metabolites contribute to all PCs. Previously, Sulc et al. used PCA in MR to summarise highly correlated anthropometric variables [125]. To our knowledge, this is the first investigation of different sparse PCA modalities in the context of MR. Our simulation studies revealed that sparse PCA methods exhibited superior performance compared to standard PCA, which had high false positive rates, and MVMR, which had high false negative rates. We additionally provide a number of ways to choose the number of components in a data-driven manner. Our proposed approach of a sparse PCA method naturally reduces overlap across components; for instance, in a paper by Sulc et al. [125], the authors use PCA and identify four independent axes of variation of body morphology. There are PCs that are driven in common by trunk, arm and leg lean mass, basal metabolic rate and BMI; a hypothetical benefit with sparse methods would be reduction of this overlap. This is an important topic for further research. When using PCA without any sparsity constraints, our simulation studies revealed numerous false positive results, at the opposite end of the nature of poor performance seen in MVMR; estimates were often misleadingly precise (false negative). Although appropriate transformations of

the exposures were achieved, we highly recommend exploring additional forms of T1E control to improve the performance of PCA. Nonetheless, sparse methods exhibited superior performance compared to both PCA and MVMR.

A previous work on sparse methods in genetics proposed their usefulness in multi-tissue transcriptome-wide association studies [126]. A finding of the study is that leveraging correlated gene expressions across tissues with sparse canonical correlation analysis (sCCA) improves power to detect SNP-trait pairs. Our approach that combines MR with sparse PCA also showed an improvement in power to detect causal effects of exposures on outcomes.

Our approach is conceptually different from the robust methods that have been developed for standard multivariable MR in the presence of weak instruments, such as MR GRAPPLE, which attempts to directly adjust point estimates for weak instrument bias, but are not a panacea, especially in the high dimensional setting discussed here. [101]. Furthermore, it reduces the need for a pre-selection of which exposures to include in a MVMR model. We present a complementary workflow through which we can include all available exposures with no prior selection, collate them in uncorrelated and interpretable components and then investigate the causal contribution of these groups of exposures. It avoids the risk of generating spurious results in such an extreme setting of high collinearity compared with MVMR IVW and MR GRAPPLE formulations. For example, a 2019 three-sample MR study that assessed 82 lipoprotein subfraction risk factors' effects on CHD used a UVMR and a robust extension of MVMR. A positive effect of VLDL- and LDL-related subfractions on CHD was reported, consistent in magnitude across the sizes of the subfractions [127]. Results

were less definitive on the effect of HDL subfractions of varying size on CHD, with both positive and negative effect estimates observed. In our study, the HDL subfractions were uniformly projected to similar subspaces, yielding a single component that was mainly HDL populated in all models, except for the SCA model 15 which projected the small/ medium and large/ extra-large HDL traits in two different components. In all cases, the association of the sPCs with CHD was very low in magnitude. Nevertheless, the direction of effects was in line with the established knowledge on the relationship between lipid classes and CHD.

Within the sparse PCA methods, there were differences in the results. The sPCA method [107] favoured a sparser model in which less than 10 metabolites per PC were used. This observation is also made by Guo et al [111]. The SCA method [112] achieved good separation of the traits and very little overlap was observed. A separation of HDL-related traits according to size, not captured by the other methods, was noted. Clinical relevance of a more high-resolution HDL profiling, with larger HDL molecules mainly associated with worse outcomes, has been previously reported [128].

4.4.1 Limitations

In the present study, many tuning parameters needed to be set in order to calibrate the PCA methods. We therefore caution against extending our conclusions on the best method outside the confines of our simulation and our specific real data example. Not all available sparse dimensionality reduction approaches were assessed in

our investigation and other techniques could have provided better results.

The use of sparsity may raise the concern of neglecting horizontal pleiotropy if a variant influences multiple components, but its weight in a given component is shrunk to zero. This would not occur for standard PCA where no such shrinkage occurs. Currently, our approach is not robust to pleiotropy operating via exposures not included in the model. Our plan is to address this as future work by incorporating median-based multivariable MR models into the second stage, as done by Grant et al [129].

Chapter 5

Examining the Causal Effects of Inflammation and BMI in Mood, Depression, and Treatment-Resistant Depression

In this Chapter, I present the results of investigations on how inflammation as proxied by C-reactive protein (CRP) and overweight independently affect depression. Parts of this work are published in *BMC Medicine* (in production).

5.1 Introduction

Depression is a highly prevalent mental health disorder, consistently ranked among the top three leading causes of disability worldwide [130]. There exist a multitude of pharmacological and psychosocial interventions that are effective and are applied in a stepwise manner [131]. The National Institute for Health and Care Excellence

(NICE) recommends that psychological interventions, such as cognitive behavioral therapy, psychodynamic psychotherapy, interpersonal therapy, and other modalities, are suggested as initial treatments for less severe depression, with a working definition of the latter being a cutoff of 16 on the PHQ9 scale. When depression is more severe, the use of antidepressant medication, particularly selective serotonin reuptake inhibitors (SSRIs), is recommended. Combining psychological therapy and medication is advised for patients with persistent depressive symptoms. Antidepressant selection should consider patient factors, such as age and comorbidities, while monitoring for treatment response and side effects. Continuation of antidepressant medication for 6 to 12 months after symptom remission is recommended to prevent relapse. Referral to specialist mental health services is suggested for severe, treatment-resistant depression or when suicidal ideation is present. Self-help interventions, such as guided self-help programs and computerized cognitive behavioral therapy, along with support from family, friends, and support groups, may be beneficial. These guidelines emphasize a comprehensive and individualized approach to depression management, taking into account the specific needs of the patient. It is important for healthcare professionals to consult the complete NICE guidelines and exercise clinical judgment when applying these recommendations in clinical practice. In people with a more severe presentation, prescription of an anti-depressant medication, coupled with psycho-social interventions, is recommended as first-line management [132]. For research purposes, response can be quantified as a meaningful reduction in a symptom severity scale, such as the Composite International Diagnostic Interview (CIDI) or Patient Health Questionnaire-9 (PHQ9) and may differ among different antidepressant agents [133]. It is recommended that, if a satisfac-

tory symptom reduction cannot be achieved within the first six weeks, then treatment can be augmented with other agents, including other antidepressants, lithium and anti-psychotics [131].

An inadequate response to at least two successive antidepressant medications, each administered for at least 6 weeks, is referred to as treatment resistant depression (TRD) and affects at least 7% of those initially diagnosed with depression [134]. Identifying contributors to treatment resistance early could potentially assist prompt management and guide appropriate interventions targeting other pathways that may predispose to treatment resistance. Recent advances in electronic health record analysis have allowed for the definition of a TRD phenotype in very large databases of routinely collected healthcare data, allowing linkage with existing genetic databases [135].

Overweight and obesity have also been shown to predict the development of depression in multiple observational studies. This relationship could be explained by worsening physical health with obesity which may in turn affect mood. A meta-analysis of 15 prospective cohort studies estimated that being overweight was associated with a 27% increase in the odds of subsequently developing depression. There is also evidence for a dose-response relationship, with obese individuals having higher risk for depression [136]. This relationship could be partially explained by social stigma due to the negative perceptions of overweight/obesity in certain cultures [137]. Another dimension of the effect is its potential appearance later in life. A meta-analysis reported a positive association only in adults older than 20 years of age but not in children and adolescents [136]. Recent studies derived a binary classification of

overweight (metabolically favourable and unfavourable adiposity) based on metabolic sequelae, namely hyperlipidemia, compromise in liver function, and sex hormone levels [138, 139]. Whilst individuals with favourable adiposity face much less of the commonly described adverse effects of high adiposity, both phenotypes appeared to exert effects of similar magnitude on the risk of multiple depression outcomes [139]. This was interpreted as a predominantly social, rather than biological, effect. Multiple studies have investigated the effects of weight loss on depressive symptoms in people with overweight or obesity. In general, caloric restriction, behavioural training, or supplements were used as interventions to incite weight loss, and weight loss was found to reduce depressive symptoms in most studies, as collated in a systematic review [140].

One aspect of the downstream metabolic consequences of overweight that is not explicitly captured by the phenotype of unfavourable adiposity is inflammation. Furthermore, evidence from genome wide association studies has suggested genetic variants important for cytokine and immune regulation predict major depressive disorder (MDD) [141]. C-Reactive Protein (CRP) is a protein synthesised by the liver as part of the inflammatory response. Measurement of CRP in serum is a common part of investigations for inflammatory conditions, e.g. microbial infections and autoimmune conditions. Given a stable general medical status, CRP levels are largely stable and multiple observational studies have investigated its potential utility as a proxy for disease progression in infectious disease [142] and autoimmune conditions. Despite its predictive value, its potentially causative role in driving the pathophysiological course of a disease has been disputed in multiple settings such as coronary

heart disease [143, 144]. In depression, recent work has indicated a higher CRP in 102 individuals with TRD compared with treatment-responsive patients and controls [145].

Despite the advantage of large sample sizes and extensive phenotyping that UKB offers, additional care has to be taken to avoid the inherent limitations of observational data. As phenotypes may be correlated due to confounding rather than a true causal relationship, the measurement of observational associations alone may not reflect a causal mechanism [146]. Mendelian randomisation (MR) is an epidemiological approach that employs genetic variants, most commonly single nucleotide polymorphisms (SNPs) as instrumental variables in order to circumvent environmental confounding [37]. By genetically predicting the levels of an exposure such as CRP by a set of relevant SNPs, the proxied levels of CRP reflect a value of CRP that may not be affected by later-life influences that could distort the value (e.g. BMI, smoking, auto-immune conditions). Associations between genetically predicted CRP and an outcome of interest can then much more readily be interpreted as a causal effect [147].

Previous work investigating the causal role of CRP, interleukin-6 (IL-6, major moderator of CRP), BMI and specific symptom dimensions of depression (sleep, appetite, suicidality) used LD score regression and a range of two-sample MR analyses. The results of this study did not find associations of CRP with any of the outcomes but report an association of IL-6 with suicidality [148]. MR has also been used to investigate how BMI and fat mass affects mood outcomes [149, 141, 150, 151].

The expansion of genome-wide association studies (GWAS) has led to the discovery of multiple new statistically significant causal effects. However, it is likely that many are false positives due to pleiotropy, the phenomenon whereby a SNP is an invalid IV due to exerting an effect on the outcome not through the exposure of interest. [152]. Multivariable MR (MVMR) can be used to assess whether an exposure causally influences an outcome *conditional* on a larger set of genetically instrumented exposures [70, 153]. Incorporating additional exposures stops them from acting as pleiotropic pathways and because of this MVMR is seen to be more robust than univariable MR (UVMR). Indeed, a wealth of evidence exists CRP as a downstream consequence of high body mass. For example, a recent GWAS of serum CRP levels on 204,402 individuals found that adjusting for BMI substantially reduced the strength of association between CRP and well known obesity genes (*FTO [154], TMEM18 [155], ABO*, previously described genes for obesity) [156].

In this paper we aim to estimate the causal contributions of CRP and BMI on TRD and other depression phenotypes using a combination of UVMR, MVMR and causal mediation analyses. We further investigate whether these highly correlated exposures exert an independent effect on depression phenotypes or if their effect is mediated, and if these relationships are constant across age distributions.

5.2 Methods

5.2.1 Data Sources

We used UK Biobank (UKB) as the primary data source for genetic and phenotypic information. UKB is a prospective cohort study that recruited approximately 500,000 individuals between the ages 37 and 73 from 2006–10 [87]. An extensive, validated questionnaire was completed by all participants gathering information on sociodemographic variables, environmental exposures and behaviours. All individuals were genotyped; specifically, single nucleotide polymorphisms (SNP) genotypes were obtained from the UKB Axiom^{TM} Array (450,000 individuals) and the UKBiLEVE array (50,000 individuals). These data have undergone rigorous quality checks [11]. Despite the public availability of better powered summary statistics, we restricted the analysis to UKB where access to individual-level data allowed for more flexibility in investigating a range of age- and sex-stratified analyses.

5.2.2 Exposures

We used BMI measurements and serum levels of CRP. Both exposures are associated with depression or TRD in observational epidemiological studies [145, 157, 134, 158]. We hypothesised that low-grade inflammation could be captured by serum levels of CRP [143]. This biomarker is part of the blood biochemistry test performed in UKB and is available in 429,141 European participants. For BMI, we used the baseline measurement taken at study enrollment. Data on 451,052 participants of European ancestry was available. Inverse normalised CRP and BMI were used to provide a more symmetric distribution than their raw values and to simplify the interpretation of resulting causal estimates as the effect of a 1 standard deviation (SD)

higher exposure on the outcome risk.

To assess the independent effects of adiposity and inflammation on depression outcomes beyond the traditional BMI measurement, we report a more granular approach where we examined two distinct phenotypes: unfavourable (UFA) and favourable adiposity (FA) [159]. FA and UFA were defined based on how SNPs that affect body fat percentage are associated with metabolic markers (high-density lipoprotein, sex hormone binding globulin, triglycerides, aspartate transaminase, alanine transaminase) [159]. Two sets of SNPs served as separate instruments to proxy body fat percentage, derived FA (36 SNPs) and UFA (38 SNPs) and perform the MR analyses described in Section 5.2.4. A previous work using MR to assess FA, UFA and depression found a differentiation between favorable and unfavorable adiposity, observing a statistically significant causal effect of FA on depression, and a modest effect of FA [139].

5.2.3 Outcomes

Multiple outcomes were derived in UKB participants based on both the mental health questionnaire (MHQ) and electronic health records. Previous works have described in detail the MHQ [160]), where questionnaires covering a range of psychological measurements (depression, anxiety, unusual experience, post-traumatic stress, substance use) were emailed to a subset of UKB participants and were completed by $n = 157,366$ individuals. An additional source of information is through linked electronic health records, including general practitioner (GP) visits. Here, we used a subset of $n = 230,000$ UKB participants and used codes to classify participants as having been diagnosed with depression. We use five outcomes: (1) GP diagnosis

of any of the read codes describing depressive disorders [135]; (2) lifetime MDD defined by the MHQ measurement [160]; (3) PHQ-9 [161] and (4) CIDI [162] depression severity measures (MHQ [160]); and (5) TRD [135]. A previous work in UKB underlined the low accuracy of self-reported depression measures and marked dilution of GWAS signals compared with clinical diagnostic phenotyping [163]. To guard against such potentially low resolution of the phenotype, we used both clinical diagnoses (GP diagnosis) and questionnaire data that was filled in by the participants (CIDI, PHQ9). These continuous outcomes (PHQ-9, CIDI) that were filled in by MHQ participants irrespectively of diagnosis both measure depression. A notable difference is that PHQ-9 focuses on the current severity of depressive symptoms in the past two weeks, whereas the CIDI targets the duration and impact of symptoms.

TRD

Linkage of the GP electronic health records and prescription data, enabled coding of TRD with information on antidepressants prescribed and TRD coded when individuals were prescribed at least two different antidepressants for six weeks. For the purpose of this study, we define the treatment interval at six weeks, which is more conservative than the four-week change encouraged by prescribing guidelines [164]. This conservative threshold helps reduce the likelihood that drug switching was due to side effects, while still allowing for adequate efficacy.

5.2.4 Statistical Analyses

Observational Associations

As a baseline model, study outcomes were directly regressed on the observed values of the exposures. For continuous outcomes (CIDI, PHQ-9), multivariable linear models were used. For binary outcomes, we used logistic regression. All models were adjusted for age, assessment centre and sex.

MR

A series of one-sample MR analyses (Chapter 2.2) were conducted within the UKB cohort. *Instrument Selection*: To avoid winner's curse bias (inflation of effect estimates due to random variation if the same dataset is used for selection and analysis [165]), external GWAS datasets were screened for genome wide significant SNPs (P$< 5x10^{-8}$). SNPs were identified that associated with CRP, BMI and MDD in publicly available GWAS studies not overlapping with UKB [166, 167, 141]. For CRP, SNPs reported by the CHARGE study were extracted [156], whilst for BMI, the Locke et al. study was used [167] with a further specification of a European-focused instrument of 73 SNPs as described by Tyrrell and co-authors [150]. Of the 97 reported SNPs in the Locke et al. study [167], we follow Casanova et al. [139] and limit this to European-specific 76 SNPs. Three further SNPs are excluded due to known pleiotropic effects leading to the final 73 SNPs that constitute the instrumental variables. Specifically, the SNPs rs11030104 (BDNF), rs13107325 (SLC39A8), are excluded because of associations with phenotypes that likely influence depression, respectively with regular smoking, with BP and HDL, and with many traits including alcohol, testosterone and cognitive domains. Clumping was performed with a

window of 50kb and an r^2 of 0.001 was used to exclude all SNPs in pairwise linkage disequilibrium (LD). This ensured our instrument set was comprised of approximately uncorrelated SNPs. For the analysis where all three CRP, BMI and MDD were genetically proxied, a genetic risk score from the 178 Levey et al. variants was used for MDD [168] in order to facilitate the computationally expensive bootstrap procedure with a single instrument that retains as much variance explained as possible.

After extraction of genotype dosages at the individual level, individual LD matrices were constructed. If any non-negligible amount of pairwise LD was observed ($r^2 > 0.05$) for two SNPs on the same chromosome, the SNP with the largest $p-$value was retained.

5.2.5 MR designs

The different MR analyses reported are visually presented in Figure 5.1. All analyses follow the one-sample MR framework, where the exposure (in our case CRP and BMI), genetic variants and the outcomes (depression and TRD) are measured in the same individuals [169]. Within this one-sample framework, we further apply: (a) univariable MR (UMVR) to estimate total causal effects; (b,c) multivariable MR (MVMR) to estimate direct effects and to perform mediation analysis; (d) pleiotropy robust MVMR as a sensitivity analysis.

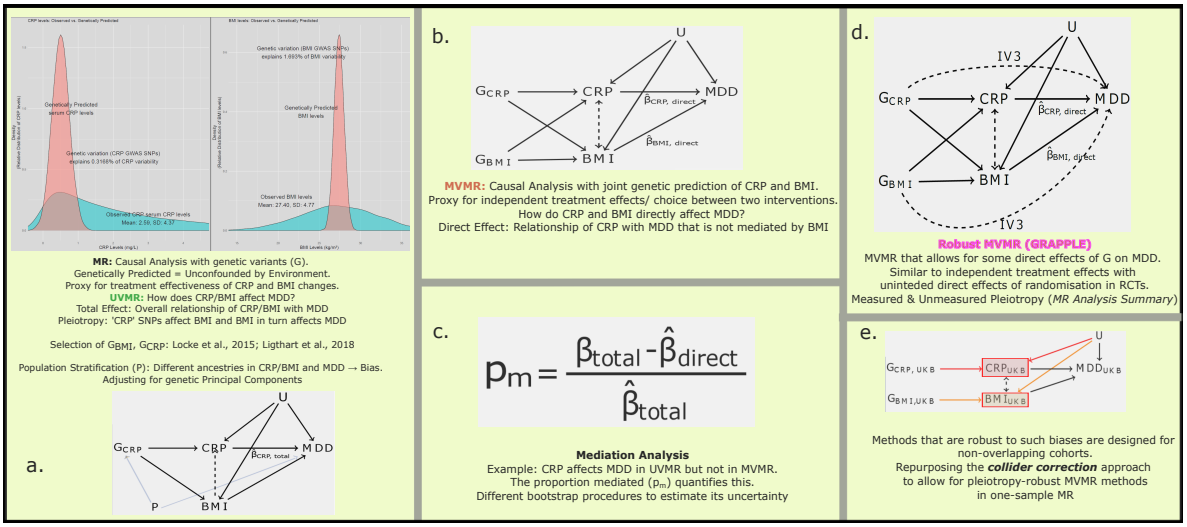


Figure 5.1: Methods Overview. **a**. Causal diagram (DAG) representing the assumed relationship between genetic variants for CRP (G_{CRP}), measured levels of serum CRP and BMI, and major depressive disorder (MDD). The dashed line between CRP and BMI represents a potential contribution of BMI to CRP levels. **b**. DAG for an MVMR analysis that genetically proxies both CRP and BMI (G_{CRP}, G_{BMI}) enables estimation of the direct causal effect of CRP and BMI on MDD. **c.** Estimation of the proportion of the CRP effect mediated by BMI (p_m). **d.** Robust MVMR to account for unmeasured pleiotropy as well as measured BMI pleiotropy. If some of the genetic variants in G_{CRP} or G_{BMI} affect MDD directly, other than just through changing CRP or BMI levels, the estimated effects will be biased. Robust methods such as MR GRAPPLE protect against this.

.

MR Analysis Summary

The following steps were taken in order to rigorously perform the MR analysis. Firstly, to measure instrument strength, we report the mean F statistic for UVMR analyses and the conditional F-statistic [100] for multivariable MVMR analyses. This latter statistic provides a measure of how well a single exposure is instrumented conditionally on the other exposures. Low values indicate the presence of multi-collinearity in the genetically predicted exposures, which leads to weak instrument bias. Secondly, individual-level data UVMR analyses in UKB were carried out using two-stage least squares (TSLS) approach for continous outcomes. For binary outcomes, the second stage linear regression was replaced with logistic regression (otherwise known as two stage predictor substitution). To implement the MVMR analysis, whilst protecting against bias due to weak instruments and pleiotropy, we employed a novel extension of the recently proposed technique of Collider-Correction [85] to the multivariable setting (Figure 5.1 d, see Appendix for further details). Thirdly, whilst MR methods are generally robust to traditional confounding, they are more susceptible to genetic confounding due to population stratification [170]. To address this, we adjust for the first five genetic principal components in all analyses [171]. This way, we also aim to partly adjust for relatedness; a stricter approach of completely excluding individuals that are related would reduce sample size in an already moderately powered context (loss of 16% of TRD cases). We present the univariable and multivariable analyses in this subset of unrelated individuals. We also use a subset of UKB that includes only individuals of European ancestry ($n = 451,025$). Finally, we assessed the extent of heterogeneity amongst causal estimates from different SNPs, a proxy of residual pleiotropy, using the Sargan test [172]. The Sargan test is useful for one-sample MR

analyses as it can indicate when multiple instruments are valid as a group, testing if they provide similar estimates as single instruments or in linear combination [62].

For continuous outcomes, causal estimates reflect the effect of a 1 SD change in the exposure on the outcome. For binary disease outcomes, causal estimates were obtained from a logistic regression and reflect the effect of a 1 SD change on the log-odds of the depression-related outcome (Chapter 2.2).

5.3 Estimation of Mediated Effects

As discussed in Chapter 2.5, the extension of multivariable MR (MVMR) allows for the joint estimation of the effect of multiple exposures on the outcome [70]. Recent developments in MVMR focus on exploring questions of mediation of the effect, that is whether the apparent effect of one exposure is mediated by another [72]. estimating the proportion of the effect mediated.

Regarding this approach, Carter et al. [72] recently presented an authoritative review on MR and mediation analysis. The process includes the sequential fitting of a univariable MR model with a genetic instrument for the exposure, a multivariable model with an instrument that jointly predicts the exposure and the mediator, and finally the estimation of the proportion of the effect mediated (π_m). The authors suggest the use of a bootstrap to obtain uncertainty estimate for π_m. Given the availability of multiple methods of performing a bootstrap and the observation that a non-parametric bootstrap tends to provide imprecise answers in the case of the binary outcomes of TRD and GP diagnosis of depression in the real data applications in this Chapter 5, we aimed to explore different ways of estimating π_m and its uncertainty. We compare the different approaches with a simulation study and provide a real-data application

for mediation of apparent physical activity (PA) effects by body mass index (BMI). We also investigate how residual pleiotropy affects the results.

The motivation for an additional investigation stemmed from observations from real data analyses, where there still was some improvement in estimating mediated effects with the Bayesian bootstrap method, in contrast with the comparable performance shown in the simulation studies. The simulation studies were constructed for a linear outcome from a normal distribution. A careful consideration of the exact applied example reveals that, in reality, the outcomes considered were binary outcomes of relatively low prevalence. These empirical findings motivated the need for another dedicated examination of the Bayesian bootstrap's performance in the specific presence of sparse outcomes, particularly in estimating mediated effects with binary data. By building upon these real-world observations, this study aimed to provide a more nuanced understanding of the Bayesian bootstrap's efficacy and its potential for enhancing precision in estimating mediated effects in Mendelian randomisation analyses with binary outcomes.

We implement our mediation analysis as described below in UKB individual-level data (where X and M are the exposures of interest (CRP, BMI) and Y is the outcome) (Fig. 5.1c.):

- Estimate the total effect of X on Y by UMVR, $\hat{\beta}_{total}$.
- Estimate the direct effect of X on Y $\hat{\beta}_{direct}$ via an MVMR and the indirect effect as $\hat{\beta}_{indirect} = \hat{\beta}_{total} - \hat{\beta}_{direct}$.
- Estimate the quantity $\hat{\pi}_m = \frac{\hat{\beta}_{indirect}}{\hat{\beta}_{total}}$ and its confidence interval via a non-parametric bootstrap of the data in order to test the null hypothesis $H_0 : \pi_m$=0. When $\hat{\beta}_{indirect}$ and $\hat{\beta}_{total}$ have the same sign, $\hat{\pi}_m$ can be interpreted as an estimate for the pro-

portion of the effect of X on Y mediated via M.

Bayesian bootstrap for uncertainty quantification and sparse binary outcomes

Carter et al [72] recommend the use of a standard non-parametric bootstrap in order to provide confidence intervals for $\hat{\pi}_m$. However, using simulations consistent with the real data we observed that it tended to over-estimate the true parameter uncertainty. For this reason, we additionally developed a method to implement Rubin's Bayesian Bootstrap [173]. Whilst the standard non-parametric bootstrap samples individuals with replacement from the original data, each iteration of the Bayesian bootstrap is always based on the complete data, but the weight they receive in the analysis is instead generated from a Dirichlet distribution. Our simulations showed that the Bayesian Bootstrap produces confidence intervals with similar coverage to the non-parametric bootstrap, while it improves performance for very sparse binary outcomes (Section 5.4.4). For further details see the Appendix. We used this method to test the hypotheses that the effects of CRP are mediated by BMI and that the effects of CRP and BMI on TRD do not operate solely through MDD, as previously investigated by Maske et al. [174]).

5.3.1 Sex Specific Effects and Age as a Moderator

Our individual level data methods enabled us to perform analyses separately in males and females, formally testing for heterogeneity in causal estimates of BMI and CRP on depression between males and females, using Fisher's z-score. In addition to sex-stratified analyses, we also explored the extent of heterogeneity in causal effects of BMI and CRP across age strata. To achieve this we split the total sample of $451,025$ European participants to seven five-year sub-samples and performed

meta-regression to assess whether age was an important predictor of causal effect heterogeneity.

We implemented the following regression model to see if there was a trend in causal estimates due to age:

$$\hat{\beta}_{XY_i} = \beta_0 + \beta_{age} a\bar{g}e_i + \varepsilon_{XYi}.$$

Here $\hat{\beta}_{XY_i}$, $i = 1, ..., 7$ expresses the causal effect estimate in the ith five-year age stratum, with β_0 the intercept and β_{age} the average age-moderation parameter of the of the mean stratum age $a\bar{g}e_i$). To assess how this model compares against a simple pooled estimate that does not take age into account, we estimated two heterogeneity statistics. Firstly, Cochran's Q statistic:

$$Q = \sum_{i=1}^{7} w_i(\hat{\beta}_{XYi} - \hat{\beta}_{All})^2, \qquad \hat{\beta}_{All} = \frac{\sum_{i=1}^{7} w_i \hat{\beta}_{XYi}}{\sum_{i=1}^{7} w_i,}$$

on 6 degrees of freedom, and Rucker's Q statistic: [175, 176])

$$Q' = \sum_{i=1}^{7} w_i(\hat{\beta}_{XY_i} - (\hat{\beta}_0 + \hat{\beta}_{age} a\bar{g}e_i))^2,$$

on 5 degrees of freedom. We then calculated their difference Q_{diff}. Under the null hypothesis that age was not a predictor of the causal effect, this statistic is distributed $Q_{diff} \sim \chi^2_{df=1}$.

5.3.2 Sensitivity Analyses

Choice of Instrument: CRP SNPs have been shown to be highly pleiotropic and affect a range of cardiovascular outcomes and serum lipid traits [166]. In addition to the data-driven pleiotropy-robust methodology described above, we perform a comple-

mentary approach of limiting the SNPs used as instruments exclusively to the *CRP* locus. This is based on the hypothesis that SNPs in this specific location represent more biologically relevant CRP variants rather than indirect associations.

We also provide an analysis using the recently proposed cis-MR approach by Patel et al. [57] which uses the complete set of highly correlated SNPs in a single area of biological relevance, instead of limiting the analysis to few independent signals. The argument is that SNPs in close proximity to a gene that codes for a precise molecular target are less likely to affect other phenotypes. We therefore use the CRP genomic region ($1 : 159712288 - 4589 \pm 100$kb) and select correlated instruments in the external study [156]. We apply no $p-$value selection threshold. We then use the method of Patel et al. [57] to test the hypothesis that CRP has no effect on the depression outcomes. In this analysis, the exposure data is from the CHARGE study summary statistics [166] and follows the two-sample MR framework.

For BMI, we aimed to locate the Locke et al. instrument to variants that affect BMI through a central mechanism and would hypothetically be more likely to affect depression through other pathways; a tissue enrichment analysis of genes associated with depression showed specific patterns of expression in the brain [168]. We follow a similar approach to Leyden and co-authors [177], employing a dedicated database [178]. The process is presented in detail in Section A.2.1.

Reverse Causality: Bias due to reverse causality may emerge when an outcome affects the risk factor (Chapter 1.3), that is a hypothetical causal effect of mood dysregulation on inflammation status and weight. There is abundant clinical literature

supporting a longitudinal, potentially bidirectional association of these phenotypes with mood disorders [179, 180] and changes in appetite, eating behaviours and unintended weight gain or weight loss all are included in the diagnostic criteria for MDD [181]. We therefore studied how low mood affects CRP and BMI, using 102 SNPs that associate with MDD [182] to genetically proxy CIDI, one of the MHQ mood questionnaire completed by a subset of UKB participants. Given the previous evidence for depression and BMI sharing a genetic component and as CRP variants also influence BMI, we used Steiger filtering to exclude MDD SNPs that associate more strongly with BMI or CRP rather than MDD, so that the SNP-set G_{MDD} consists of SNPs that predict CIDI more strongly than CRP or BMI [183]. As the instrument strength of G_{MDD} suggested that there may be dilution bias due to weak instruments, a combination of the collider correction approach [85] and the weak-instrument robust MR RAPS approach were also reported [184].

5.4 Results

5.4.1 Patient Characteristics

Table 5.1 reports the individual characteristics of the UK Biobank participants. The TRD phenotype as previously curated was available for $n = 189,917$ controls and $n = 2199$ TRD individuals. At the observational level, participants that went on to be diagnosed with TRD had a higher CRP and BMI. The baseline measurements of mood indicated that they scored higher both for CIDI and PHQ9 (Table 5.1). The proportion of females in the GP depression and TRD groups is higher than in the control group. The majority of individuals have CRP levels that are not consistent with clinically active inflammation, however it seems that there is variability of CRP

levels according to depression status, with people with depression and TRD having a 0.55 and 1.1mg/L higher CRP on average respectively than controls.

	Controls	GP-based depression cases	TRD cases	Group Comparison (Value, p)
N	173,786	18,330	2,199	
Age	57.42 (±8.11)	56.11 (±7.95)	56.43 (±7.82)	224.6 ($< 10^{-10}$)
$\%$ Female	0.485	0.636	0.724	1804 ($< 10^{-10}$)
BMI	27.38 (±4.67)	28.26 (±5.37)	29.41 (±5.96)	359.1 ($< 10^{-10}$)
CRP (mg/L)	2.54 (±4.35)	2.99 (±4.57)	3.64 (±5.42)	109.7 ($< 10^{-10}$)
CIDI_MDD*	2 (±10.43)	5.45 (±11.55)	6.63 (±1.68)	3290 ($< 10^{-10}$)
PHQ9*	2.11 (±2.89)	4.52 (±5.09)	9.06 (±6.83)	2240 ($< 10^{-10}$)

Table 5.1: Individual characteristics. Mean (±SD). *: CIDI and PHQ9 were measured in a different, partly overlapping subset of UKB participants ($n = 146,067$). Comparisons across groups are performed with analysis of variance (ANOVA) tests, and F-values and p-values are reported. A χ^2 test was used to compare the proportion of females across the three groups.

5.4.2 Univariable & Multivariable MR

BMI: Figure 5.2 presents the causal estimates for BMI on a range of depression outcomes. In the UVMR analyses of BMI in green, all estimates suggest a robust causal effect, with 95% confidence intervals excluding the null. A 1-SD increase in genetically proxied BMI was associated with 13.9$\%$ (95$\%$ CI$_{OR}$: 8.3$\%$, 22.1$\%$) higher odds of a lifetime diagnosis of depression, 19.7$\%$ (5.1$\%$, 32.3$\%$) higher odds of a GP diagnosis of depression, and 41.9$\%$ (2.0$\%$, 95.4$\%$) higher odds of TRD (Figure 5.2). In the MVMR models (in red), this pattern persists, with the analyses with external weights [166, 167] agreeing. The repeat analysis under a stricter approach of completely excluding related individuals also supports that BMI has a positive effect on all outcomes (Figure A.4). In the sex-stratified analysis in females (Figure A.5), there is evidence for a causal effect of BMI on all outcomes except for TRD whereas in males, a robust effect is observed only on PHQ9 (severity). Although the BMI point estimates appear larger in magnitude for females, the z-test does not suggest statistical significance at the 95% level (Table A.3).

Repeating the analyses in each five-year age stratum, there was an attenuation of the causal effect of BMI on PHQ9 with age ($\beta_{age} = -0.025$ PHQ9 total units per age stratum, $p = 0.011$) (Figure A.7). This effect was also nominally statistically significant for a non-linear trend ($\beta_{age^2} = -0.22, p = 0.0397$, Table A.5). The heterogeneity statistics indicate a better fit for the model that includes age for this comparison ($Q_{diff} = 6.494, p_{Q_{Diff}} = 0.011$). The evaluation of instrument strength in specific groups suggested a lower strength of association in the $69 - 74$ age group (Table A.2).

CRP: In UVMR, CRP displays an effect on all outcomes (Figure 5.2). A 1-SD increase in genetically proxied CRP was associated with 12.7% (1.0%, 22.3%) higher odds of a lifetime diagnosis of depression, 20.9% (95% CI: 2.0%, 40.5%) higher odds of a GP diagnosis of depression, and 63.2% (13.9%, 146.0%) higher odds of TRD (Figure 5.2). In the sex-stratified analyses, strong causal estimates are present for GP diagnosed depression and TRD in females only. All these associations do not persist in the MVMR analyses, where upon jointly predicting BMI and CRP in MVMR models, the estimated effect moves close to the null and significance is lost (Figure 5.2). Of note, the consistency of the results remains unchanged even when only unrelated individuals are considered, as illustrated in Figure A.4. The analysis with external weights suggests more modest CRP effects (Figure A.6). There does not seem to by a clear modification of the estimated effects by age but there seems to be heterogeneity on the age-specific effects on GP diagnosis (Figure A.7).
In Table A.4, applying the Sargan test revealed sstatistically significant heterogeneity for all analyses. Therefore, these analyses could be potentially affected by at

least one SNP exerting a pleiotropic effect. The reported estimate standard errors account for this heterogeneity, and the estimates themselves are valid under the assumption that the pleiotropy is balanced. A repeat test with the CRP-BMI pair of exposures in an MVMR model (Table A.4) indicates that heterogeneity is likely even in the joint model, motivating a pleiotropy-robust method as a more appropriate modelling choice.

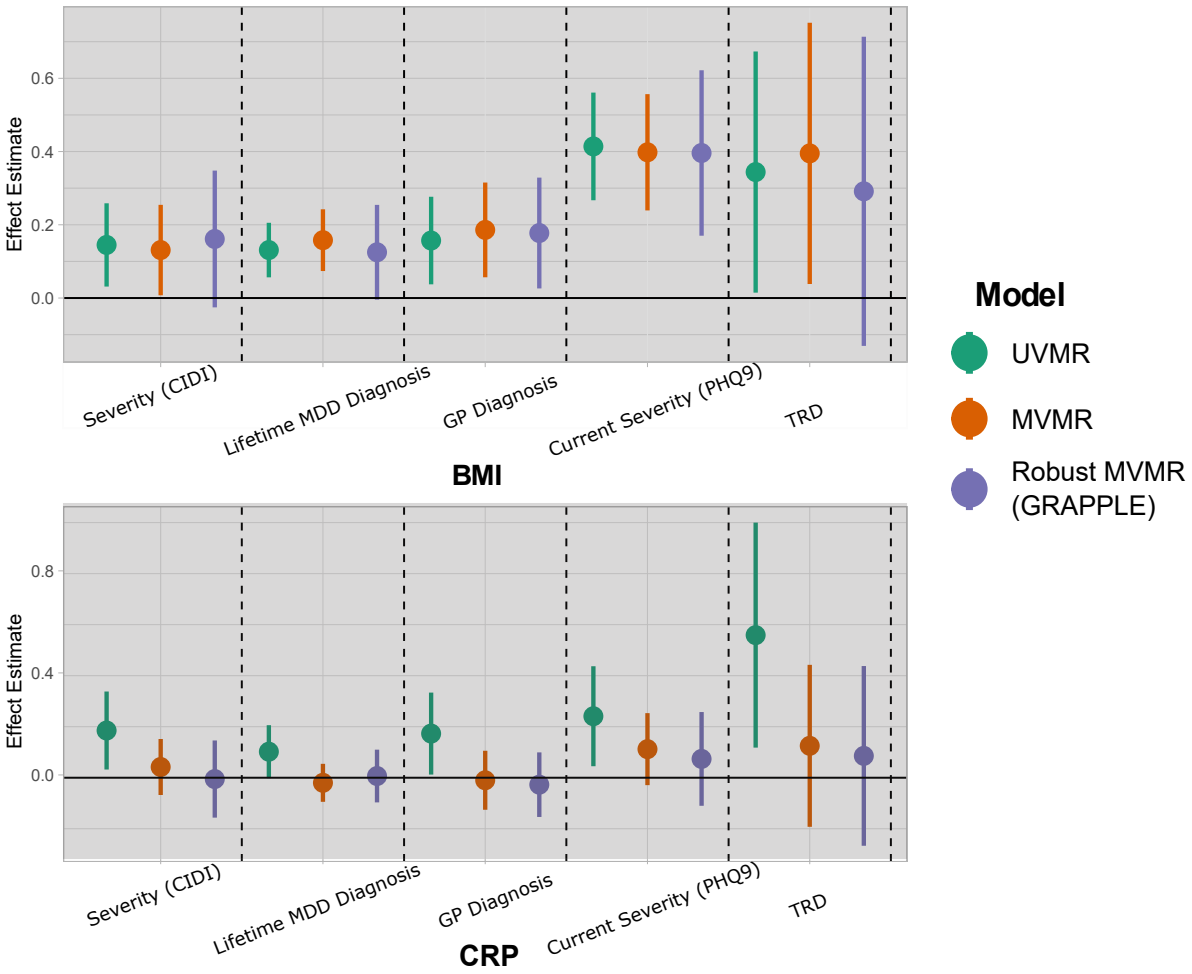


Figure 5.2: Effects of BMI and CRP on various depression-related outcomes as measured by univariable and multivariable MR models. The CRP effect is measured by using 45 CRP SNPs as instruments. In the horizontal axis, the five outcomes are presented; in the vertical axis, the effect size is shown as a point and the line denotes the 95% confidence intervals (CIs). UV: Univariable MR; MV: Multivariable MR; GRAPPLE: Robust Multivariable MR with MR GRAPPLE; CIDI: Composite International Diagnostic Interview; PHQ9: Patient Health Questionnaire-9; TRD: treatment-resistant depression

.

5.4.3 Sensitivity Analyses

Favourable and Unfavourable Adiposity

In Figure A.8, we substituted BMI with unfavourable and favourable adiposity, and repeated the assessment of how these affect the outcomes and how the effect of CRP changes. Unfavourable adiposity appears to influence all outcomes, while CRP showed an effect on TRD, PHQ9, and CIDI. In the multivariable models, we found that the point estimate of CRP remained relatively stable, indicating a consistent association with the outcomes. These findings highlight the independent contributions of both unfavourable adiposity and CRP to depression, emphasizing the need to address these factors in the prevention and treatment of depressive symptoms.

Robust MVMR

Results for the pleiotropy-robust GRAPPLE implementation of MVMR are presented in Figure 5.2 (purple). This method simultaneously accounts for weak instrument bias, imbalanced pleiotropy (via penalization of outliers) and sample overlap [185]. The results appear to be largely concordant with those of the MVMR, with slightly lower precision and lower magnitude of effects. In this analysis, BMI is associated only with a GP diagnosis of depression and PHQ9. There was insufficient evidence to confirm an effect of CRP with any of the outcomes in the multivariable models.

CRP Gene-Specific Instrument and Tissue-Specific BMI SNPs

Specifying the search for valid CRP instruments in the area around the *CRP* gene, 194 variants were identified in the selection sample [156]. After clumping, four were retained as independent (*rs11585798, rs2794520, rs3934775, rs12727193*) and one

(*rs2794520*) was strongly associated with CRP (β ($95\%CI$): -0.182 (-0.189, -0.176) with *T* and *C* as the effect and non-effect alleles, $p = 1.2 \times 10^{-305}$). This SNP was carried forward for the analyses in UKB. As in the selection study, *C* carrier status was associated with lower CRP serum levels, more strongly in females (-0.172 (-0.176, -0.167) in all, -0.185 (-0.191, -0.178) in women, -0.158 (-0.162, -0.153) in men; Fisher's $z = 6.01$, $p_{diff} < 1.9 \times 10^{-9}$). Using only this SNP as an instrument, UVMR indicates a negative association of CRP with a GP diagnosis of depression (-0.155 $(-0.223, -0.087)$). The sex-specific analysis implied a stronger effect in females (-0.308 $(-0.570, -0.046)$) and an effect in males only for PHQ9 reaching statistical significance at the 5% level. (-0.31 $(-0.580, -0.039)$).

In the MVMR analysis where we also proxy BMI with 73 SNPs, CRP is judged to negatively influence GP diagnosis of depression (-0.151 (0.051)); this is independent of BMI. This effect is also observed in males (-0.211 (0.089)), whereas in females it does not reach statistical significance at the 5% level (-0.113 (0.062)). In the robust MVMR analysis, a statistically significant negative effect was estimated (-0.144 (0.055)).

Alternative Instruments

Regarding CRP, the focused search in the *CRP* genetic region yielded 194 SNPs. Clumping greatly restricted the available variants to guarantee independence and only one variant would be used (*rs2794520*). Using the recently proposed approach of Patel et al., it was possible to retain all 194 variants, extract them at the individual allele dosage level and decompose them in independent genetic signals [57]; namely, the variants presented in Figure A.11 were projected in 10 principal compo-

nents which were then used for MR inference. The results are shown in Table A.6. Although the estimates were more precise than those reported in Section 5.4.3, their magnitude was lower and consequently none surpass the conventional significance threshold.

The tissue-specific MR analyses are presented in Section A.2.1. Of the 73 BMI SNPs, 23 were retained as being preferentially expressed in brain-related tissues and 31 others that were mapped to coding regions were expressed in the periphery. Similar estimates were obtained for all outcomes. For TRD, UVMR and MVMR suggested a positive effect of BMI on depression only when the peripherally focused instrument was used ($\beta_{UVMR}(CI) : 0.722(0.246, 1.123)$, $\beta_{MVMR}(CI) : 0.745(0.251, 1.229)$). In contrast, the UVMR and MVMR estimates from the instrument that included brain-expressed genes failed to reject the null ($\beta_{UVMR}(CI) : 0.139(-0.472, 0.750)$, $\beta_{MVMR}(CI) : 0.136(-0.481, 0.752)$). Both instruments provided similar results for all outcomes including TRD in the pleiotropy-robust MR GRAPPLE method.

Reverse Causality assessment using UVMR

Applying the Steiger filtering routine for BMI, 16 of the 102 SNPs reported by Howard et al. [182] were excluded. For the remaining 86 SNPs, instrument strength for the genetic prediction of CIDI was estimated at $F_{stat} = 5.337$. In 2SLS, the causal effect of genetically predicted CIDI on BMI was estimated as β (SE): $0.077(0.016)$, but was likely to be affected by weak instrument bias. Using MR-RAPS, the uncertainty in this result increased substantially (β (SE):$0.074(0.045)$. For CRP, Steiger filtering indicated that there were 27 SNPs that predict a larger proportion of the CRP variance compared with the CIDI variance. In the remaining 75 SNPs ($F_{stat} = 4.132$), an asso-

ciation of genetically proxied CIDI with CRP was observed (β (SE): 0.080 (0.020)). Applying MR-RAPS as above, the resulting estimate did not maintain significance at the 95% level (β (SE): 0.101 (0.053)). In summary, a potential causal effect of the genetically predicted mood questionnaire score on BMI and CRP was found in 2SLS, but these estimates are likely affected by weak instrument bias.

Mediation Analysis

5.4.4 Improved performance of the Bayesian Bootstrap: simulation evidence

The investigation was prompted by observations from real data analyses, where the proportion mediated is relatively imprecise. We investigated three different methods of obtaining $\hat{\pi}_m$ and its uncertainty. The data generating model is visualised in Figure 5.3.

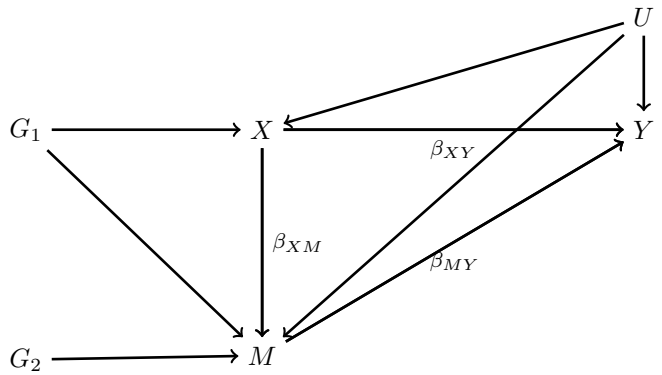


Figure 5.3: Directed Acyclic Graph. The exposure X and the mediator M exert two independent effects on Y. Genetically proxying only X can result in an inaccuracy in the estimation as $\hat{\beta}_{XY,Univariable}$ will be capturing the *total* effect ($\beta_{XY} + \beta_{XM} * \beta_{MY}$).

The following data-generating models for the exposure X, mediator M and outcome Y were used, consistent with the DAG in Figure 5.3:

$$
\begin{aligned}
X_i &= \sum_{p=1}^{P} \gamma_{Xp} G_{1ip} + U_i + \varepsilon_{Xi}. \\
M_i &= \beta_{XM} X_i + \sum_{p=1}^{P} \gamma_{1M} G_{1ip} + \sum_{p=1}^{P} \gamma_{2Mp} G_{2ip} + U_i + \varepsilon_{Mi}. \\
Y_i &= \beta_{XY} X_i + \beta_{MY} M_i + U_i + \varepsilon_{Yi}.
\end{aligned}
$$

The UVMR estimate for the causal effect of X on Y, β_{XY},is biased if we exclusively use the set G_1 as instruments for X, since the SNPs exert a pleiotropic effect on Y through M. If, however, we genetically proxy both X and M in an MVMR analysis, this bias can be removed and we can also estimate the extent to which the effect of X on Y is mediated through M [72]. In this setting, the target of the bootstrap methods is to quantify the uncertainty in the mediated proportion estimate:

$$
\begin{aligned}
\hat{\pi}_m &= \frac{\hat{\beta}_{XM}\hat{\beta}_{MY}}{\hat{\beta}_{XY} + \hat{\beta}_{XM}\hat{\beta}_{MY}} \\
&= 1 - \frac{\hat{\beta}_{XY}}{\hat{\beta}_{XY} + \hat{\beta}_{XM}\hat{\beta}_{MY}},
\end{aligned}
$$

where in order to interpret this quantity as a proportion we require that the total effect and mediated effect are of the same sign. We apply three methods to estimate the sampling distribution of $\hat{\pi}_m$.

1. Standard Non-parametric bootstrap: in each iteration, a bootstrapped sample of equal size is constructed by sampling with replacement from the original data;

2. Quasi-Bayesian estimation of confidence intervals [186] implemented with the *mediation* R package;

3. A fully Bayesian bootstrap [173]: in each iteration a vector of Dirichlet weights

($w_i = \frac{\Gamma_i(1,1)}{\sum_{i=1}^{N}\Gamma_i(1,1)}$) is generated and used to fit the the UVMR and MVMR regressions on the original complete data, using regression weights w_i for subject i.

Results for our simulation are presented in Figure 5.4 in which we show the mean (point estimate) and 95% confidence interval for $\hat{\pi}_m$ (top) and its coverage (bottom) based on the three bootstrap procedures. Similar to our previous simulation, data were generated to give a range of conditional F statistic values for X between 2 and 30. We see that the three methods perform comparably in terms of bias, with small differences in uncertainty estimation and coverage. The fully Bayesian bootstrap procedures performed well, furnishing a sampling distribution for $\hat{\pi}_m$ centred on the true value, from which could be derived confidence intervals with the correct nominal coverage. We therefore conclude that all methods perform comparably in this particular context.

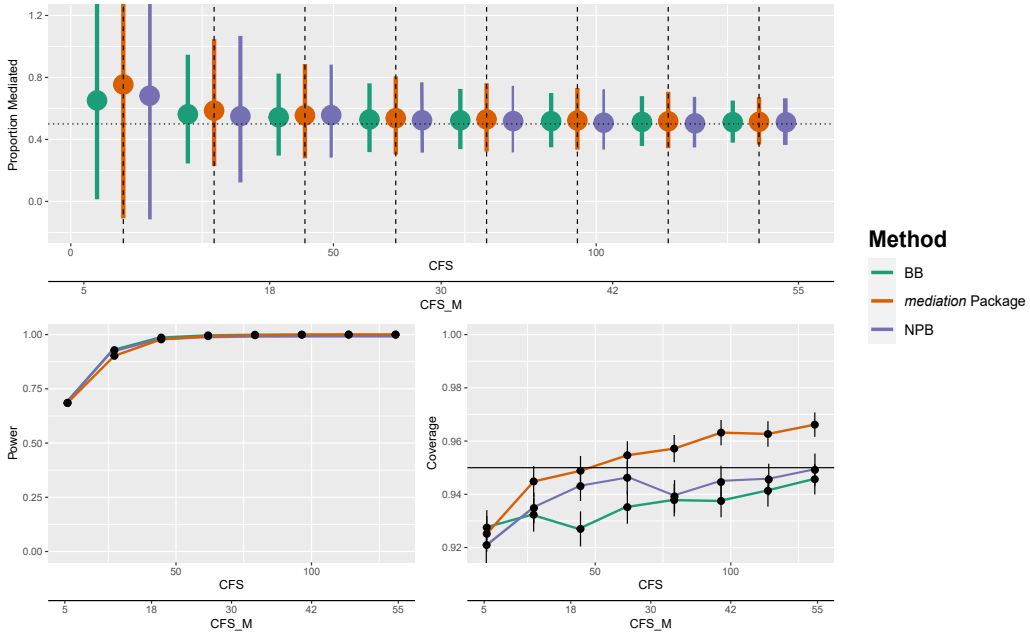


Figure 5.4: Uncertainty in estimating the proportion of mediated effects, simulation Results. CFS: Conditional $F-$statistic for the exposure X in Figure 5.3; CFS$_M$: Conditional $F-$statistic for the mediator M. Error bars in the coverage and power plots represent the Monte Carlo error for $s = 6000$ simulations. BB : Bayesian bootstrap; *mediat_package*: implementation with the *R mediation* package; norm: non-parametric bootstrap

A further investigation was prompted by observations from real data analyses, where the Bayesian bootstrap method showed some improvement in estimating $\hat{\pi}_m$ despite comparable performance in simulation studies. The simulation studies used linear outcomes from a normal distribution, and we hypothesised that the difference could be attributed to the low prevalence of the binary outcomes (TRD) in the real data. We therefore performed a dedicated simulation study with a binary outcome that is sparse (or rare), keeping all other parameters (e.g. SNP-exposure strength of association, sample size, confounder effects) fixed. We apply the BB (BB1), the NPB, and a Bayesian bootstrap that resembles the structure of the *mediate* function (BB2). The results shown in Figure 5.5 suggest a benefit for BB when the outcome prevalence is below 2%, but with a very similar performance for more frequent outcomes.

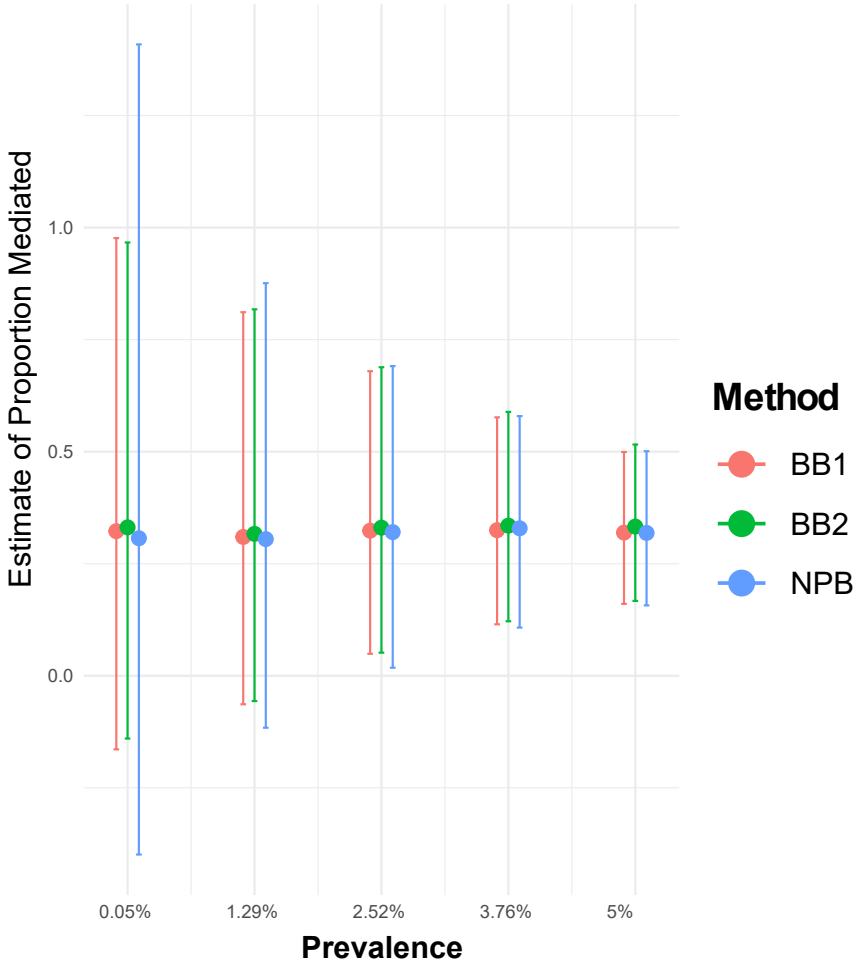


Figure 5.5: Simulation Results ($\hat{\pi}_m$ and 95% Confidence Intervals) for Increasing Prevalence of Binary Outcome.

The results for the mediation analysis are presented in Figure A.9. When using the discovery set of CRP SNPs ($n_{CRP} = 45$), the effect of CRP on GP diagnosis and TRD appears to be mediated by BMI ($107.82\%(56.1\%, 357.02\%)$ and $78.87\%(33.35\%, 197.23\%)$ respectively). For the ever depressed outcome, a similar magnitude was observed but the CIs do not indicate significance at the 95% level ($82.87\%(-208.39\%, 250.43\%)$) (Figure A.9). In the sex-specific mediation analysis, BMI appears to mediate the effect of CRP on TRD and GPD in females and on the ever depressed in males (Figure A.10). With the BB method, the CIs for π_m are somewhat narrower for TRD.

As discussed above, there is a concern for residual pleiotropy in this particular set of SNPs. Since 2SLS models are used throughout and CRP SNPs are known to be highly pleiotropic (Section 5.4.3), the estimates with these 45 SNPs may be biased. In the multivariable MR, this is partly alleviated, however there is a possibility of other pathways not related to BMI affecting the outcome. As a result, the estimates of the proportion mediated may be distorted with the full set of 45 SNPs. In a repeat mediation analysis with only one SNP in the CRP locus used to genetically proxy serum CRP levels (*rs2794520*, Section 5.4.3), the results are not statistically significant at the 95% level (green lines, right panel, Figure A.9).

In the mediation analysis of genetically proxying CIDI and assessing how the CRP and BMI estimates change in a MVMR model with all three as exposures, relatively imprecise results were obtained. This is possibly due to a lack of power ($n = 58,586$ individuals with a GP record, follow-up MHQ measurements and baseline CRP and BMI, of whom 514 TRD cases). As expected, CIDI appeared to contribute to TRD in the full model (0.757 $(0.382, 1.132)$), suggesting that a unit increase in CIDI doubles the odds of TRD independently of CRP or BMI status. In this model, BMI and CRP were positively and negatively associated with TRD ($+0.533(-0.203, 1.268)$ and ($-0.093(-0.434, 0.247)$ respectively). In comparison, the corresponding estimates in the univariable model were slightly larger in magnitude ($+0.567(-0.121, 1.255)$ and $-0.127(-0.459, 0.206)$). The mediation analysis results suggest that the independent BMI and CRP effects (Figure 5.2) are not mediated by CIDI ($p_{m,CRP}$=$0.231(-3.372, 4.041)$, $p_{m,BMI}$=$0.288(-2.375, 3.221)$).

5.4.5 Other Inflammatory Mediators

Apart from CRP aberrations, there is a multitude of inflammatory mediators with distinct roles, that constitute a highly interconnected network. We use a study that investigated the genetics of a wide range of these proteins, where GWAS studies (Chapter 2.3.1) for the serum levels of $n = 41$ cytokines were reported [187]. The study used data from two large cohort studies, with individuals predominantly of Finnish ancestry, that are external to UKB (Cardiovascular Risk in Young Finns Study (YFS), FINRISK surveys (1997, 2002)). A lenient threshold ($p < 10^{-6}$) and a standard clumping procedure (distance $10,000$, correlation $r^2 < 0.01$) were applied to identify independent genetic variants that associate with each of the n cytokines. We used the major depressive disorder summary statistics of a meta-analysis of three cohort studies ($N_{Total} = 807,553$ with $246,363$ cases, $561,190$ controls) in which 102 independent variants were statistically significantly associated with depression [182]. Sample size for the inflammatory markers ranged from 840 to 8293 (Table A.8). The results indicate that three of the investigated mediators (fibroblast growth factor levels, macrophage colony-stimulating factor, cutaneous T cell-attracting chemokine) have an estimated effect on major depression (Figure A.14).

A further study with CRP as the exposure and these 41 mediators as outcomes one at a time suggests that there is an estimated nominal effect on IL-6 and GCSF; at the same time, when we consider the reverse direction and we choose instruments that are genetically predicting one of the 41 mediators, only interferon-γ appeared to decrease CRP. There is some evidence for IFN-γ suppressing CRP expression by inhibiting the production of interleukin-6 (IL-6), the key regulator of CRP synthesis in the liver. Other studies have found that IFN-γ can actually increase CRP levels

in certain settings, such as in patients with chronic obstructive pulmonary disease (COPD) or systemic lupus erythematosus (SLE). Another marginally statistically non-significant result was that of IL-18 on CRP (β (SE) 0.021 (0.009), p=0.065), for which there is also evidence for inducing CRP production.

Seeing that the set of inflammatory markers was high-dimensional, we applied the two PCA-based approaches proposed in Chapter 4. The goal was to transform the markers to groups, and then use those groups for inference. For the common MVMR instrument of $n = 530$ SNPs that associate with at least one of the inflammatory markers, low conditional instrument strengths were observed [99], ranging from 0.16 (MIPIβ) to 1.59 (IL1β). A total of $k = 14$ PCs were retained as informative, suggesting some potential for dimension reduction. None of the k PCs appeared to exert an effect on depression, at the nominal threshold and after correction.

5.5 Discussion

In this study, we aimed to investigate the causal effects of CRP and BMI on depression-related phenotypes, including TRD, using various MR methods to overcome a series of methodological issues [185, 57, 188]. We show that apparent statistically significant findings from univariable MR analyses of CRP do not persist in MVMR analyses adjusting for BMI, indicating that BMI may be the primary driver of the observed association between CRP and depression outcomes. We also found evidence for BMI exerting a positive causal effect on investigated outcomes, particularly in women, and that age may attenuate the effect on severity of depression. This evidence is subject to the limitations of BMI and CRP as imperfect but adequate proxies for adiposity and inflammation, respectively.

We find evidence further supporting the influence of BMI on depression, with larger effect sizes observed in women. An attenuation of this effect on severity of depression was also found. Although the sex difference was not statistically significant, it aligns with previous research by Tyrrell et al. [150] on the role of BMI in depression. Our findings suggest that BMI directly influences depression beyond the inflammatory consequences of overweight, addressing a limitation in prior studies. The influence of social processes, including social stigma, may play a crucial role in the relationship between BMI and depression. Our results are in line with a recent work that suggested a causal relationship between trauma and MDD, that is independent of BMI [189].

Sensitivity analyses using brain-specific and periphery-specific instruments yielded similar effects of BMI on depression measures, indirectly highlighting a social etiology that is independent of metabolic or inflammatory status. For resistance to antidepressant treatment, peripherally expressed BMI SNPs indicated a positive effect on TRD, although uncertainty in the estimates limits strong conclusions. This could be related to a differential metabolic breakdown of antidepressants in those with a periphery-driven difference in adiposity. Future works could also look into more detailed liver and kidney markers, considering their influence on antidepressant metabolism and effectiveness.

Discrepancies across different models were seen for CRP. While a positive effect on depression outcomes was found in the main analyses, this did not persist when jointly estimating BMI and CRP effects for all depression outcomes, and additionally when using cis-MR. Although conditional instrument strength is adequate in these

models [153], residual pleiotropy (i.e. other than through CRP or BMI) was still a concern. Indeed, using pleiotropy robust MVMR provided results consistent with the null. One way to address this issue is to use a stricter selection of variants based on their biologically plausibility as instruments. For example, Kappelmann et al [148] assessed the role of CRP and IL-6 in individual symptom domains of depression, using a range of instrument selections rules. Another dimension that could explain the indirect effects of inflammation is early or later-life traumatic events [189].

In the causal analyses for the reverse pathway (from depression outcomes to BMI and CRP), a significant proportion of SNPs were excluded in Steiger filtering as they appeared to be more strongly associated with CRP or BMI. Low predictive capacity of the 102 SNPs was observed for the mood questionnaire that quantified depressive symptomatology. We bypassed this issue by implementing a weak-instrument and pleiotropy-robust method [184], which suggested that there is no effect of depressive symptomatology to BMI or CRP. The estimates were in line with those in 2SLS but uncertainty was larger.

Strengths of our study include the adjustment of novel MR methods to address the issues of mediation and pleiotropy. We employed a range of cutting-edge MR methods including GRAPPLE [185] and cis-MR [57], and overcome a series of methodological issues, including the application of pleiotropy-robust methods in MVMR through extending the Collider-Correction algorithm [188]. We also assess different methods of precise mediation analyses and provide a new method that performs better in cases of sparse binary outcomes (Bayesian bootstrap).

Our study has limitations to be considered. A limitation of the BMI and CRP pheno-

types is that they do not fully capture the metabolically harmful aspects of adiposity and the concept of inflammation respectively. Although we observe similar effects when we proxy favourable and unfavourable adiposity, there still is room for improvement for the inflammatory aspect. As data on proteomics is becoming available at a large scale, more refined analyses will be feasible. Assortative mating plays a significant role in the genetic correlation of depression and various psychiatric disorders [190] and it could ostensibly also account for some of the observed effect of body mass and depression outcomes as the BMIs of partners also tends to be phenotypically correlated. This distorts heritability estimates [191] and, to a degree, the magnitude of the effects of the SNPs used as instruments. Future studies of BMI and depression could control for assortative mating by dedicated matching techniques, so as to compare individuals who are less likely to have mated non-randomly.

Chapter 6

Discussion

In this PhD thesis, I have proposed new methods for performing MR in a range of challenging scenarios. These include a method to account for pleiotropy with minimal distributional assumptions (Chapter 3, [73]), a method to apply existing pleiotropy-robust methods in multivariable MR in a single sample (Chapter 5), and a method to apply multivariable MR in high-dimensional datasets (Chapter 4, [103]). My research underlines the importance of carefully considering the MR assumptions before making strong interpretations from the available estimates. As MR is now feasible and all the tools and datasets to perform it are publicly available, methodological scrutiny is important for principled investigations of epidemiological questions. In this Chapter, I highlight the main contributions and compare them to relevant literature. I then provide a discussion of the limitations of the works and how these can guide future directions of relevant research.

In Chapter 3, I considered the case where the utilised genetic variants influence the outcome directly and in a manner that invalidates other pleiotropy-robust approaches. At present, MR Egger [67] requires the InSIDE assumption to be satisfied (pleiotropic effects independent of instrument strength), and MR Median and MR

Mode require a subset of SNPs to be completely 'valid' (zero pleiotropy) . We replace the InSIDE assumption by stipulating that SNPs are sex specific (i.e. there is a gene-sex interaction that is apparent for the exposure). We show that this interaction-based method (Chapter 3) performs well even if all SNPs are pleiotropic and violate InSIDE by showing that each pleiotropic effect can be cancelled out. Our work builds on that of Spiller et al's MR-GxE method [75], but with some important differences. In comparison with the GxE method, we focus on biological sex rather than more composite, later-life variables (such as deprivation indices) which is less likely to pose issues in justifying the data generating models. For example, biological sex determination causally precedes the exposure and the outcome, whereas socioeconomic status may not. This protects against possible collider bias if the direction of causality is reversed in the real data.

In addition, while the desirable pleiotropy-cancelling effect is stronger for phenotypes with pronounced sex specificity, I have explicitly worked throughout to accommodate weaker sex-specific phenotypes that would otherwise induce weak instrument bias. The current published work [73] includes a thorough applied analysis with the flagship sex-specific phenotype of waist-to-hip ratio; another markedly sex-specific phenotype that could also be used is serum levels of testosterone [192]. An essential future step to show the utility of the method would be to apply it more broadly in other applied instances. A work in progress is the application of the method to assess how grip strength affects the risk of falls.

While biological sex is well suited for reasoning a solid interaction variable, the concept of cancellation is general and can include any binary interacting variable. An interesting example that could serve as such a variable and I have made preliminary

investigations on is the regulatory elements in the genome [193]. Since regulatory genes are assorted in even earlier developmental stages, the same argument of temporal precedence that is made for sex also holds. An example that I've investigated is the *FTO* gene and a regulatory element (*NR3C1*) that influences the levels of expression of *FTO* mRNA and FTO protein. The underlying assumption was that, if we can identify one variant (or a few as a GRS) that substantially affects the levels of NR3C1, then this might in turn affect FTO expression. Thus, two strata of a distant genotype, a fixed variable akin to sex in Chapter 3, would show different functionalities of FTO and possibly differential associations with BMI. While it is feasible to identify such influential variants with dedicated transcription factor databases, this applied example did not yield statistically significantly different FTO-BMI associations in people with different *NR3C1* levels. In the future, I plan on expanding the search to include many possible regulatory elements and describe the method for gene-by-gene interactions.

PCA: Through the PCA work on using dimensionality reduction with MR (Chapter 4), the major contribution is providing a reliable method that works in the challenging setting of many correlated exposures. I present a detailed investigation into first how exactly the ability of genetic variants to discriminate among exposures severely attenuates when many highly correlated exposures are considered (lack of conditional instrument strength), and how we can rescue this by transforming the exposures into blocks of exposures. The published paper contains a comprehensive report of a wide range of simulation scenarios, and I also provide convenient R functions on a GitHub repository [103]. The central message of this study is that sparse PCA methods outperform conventional multivariable MR when used for this purpose. Sparse methods

enable us to populate the resulting principal components, with with fewer mutually exclusive sub-groups of exposures, making them much more interpretable [108].

Compared with works that tackle similar problems that were published while the work was in progress, our approach has advantages. A study that decomposes anthropometric measurements to independent principal components that drive obesity follows a similar approach and reports differing effects in cardiovascular outcomes [194]. One theoretical advantage of our approach is that the observed overlap of contribution of certain traits, such as height and BMI, to many PCs could be avoided with sparsity and inference on what drives the MR results could be more readily attributable to the remainder of the contributors.

There are certain extensions to be considered. In the published version of the paper, we focus on a scenario where the underlying data generating mechanism does not include any other pleiotropic pathways. Nevertheless, it would be meaningful to explore the use of pleiotropy-robust methods in conjunction with dimensionality reduction approaches in future works. Despite the high dimensionality of such NMR or imaging datasets, they are often focused on one metabolic pathway or anatomical structure and hypothetically some diffusely pleiotropic variants [94] could affect other pathways. In its current form, the method implements an inverse-variance weighted meta-analysis which would be susceptible to pleiotropy (Chapter 2.4). A straightforward extension would be to use one of the available pleiotropy-robust methods [67, 69, 129] based on summary data to replace the current IVW method. However, without further investigation, it is unclear how the pleiotropy itself is transformed and how this affects the estimates. Another area of improvement would be to amend the existing approach in order to further improve the interpretability of the estimates;

currently, the exact effect size is transformed and the recommended area of application is hypothesis testing. We believe this is related to the well-described issue of the interpretability of principal component regression, where the transformation of the data leads to an unclear interpretation for the magnitude of any identified associations [195].

In Chapter 5, I investigated the independent and dependent causal effects of CRP and BMI across a range of depression phenotypes. The key finding was the attenuation of the direct CRP effects in joint models compared to their estimated total effects. These findings replicated previous results on BMI affecting mood. I also provide further evidence on the stronger effect of BMI on mood in women. I utilized a series of cutting-edge MR models to support these findings and enabled two-sample pleiotropy-robust methodology such as MR-GRAPPLE [185] to be applied by generalising the technique of Collider Correction [85] to multivariable MR. This allowed for more flexibility than the one-sample MR framework allows, including stratification by sex and age, together with the robustness of two-sample MR methods. The key applied message from the published study is that the effect of inflammation reported in previous studies is likely to be, at least to some degree, attributable to BMI. However, there are some limitations to the study. Inflammation is inherently a vague concept, and CRP is not a perfect instrument, having been criticised for its inter-individual variability across time, and for lack of sensitivity and specificity [196]. Thus, a more granular view of the immune system is warranted, particularly since the target of the analyses is a type of inflammation that is chronic and low-grade in nature. Preliminary analyses performed with a dataset of 41 immune mediators (Chapter 5.4.5,

[187]) were underpowered, and crucially, there was a lack of conditionally strong instruments. Nonetheless, an interesting result was a positive effect of FGF levels on depression; this protein belongs to a family of growth factors that are crucial in embryological development in many structures [197], including the neural ectoderm. Schematically, the prior stage of ectoderm is stopped from developing to neural ectoderm by Bone Morphogenetic Protein (BMP) signaling. In early animal models, FGF signaling inhibits this BMP inhibitory effect and there is in vitro studies providing a more nuanced picture of exactly how these events shape neural development [198]. In summary, future work should address these limitations and aim to provide a more comprehensive view of the role of inflammation in mood disorders.

Another limitation to be considered is the inherent disadvantage of MR in targeting exposures that are limited in time at one point or period in life. The estimated causal effect is more accurately interpreted as how a lifelong propensity to an increased level of the exposure (in our case, body weight and inflammation) affects the outcome (depression). In recent observational studies, it has been discussed whether the acute phase of mental illness has an inflammatory imprint, with variable findings [199]. If the effect of inflammation is temporally located around the depressive episodes or is preceding it, as has been investigated but not reliably shown by works on predictive modelling [200], then this may be diluted in the lifecourse MR effects that we are reporting. These types of effects are more likely to be captured by different experimental designs; for instance, precipitation of pro-inflammatory conditions in various patient groups has been shown to elicit fatigue and low mood [201]. There also exist approaches that temporally dissect an exposure to early- and later-

life components and report their direct effects through multivariable MR [202], but such models make strong assumptions and are not uniformly accepted.

The main scientific contribution of the CRP and BMI work is that it provides another line of evidence suggesting an effect of body shape on mood to add to the previous evidence base. It appears that the primary driver of the observed association between CRP and depression outcomes is body mass, in line with other works that have investigated this question, looking in particular in how reported trauma [148] and interleukin-6 [189] modulate the associations. There is evidence for reported trauma amplifying the genetic heritability of MDD, indicating a stronger influence of genetic factors in individuals exposed to trauma compared to those unexposed. Evidence from meta-analyses of observational studies suggests that the BMI positive effect on depression outcomes is more pronounced in women [136]. We find that age may attenuate the effect of BMI on depression severity. To contextualise this, we compared it with an estimate of a meta-analysis of longitudinal studies that suggested that baseline overweight was associated with depression in subjects 20 years or older, but not in younger individuals [136]. This is complementary to our result since not many individuals in childhood and adolescence are represented in UKB where our sample comes from. Adolescence is an important developmental milestone, containing a range of important life events such as transition to self-sufficiency as well as a distinct neurological imprint with events such as changes in cortical myelination, and the first manifestation of the major psychiatric diagnoses is condensed in this period [203]. What later large meta-analyses indicate is an agreement in a harmful effect of overweight and obesity on depressive mood, with no clear moderation of the effect in different early life age groups [203]; what our study then adds

is a more granular view of later life and how the effect of BMI on diagnosed clinical depression as well as most mood questionnaires seems to be stable. The exception is the severity of depression, with younger individuals experiencing a stronger effect of BMI. Furthermore, we observe a lowering of the predictive capability of the BMI instrument in higher age groups [100]; this is in line with the general concept of an aggregation of environmental exposures during the lifecourse, giving non-genetic factors time to accrue and exert their effects on BMI and direct metabolism and weight away from the genetically predicted value; a 2012 meta-analysis of 88 independent estimates of BMI heritability, including 140,525 twins and 42,968 family members, found that one of the factors that influenced heritability heterogeneity was the mean age in a range of studies assessing individuals of 10 to 67 years of age [204]. A non-linear pattern was observed, with increasing heritability values as adolescence progresses and decreasing values afterwards throughout adulthood. This is in accordance with our finding on a decreasing proportion of BMI variance explained in older age groups, but not for CRP. It can be conceived that CRP is a less complex trait and therefore there may be a different age-dependent variability in its genetic component than BMI. Additionally, the causes of an increased CRP, such as autoimmune conditions and infections are occuring in later life and they may provide the contributing conditions required for a high CRP.

Our mediation analysis indicates that the apparent positive effect of CRP levels on depression is predominantly mediated by BMI in women. We also replicate the results of a previous work on FA and UFA [139] and find comparable effects of the two measures on depression, adding the angle of a careful look at the inflammation aspect.

There is one additional project that has not been presented as a standalone chapter but was a learning experience and I will be describing in this paragraph. Trying to mitigate weak instrument bias in multivariable MR analyses, I explored selection algorithms as a possible solution. Due to widespread effects of SNPs on multiple phenotypes, multivariable models where two exposures share a large part of their genetic composition could lead to low discriminatory ability of the instrument for each exposure, leading to substantial and imprecise estimates of effects. Based on observations that excluding some of the SNPs in joint MVMR models appeared to improve the conditional $F-$statistic [99], I aimed to investigate how selection algorithms could be used to improve conditional instrument strength. Progress in this was presented as preliminary findings in the 2021 Mendelian randomisation Conference in Bristol [104]. After describing and developing software with algorithms that explore the space of possible SNP combinations more efficiently than other approaches, and validating this performance in simulation studies and real data examples on inflammation, BMI, and depression, it was clear that there exist SNP subsets that improve conditional instrument strength. However, there was a drastic decrease in accuracy, with biased results in the models that leveraged those subsets in place of the original set. A range of follow-up amendments, such as stopping the iterative procedure early, did not salvage this bias. We believe that this is related to the well-described effect of the selection process itself causing bias. While there exist advanced methods to include this selection procedure in a more encompassing modelling framework [205], given the labour-intensive nature of the project and the very modest returns observed, we made the decision to not pursue this line any further, instead focusing on the other works presented here that showed more consistent

results and in the end led to publications. In the second year of my studies, I revisited the project and applied a post-selection correction algorithm in one applied example involving the relationship between depression, BMI, and type 2 diabetes (positive control outcome). Although this approach yielded only a minimal increase in conditional instrument strength and provided sensible causal effects, it is not validated by simulation studies in a range of settings as the other works in the included Chapters. Therefore, we chose to briefly describe the research process here and reflect on the experience. This taught me several lessons. First, it became clear how important clear description of interim results is, and how starting projects with simple scenarios is a prerequisite for progressing further. Most importantly, having the courage to abandon projects that offer only marginal or no meaningful returns. While the exploration of selection algorithms for weak instruments in MR showed promise, the challenges faced and the limited improvements observed guided us to pursue alternative avenues for future research.

6.1 Summary

In this thesis, I have developed cutting edge methods for Mendelian randomisation to extend its reach to settings where the MR assumptions do not hold. I first show how we can effectively address pleiotropy by leveraging gene-by-sex interactions in the exposure to cancel it out. Building on multivariable MR, I have provided a thorough examination of ways to appraise high-dimensional data sets, using established sparse PCA methods to transform them to meaningful groups rather than analyse them individually. Through simulation studies and positive control analyses, biologically informative groups can be identified and the estimated effects are salvaging

power at the cost of estimation. In my applied work, I examined the impact of BMI and CRP on depression outcomes with a multitude of causal inference approaches. Apparent influences of CRP on a range of clinical and self-reported depression measures were challenged, and the mediating role of BMI was underlined. I also showed that methodological amendments of existing methods can obtain more sensible results in a parallel manner to biologically informed exclusions. These findings replicate previous notions on the understanding of the common sources of inflammation and body weight and how these influence mood and clinical diagnosis of depression, and provide suggestions on how to model such relationships in MR. In conclusion, I propose novel methods in pleiotropy-robust MR and multivariable Mendelian randomisation, showing their applications in investigating the relationships between metabolic health and depression.

Appendix A

Appendix

A.1 Sex-Stratified MR

A.1.1 Stratified F-statistic

As described in Section 3.3.3, we expect that the F_{strat} statistic is to be interpreted in a similar manner as the $F-$statistic in one-sample and two-sample MR. To test this hypothesis, we perform a simulation study to assess the dilution bias as predicted by F_{strat} and the observed estimate. The stratified IVW method is used for the latter since it is not robust to this weak-instrument bias and therefore this method is where dilution would manifest. An alignment of this stratified IVW estimate with a predicted value biased towards the null by a factor of $\frac{\hat{F}_{strat}-1}{\hat{F}_{strat}}$ would hence indicate that F_{strat} is a good predictor for interaction strength (section 3.3.3). As can be observed in Figure A.1, the predicted mean estimate is in close agreement with the observed estimates from the stratified IVW method.

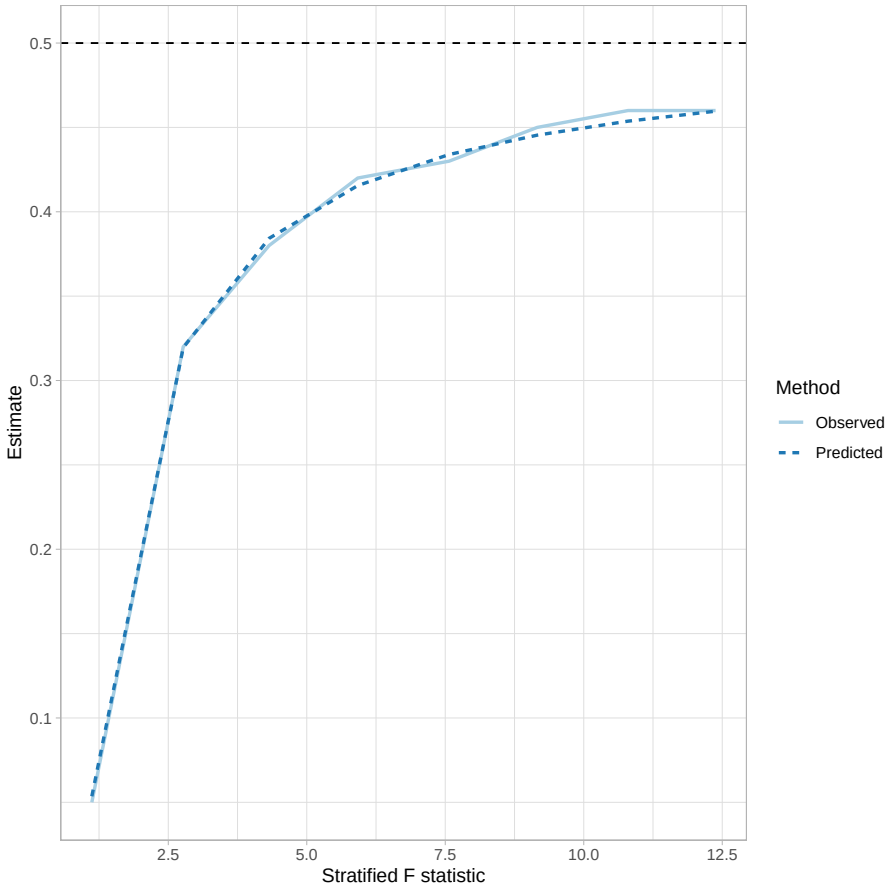


Figure A.1: dashed black line: causal effect of X on Y ($\beta = 0.5$); Observed: mean estimate of the causal effect across all simulations $\bar{\hat{\beta}}_{IVWstrat}$; Predicted: $\beta \times \frac{\hat{F}_{strat}-1}{\hat{F}_{strat}}$.

A.1.2 Plots of SNP-WHR - SNP-Y^* associations

We present a scatter plot of the differences in the sex-specific SNP-WHR estimates γ_{j1} and γ_{j0} against the respective collider-biased SNP-Y estimates ($\hat{\alpha_1}^*$ and $\hat{\alpha_0}^*$ in Eq. 3.8). The slope of the line represents the biased causal effect estimate of WHR on the outcome Y, which is then used to correct the observational association ($\widehat{\beta - \beta^*}$, Section 3.3.3). For ease of interpreation, all SNP-WHR association estimates are reoriented so that the WHR-raising allele is the reference allele.

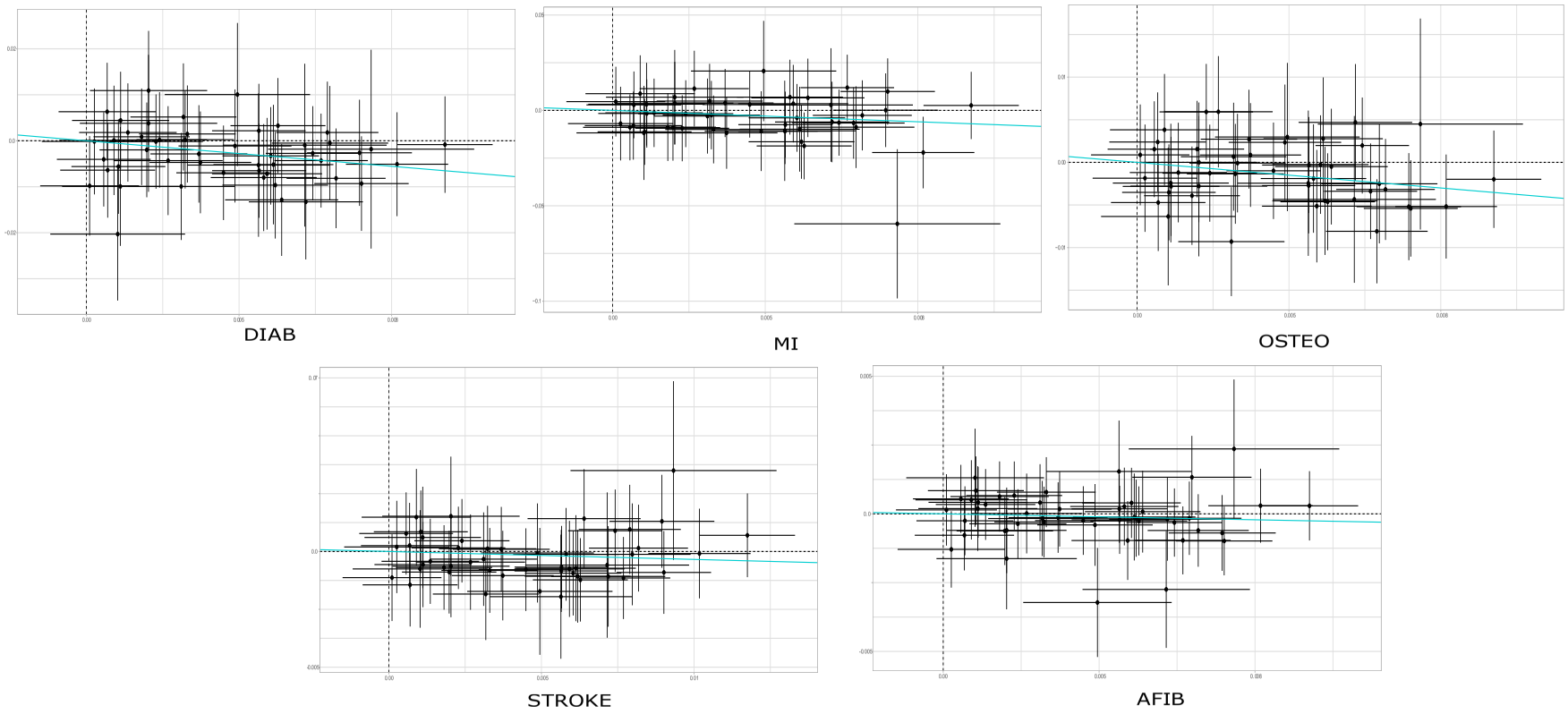


Figure A.2: Binary outcomes. Scatter plot of summary estimates, with the differences in the sex-specific estimates for BMI in the y axis. Blue line: collider-biased estimate.

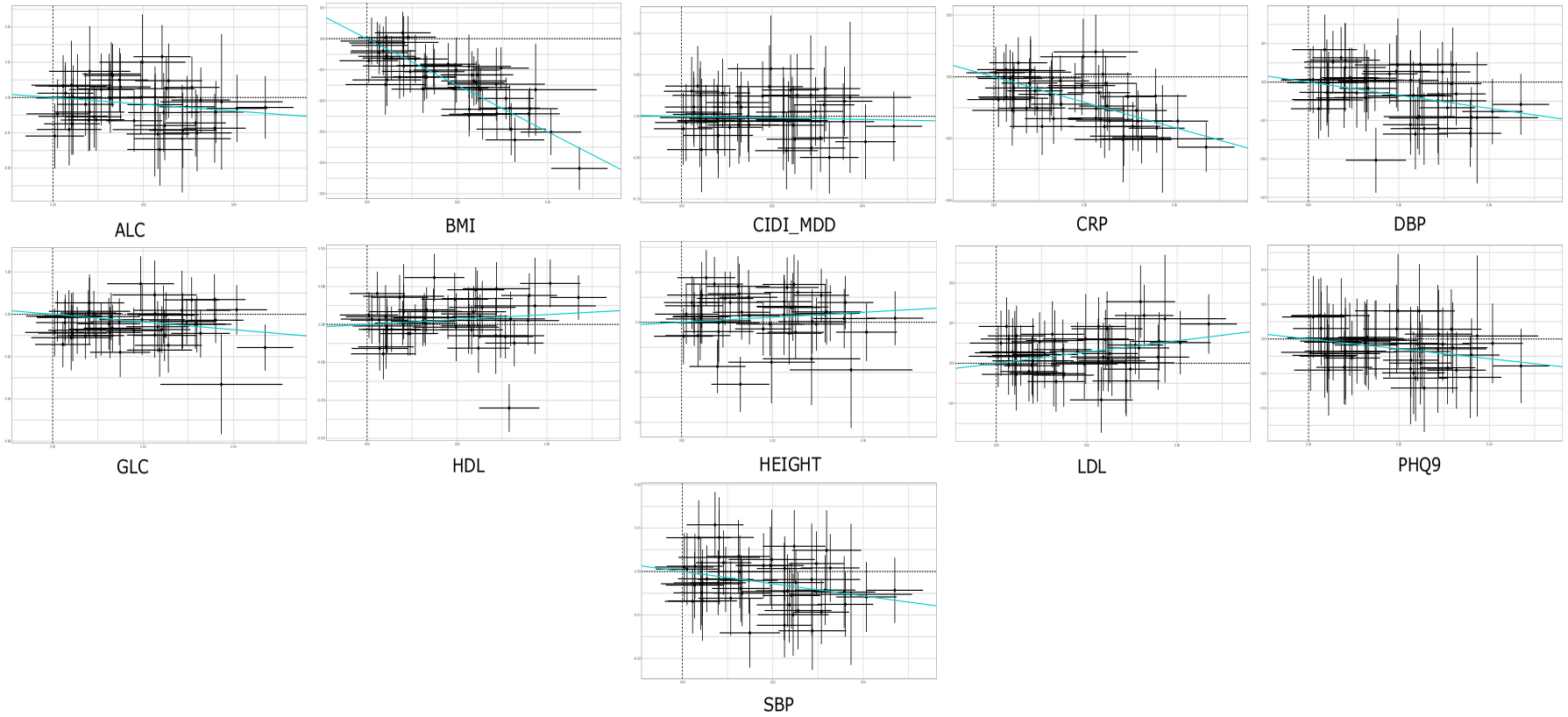


Figure A.3: Continuous outcomes. Blue line: collider-biased estimate.

A.2 Real Data Applications

PHQ-9 Items	CIDI-SF Items	Description
1. Little interest or pleasure in doing things	Loss of interest in activities Decreased pleasure in daily activities	Lack of enjoyment in activities Reduced enjoyment in daily activities
2. Feeling down, depressed, or hopeless	Depressed mood Feeling hopeless about the future	Feeling sad or low in mood Pessimistic outlook on the future
3. Trouble falling asleep, staying asleep, or sleeping too much	Sleep disturbances	Difficulties with sleep patterns
4. Feeling tired or having little energy	Fatigue or loss of energy	Lack of energy or fatigue
5. Poor appetite or overeating	Appetite changes	Changes in eating habits
6. Feeling bad about yourself or that you are a failure or have let yourself or your family down	Low self-esteem or feelings of worthlessness Self-critical thoughts Guilt or self-blame	Negative self-perception or low self-worth Critical thoughts about oneself Feelings of guilt or self-blame
7. Trouble concentrating on things	Difficulty concentrating Difficulty making decisions	Problems with concentration Challenges in decision-making
8. Moving or speaking slowly or being fidgety or restless	Psychomotor changes Restlessness or feeling slowed down	Changes in motor activity Restlessness or slowed movements
9. Thoughts of being better off dead or of hurting yourself	Suicidal thoughts Suicidal ideation	Thoughts of self-harm or suicide Ideas of ending one's life

Table A.1: Comparison of the two Mental Health Questionnaire Items sent out to UK Biobank participants [160]. PHQ-9: Patient Health Questionnaire-9; CIDI-SF: Composite International Diagnostic Interview short-form.

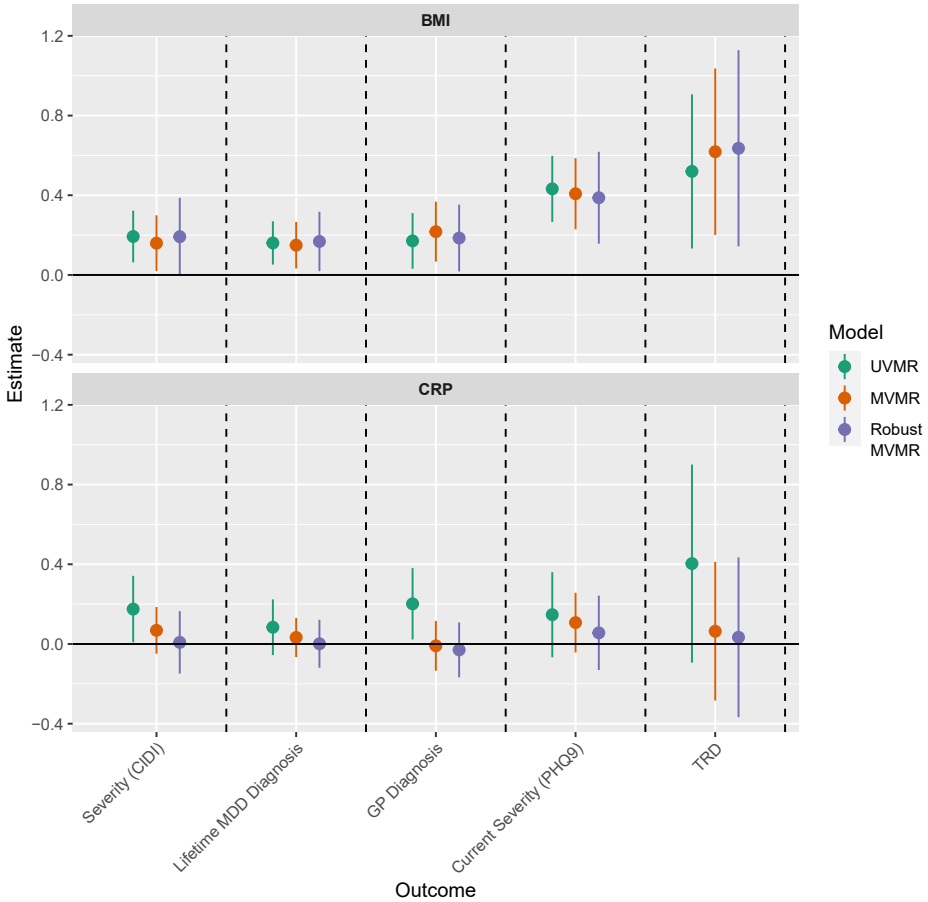


Figure A.4: Estimated effects of CRP and BMI on a range of depression-related outcomes in a subset of unrelated individuals (PHQ9 and CIDI $n = 52,510$, GP Diagnosis of MDD and TRD $n = 165,378$.)

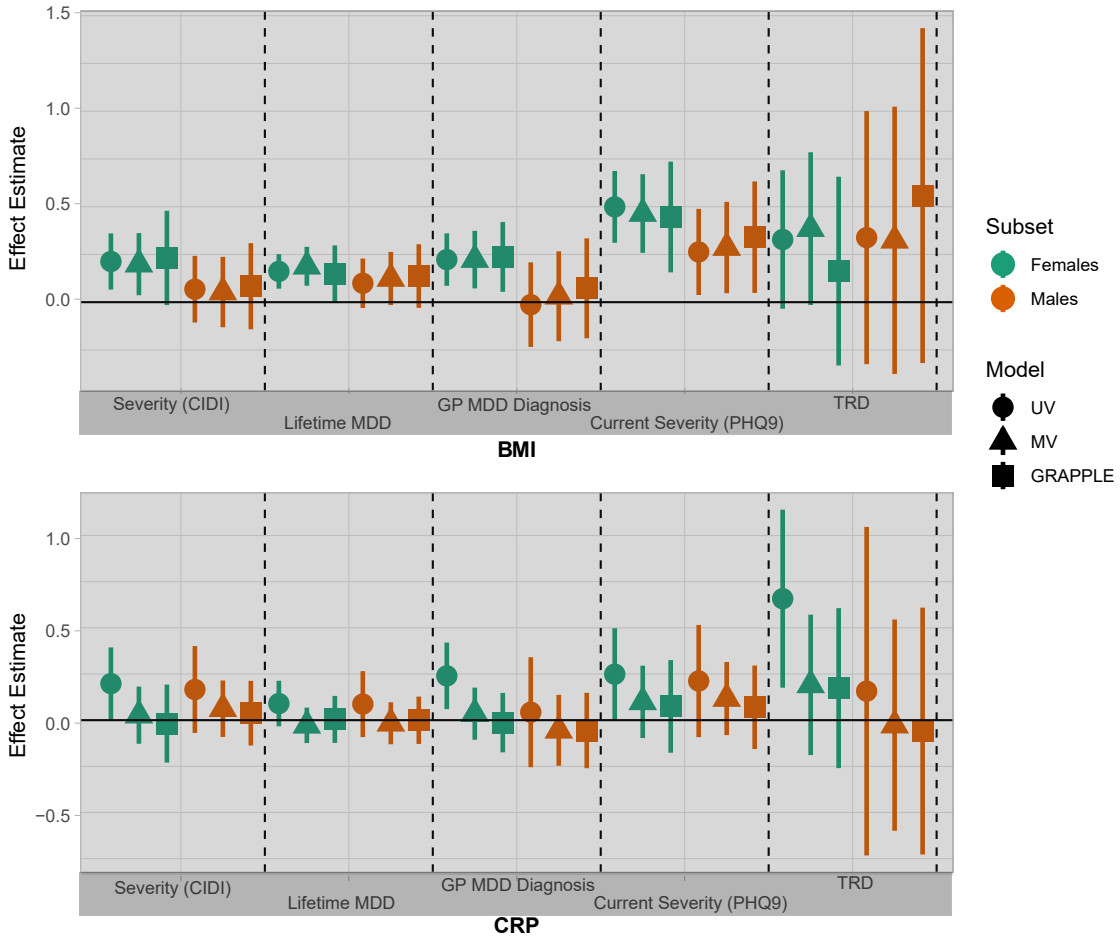


Figure A.5: Sex-Stratified Analysis for all outcomes reported in Figure 5.2. Estimates from univariable MR (UV), multivariable (MV), and pleiotropy-robust multivariable MR (GRAPPLE) are reported for females and males separately.

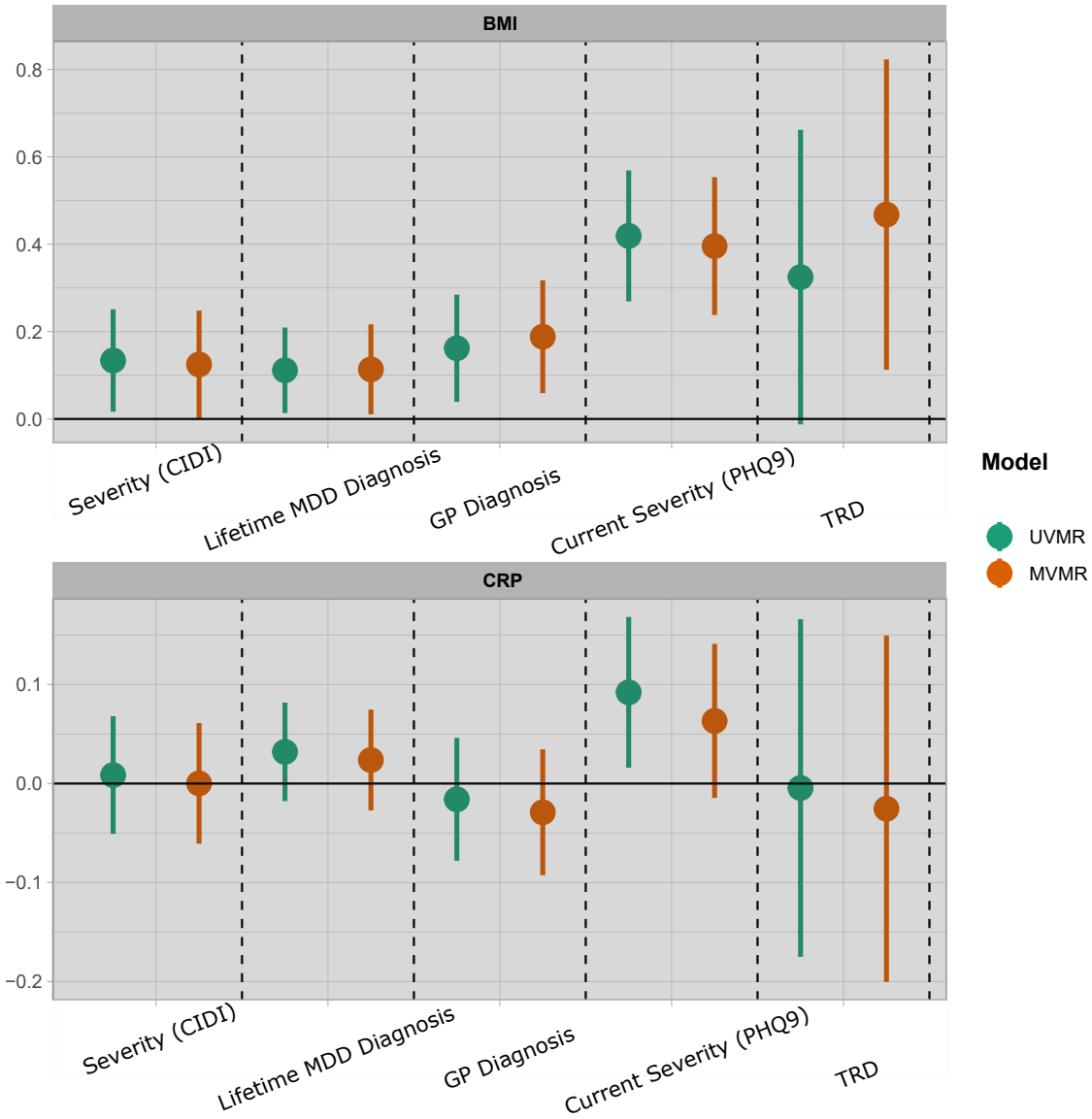


Figure A.6: MR Analysis with external weights for BMI [167] and CRP[166]

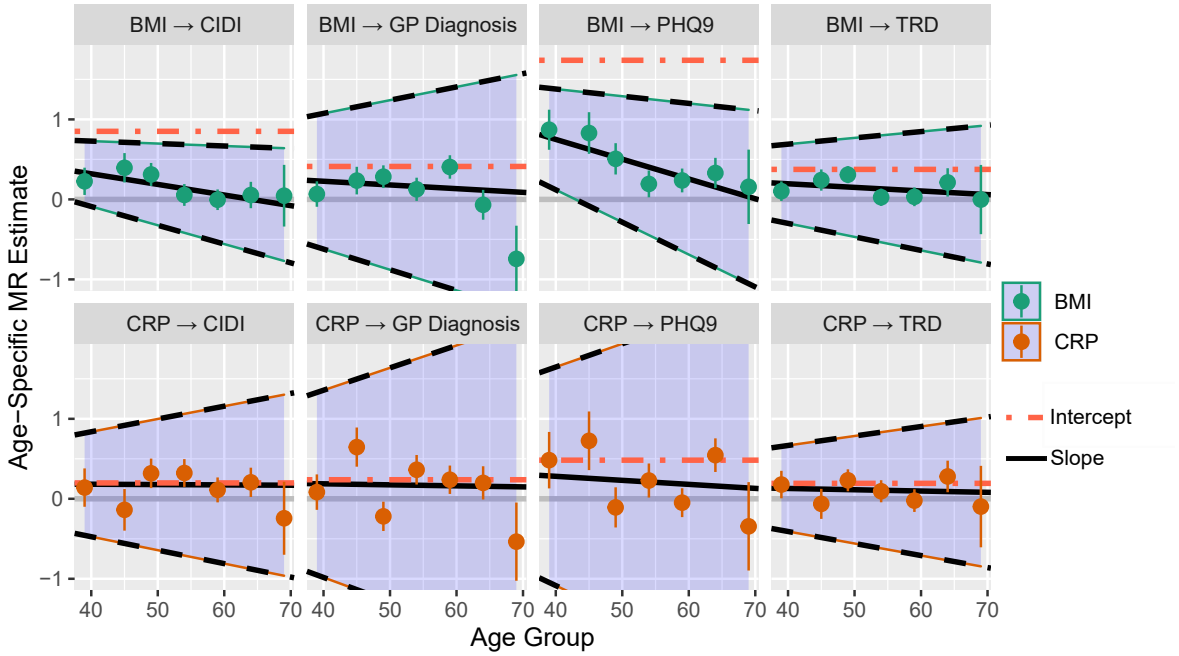


Figure A.7: Age as a Moderator of the causal effects of CRP and BMI with mood outcomes. In the visualisation of the meta-regression slope, if the intercept falls within the confidence region of the age slope, then the result is not statistically significant.

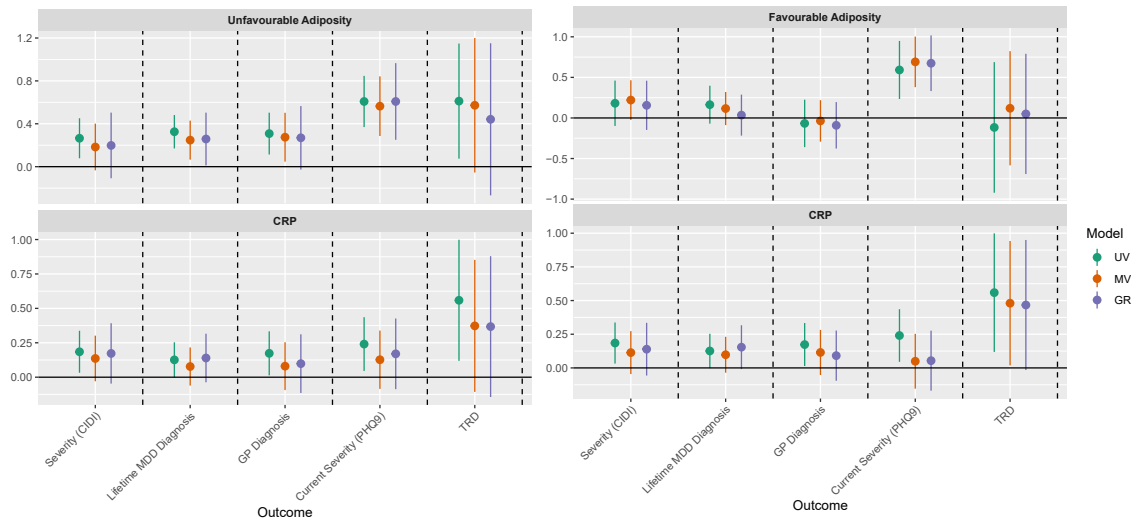


Figure A.8: Forest Plot for the Effect Estimates of Favourable and Unfavourable Adiposity and CRP on Depression Outcomes. CIDI: Composite International Diagnostic Interview; MDD: Major Depressive Disorder; PHQ9: Patient Health Questionnaire-9; TRD: treatment-resistant depression.

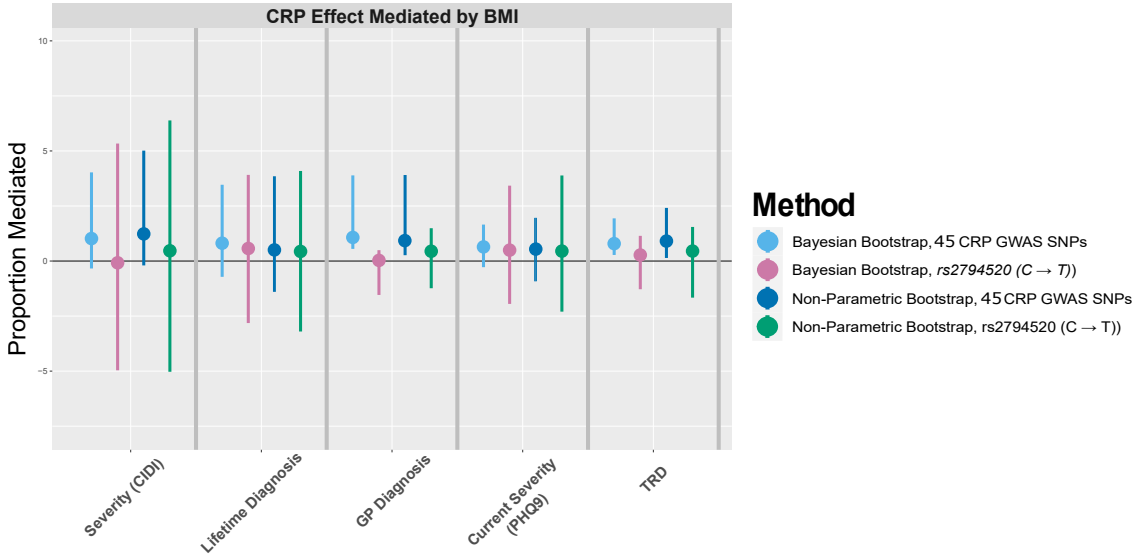


Figure A.9: Proportion of CRP effect mediated by BMI, defining the CRP effect with two different instruments (45 CRP SNPs, rs2794520 (C → T)) in the *CRP* region). Two methods of bootstrapping are used to estimate the uncertainty (Bayesian bootstrap, non-parametric bootstrap). CIDI: Composite International Diagnostic Interview; MDD: Major Depressive Disorder; PHQ9: Patient Health Questionnaire-9; TRD: treatment-resistant depression.

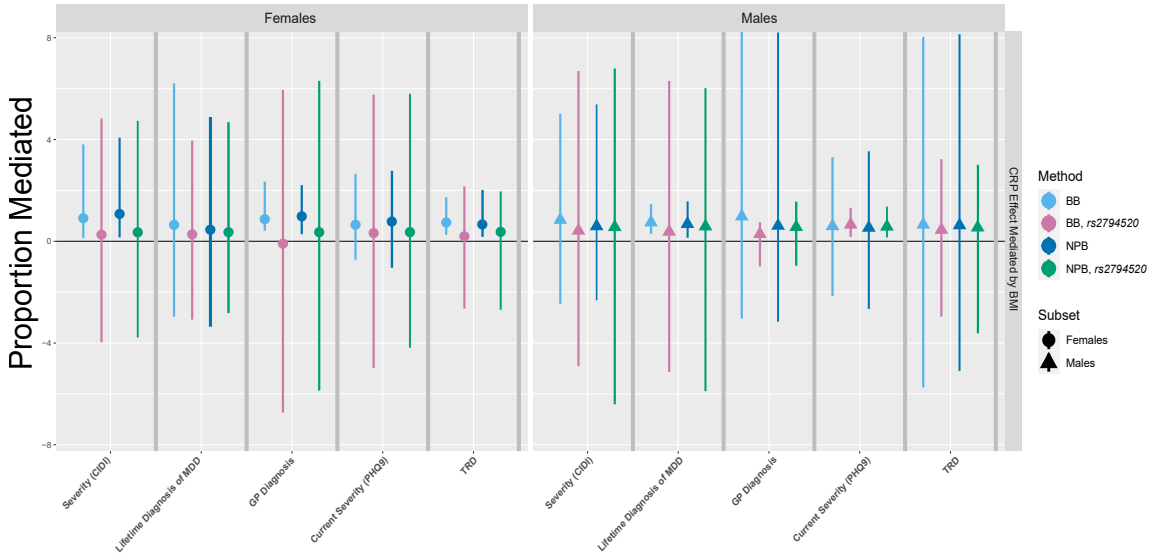


Figure A.10: Proportion of CRP effect mediated by BMI in males and females. The CRP effect is defined by two different instruments. cisMR: CRP effect is estimated with one SNP as instrument (rs2794520).

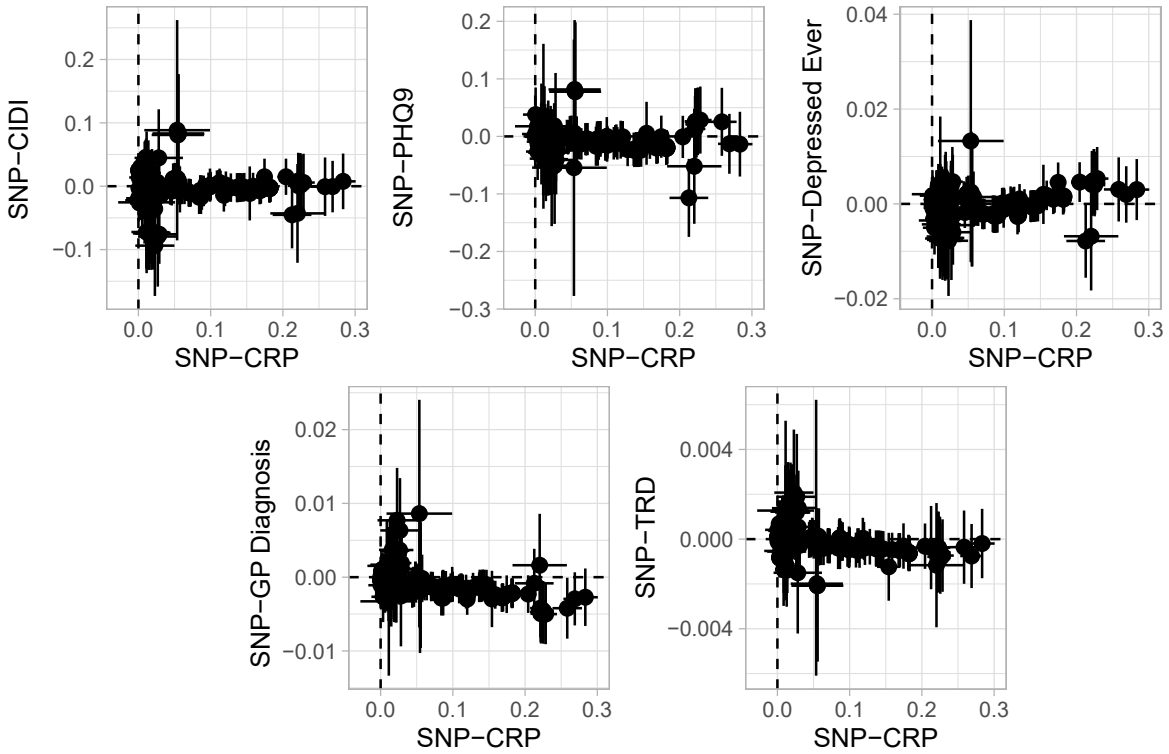


Figure A.11: SNP-CRP associations and SNP-depression associations for $n = 194$ SNPs in LD. These genetic associations are then projected to independent genetic components (cis-MR, [57]).

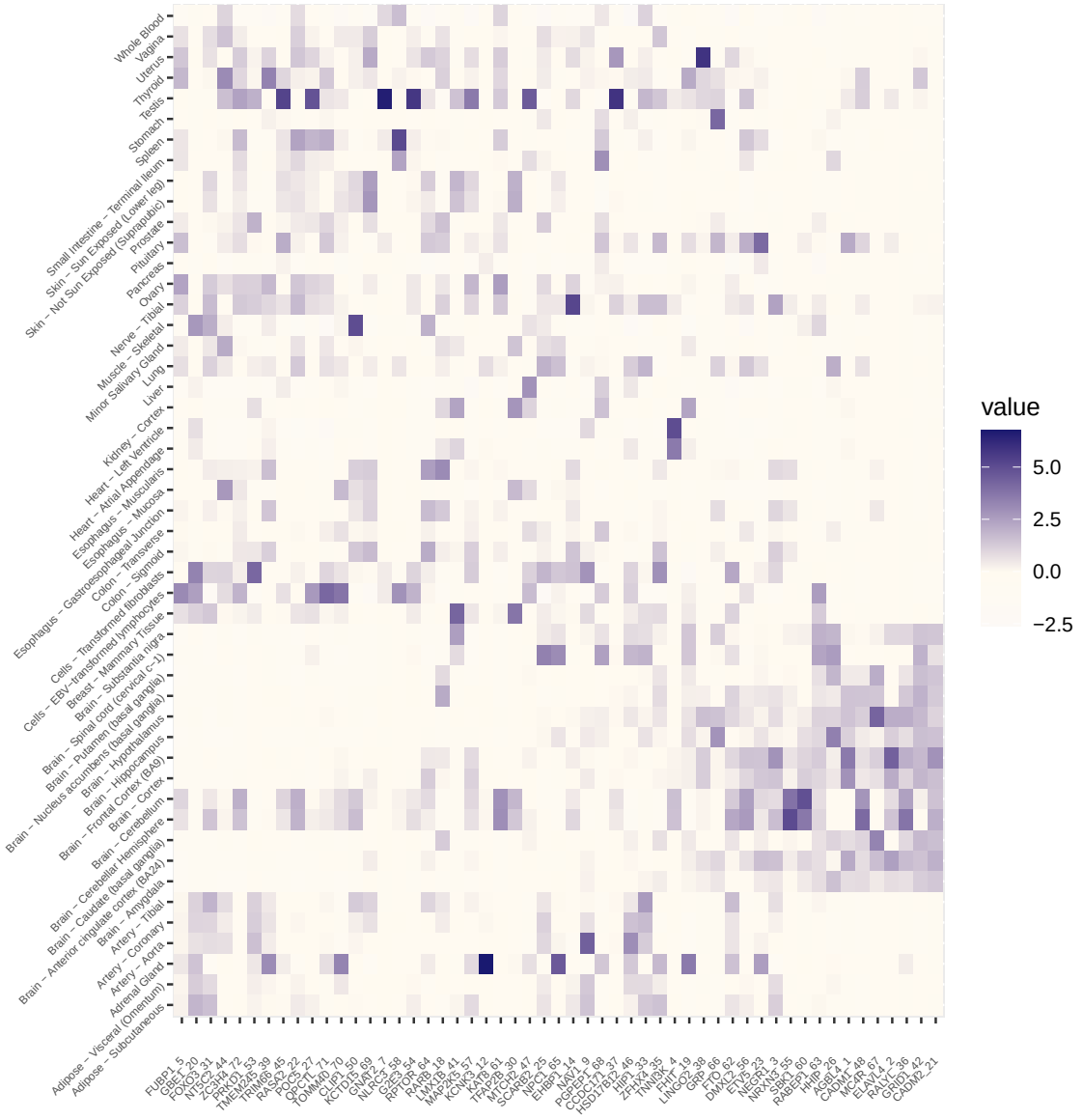


Figure A.12: Tissue Expression for SNPs in Locke et al. Transcript per million (TPM) data, scaled per gene, are presented. Ordering follows the sum of scaled TPM across brain regions.

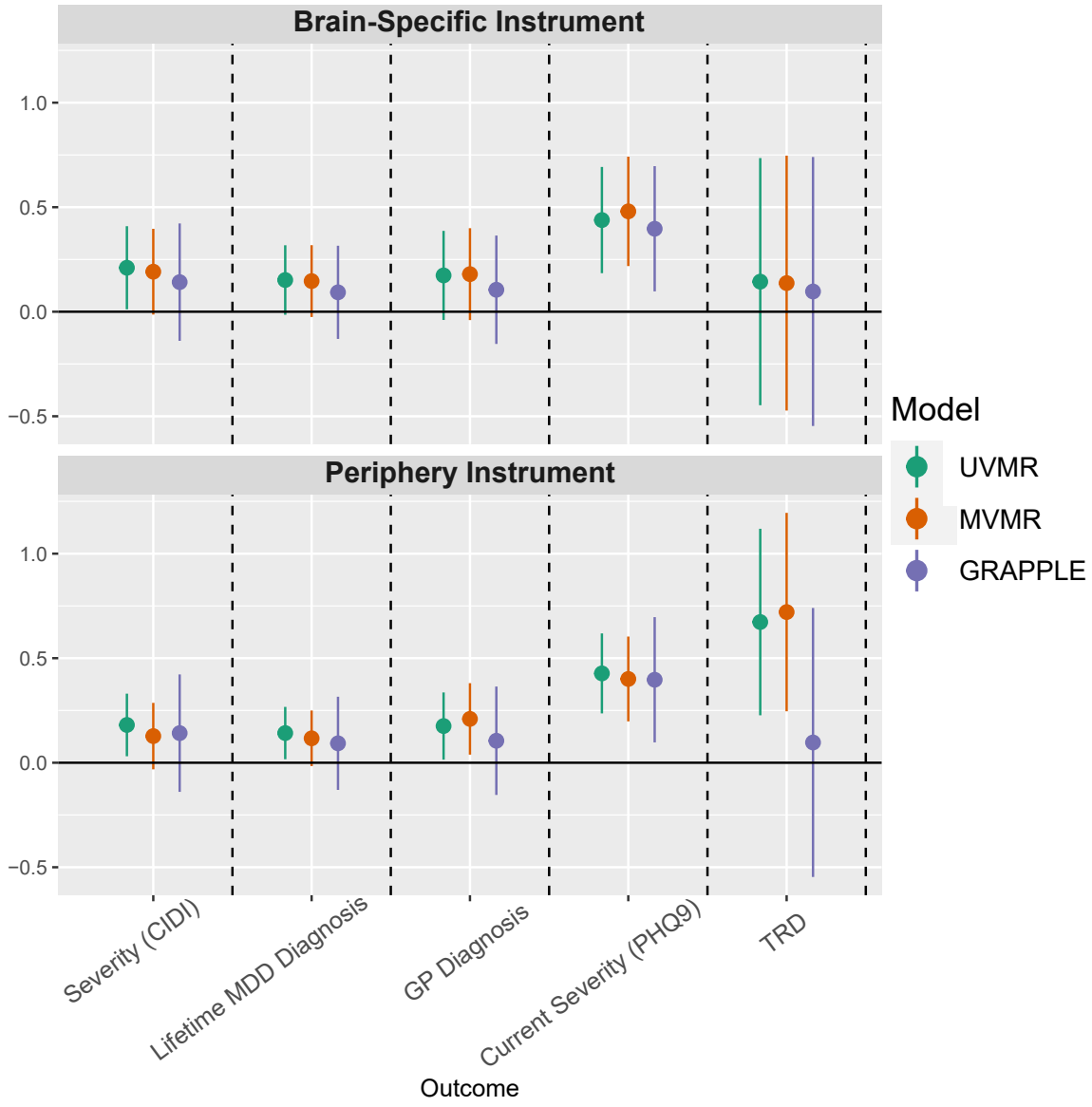


Figure A.13: Estimates of the Effect of BMI on the depression outcomes when two different tissue expression-informed instruments are used. In the top panel, the top 20 SNPs of genes that are predominantly expressed in the brain are shown (Brain); in the bottom panel, genes that are expressed in the periphery constitute the instrument.

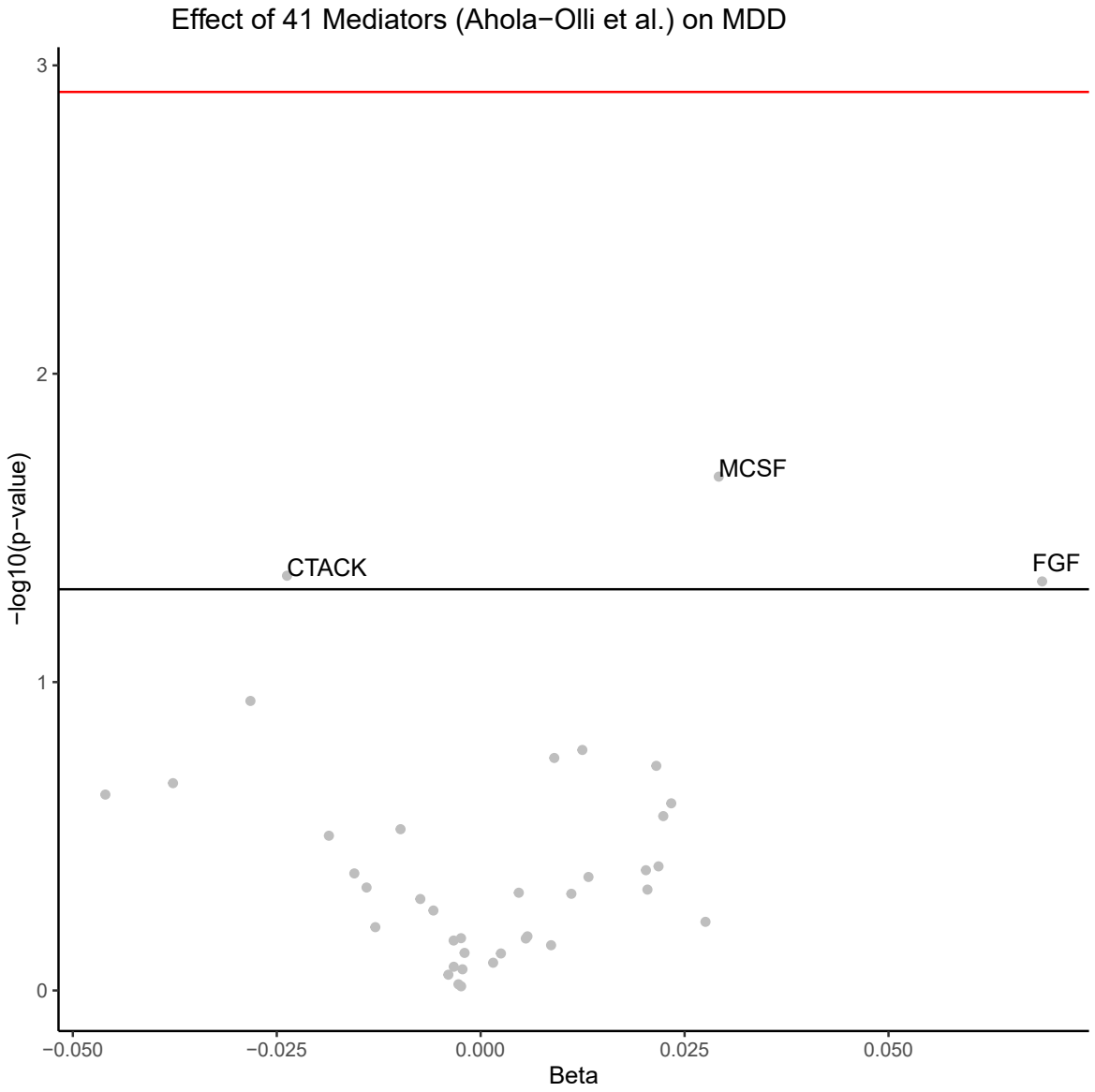


Figure A.14: Volcano plot for the two-sample MR estimates (inverse variance weighted, IVW) of the effect of 41 inflammatory mediators [187] on major depressive disorder [182]. Horizontal axis: effect size; vertical axis: -$\log_{10}$p-value. CTACK: cutaneous T cell-attracting chemokine levels; FGF: fibroblast growth factor levels; MCSF: Macrophage colony-stimulating factor.

		CRP GWAS			CRP rs2794520			BMI			
Age Group	**N**	***F***	**CFS**	R^2	***F***	**CFS**	R^2	***F***	R^2	**CFS, rs2794520**	**CFS, CRP GWAS**
38-45	43,191	59.51	24.06	0.065	572.17	14.69	0.014	13.05	0.021	10.76	7.609
45-49	43,738	60.77	25.13	0.065	505.94	13.862	0.012	10.75	0.017	9.197	6.48
49-54	64,289	88.31	36.11	0.065	715.76	20.173	0.012	16.38	0.018	13.782	9.696
54-59	76,918	104.54	43.19	0.064	1014.21	26.006	0.014	16.94	0.015	14.355	10.171
59-64	108,124	154.10	61.55	0.067	1453.36	37.431	0.014	20.71	0.013	17.864	12.375
64-69	90,975	143.60	57.14	0.074	1228.65	33.846	0.014	15.63	0.012	13.776	9.533
69-75	16,357	26.58	11.18	0.076	263.45	7.266	0.017	3.72	0.016	3.324	2.592

Table A.2: Instrument Strength (Mean $F-$statistic (*F*), conditional *F*-statistic (CFS), R^2) for each age group. N: sample size for each group.

Outcome	Model	BMI Fisher's z	BMI p value	CRP Fisher's z	CRP p value
TRD	UVMR	0.0038	0.979	1.0051	0.332
TRD	MVMR	0.1481	0.882	0.5511	0.534
TRD	GRAPPLE	0.782	0.441	0.5237	0.567
PHQ9	UVMR	1.3176	0.115	0.1597	0.854
PHQ9	MVMR	1.2152	0.269	0.1541	0.899
PHQ9	GRAPPLE	0.5038	0.615	0.0273	0.978
Ever Depressed	UVMR	0.7316	0.441	0.0274	0.978
Ever Depressed	MVMR	0.7348	0.467	0.2168	0.885
Ever Depressed	GRAPPLE	0.0835	0.917	0.0054	0.96
GP Diagnosis	UVMR	1.6839	0.076	1.2134	0.268
GP Diagnosis	MVMR	1.3196	0.178	0.7387	0.459
GP Diagnosis	GRAPPLE	1.0069	0.313	0.3348	0.747
CIDI	UVMR	1.1051	0.213	0.2325	0.838
CIDI	MVMR	1.2101	0.244	0.3163	0.753
CIDI	GRAPPLE	0.7752	0.385	0.4188	0.687

Table A.3: Results for Fisher's z statistic to assess sex specificity of CRP & BMI estimates on mood outcomes.

	CRP	BMI	CRP & BMI MVMR (45 GWAS hits)	CRP & BMI MVMR (rs2794520)
Severity (CIDI)	476.07, 0	1097.50, 0	1155.62, 0	1102.74, 0
Current Severity (PHQ9)	1356.22, 0	2891.75, 0	3007.66, 0	2962.58, 0
GP Diagnosis of Depression	350.86, 0	92.48, 0.0372	703.59, 0	675.73, 0
TRD	162.73, 0	368.44, 0	413.20, 0	386.74, 0
Lifetime MDD Diagnosis	332.43, 0	700.57, 0	765.54, 0	715.02, 0

Table A.4: Sargan Test for heterogeneity.

Exposure	Outcome	β_{age^2}	**SE**	**p-value**
CRP	PHQ9	-0.0979	0.1825	0.5915
BMI	PHQ9	**-0.2240**	**0.1089**	**0.0397**
CRP	TRD	-0.0047	0.0992	0.9629
BMI	TRD	-0.0602	0.0920	0.5127
CRP	GP Diagnosis	-0.1147	0.1713	0.5030
BMI	GP Diagnosis	-0.1574	0.0999	0.1152
CRP	CIDI	-0.0325	0.1060	0.7592
BMI	CIDI	-0.1444	0.0867	0.0953

Table A.5: Age as a Moderator of the causal effects of CRP and BMI with mood outcomes, results for a quadratic effect (β_{age^2}). The cohort is split in seven age strata, as shown in Figure A.7.

	Estimate	LCI	UCI
CIDI	0.003	-0.097	0.103
DEP_EVER	0.011	-0.015	0.014
GP Diagnosis	-0.014	-0.01	0.01
PHQ9	-0.046	-0.129	0.13
TRD	-0.003	-0.004	0.004

Table A.6: Effect of CRP on depression outcomes, results for the cis-MR approach [57] that includes $n = 194$ SNPs in LD near the *CRP* gene.

Gene	SNP	Chromosome	Position	Total Expression in Brain Regions
CADM2	rs13078960	3	85758440	20.39608
GRID1	rs7899106	10	85651147	18.72607
RALYL	rs2033732	8	84167474	18.19608
ELAVL4	rs11583200	1	50094148	15.36253
MC4R	rs6567160	18	60161902	14.90682
CADM1	rs12286929	11	115151684	14.09107
AGBL4	rs657452	1	49124175	14.04236
HHIP	rs11727676	4	144737912	14.02281
RABEP1	rs1000940	17	5379957	13.7605
SBK1	rs2650492	16	28322090	13.31439
NRXN3	rs7141420	14	79433111	11.06906
NEGR1	rs3101336	1	72285502	9.719202
ETV5	rs1516725	3	186106215	8.168417
DMXL2	rs3736485	15	51456413	7.609515
FTO	rs1558902	16	53769662	7.357219
GRP	rs7243357	18	59216087	5.22611
LINGO2	rs10968576	9	28414341	4.872951
FHIT	rs2365389	3	61250788	4.049235
TNNI3K	rs12566985	1	74536509	1.882733
ZFHX4	rs17405819	8	75894349	1.520119
HIP1	rs1167827	7	75533848	0.527404
HSD17B12	rs2176598	11	43842728	0.202004
CCDC171	rs4740619	9	15634328	-0.07163
PGPEP1	rs17724992	19	18344015	-0.2375
NAV1	rs2820292	1	201815159	-0.2385
EHBP1	rs11688816	2	62825913	-0.45264
NPC1	rs1808579	18	23524924	-0.61269
SCARB2	rs17001654	4	76208415	-0.66507
MTCH2	rs3817334	11	47629441	-1.35257
TFAP2B	rs2207139	6	50877777	-1.73758
KAT8	rs9925964	16	31118574	-1.9354
KCNK3	rs11126666	2	26705943	-1.95684
MAP2K5	rs16951275	15	67784830	-2.02632
LMX1B	rs10733682	9	126698635	-2.06718
RARB	rs6804842	3	25064946	-2.13561
RPTOR	rs12940622	17	80641771	-2.52459
G2E3	rs11847697	14	30045906	-2.65925
NLRC3	rs758747	16	3577357	-3.49498
GNAT2	rs17024393	1	109612066	-4.06058
KCTD15	rs29941	19	33818627	-4.39511
CLIP1	rs11057405	12	122297350	-4.56075
TOMM40	rs2075650	19	44892362	-5.16283
QPCTL	rs2287019	19	45698914	-5.25288
POC5	rs2112347	5	75719417	-5.41987
RASA2	rs16851483	3	141556594	-5.63452
TRIM66	rs4256980	11	8652392	-5.84486
TMEM245	rs6477694	9	109170062	-6.04455
PRKD1	rs12885454	14	29267632	-6.36227
ZC3H4	rs3810291	19	47065746	-7.65248
NT5C2	rs11191560	10	103109281	-8.33572
FOXO3	rs9400239	6	108656460	-8.62294
GBE1	rs3849570	3	81742961	-9.87441
FUBP1	rs12401738	1	77981077	-10.2271

Table A.7: Matching of Locke et al. SNPs to genes and total brain expression in GTEx.

A.2.1 Supplementary Methods

Sensitivity Analysis: Classification of BMI SNPs: Given the established contribution of regulatory neural pathways to appetite control and subsequent changes in BMI, we aimed to split the Locke SNPs to those that are preferentially expressed in the brain or other tissues. The hypothesis is that those genes that affect BMI through a brain-related mechanism might be more likely to also affect depression in a direct manner, that is not through BMI modulation, i.e. they are pleiotropic. This is a complementary approach to the data-driven ways to guard against pleiotropy.

We use a tissue gene expression database to obtain information for the RNA measurement at the tissue level as measured in post-mortem samples [178]. We first identify the gene that corresponds to the 73 SNPs used as instruments. Of those, 59 could be mapped to coding regions (Table A.7). We then query the database with those genes [206] and present the trancript-per-million results for these genes for all available tissues. The results are presented in Figure A.12. Data was available for 53 genes. A total of 20 SNPs were chosen as having substantial expression in brain tissues (first 20 in Table A.7). These were then used as the brain-specific instrument for BMI. The rest of the SNPs constituted a single category of periphery-specific instrument and the analyses of the BMI effect on depression-related outcomes were repeated for comparison.

The results of the two subsets with UVMR, MVMR and MR GRAPPLE are presented in Figure A.13. The brain-specific and periphery-specific instruments provide similar estimates for PHQ9, CIDI, Lifetime MDD Diagnosis, and GP diagnosis. For TRD, the periphery-specific instrument estimates a positive causal effect of BMI, whereas the brain-specific analysis fails to reject the null in both UVMR and MVMR. In MR

GRAPPLE, the estimated effect of BMI seems to be more suggestive only for PHQ9, with both the brain and the periphery instruments.

A.3 Dimensionality Reduction Approaches in MVMR

A.3.1 Instrument Strength for PCs

Since we transform $\hat{\gamma}$ and obtain a matrix of lower dimensionality, formula 4.1 can't be used as there is no longer a one-to-one correspondence of the SEs with the PCs. Likewise, a conditional $F-$statistic for the PCs also cannot be computed for this reason. We aim to arrive at a modified formula that bypasses this issue. For this purpose, we take advantage of two concepts, first an expression of the $F-$statistic for an exposure k (F_k) in matrix notation and, second, the use of this expression to estimate $F-$statistics for the PCs (F_{PC}) from $\hat{\gamma}$ decomposition.

We make the assumption that the uncertainty in the $\hat{\gamma}_{G,X_K}$ estimates is similar in all K exposures, i.e. $\hat{\gamma}_{G,X}$ uncertainty estimates do not substantially differ among exposures. This is not implausible as the uncertainty is predominantly driven by sample size and minor allele frequency [207]. Specifically, the authors of [207] show that

$$Var(\hat{\gamma}_{X_k}) = \frac{1}{n_k Var(X_k) MAF(1-MAF)},$$

where MAF is the minor allele frequency, n_k is the sample size in exposure k and $Var(X_k)$ is the phenotypic variance. What this means is that, in experiments such as [102] where n_k is the same across all exposures and $Var(X_k)$ can be standardised to 1, the main driver of differences in $Var(\hat{\gamma}_{X_K})$ is differences in MAF. As MAF is the same for each SNP across all exposures, the collation of SEs across exposures per SNP is well motivated.

We can then define a matrix Σ as follows.

$$\Sigma = \begin{bmatrix} \bar{SE}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \bar{SE}_2^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \bar{SE}_p^2 \end{bmatrix}, \; \bar{SE}_j^2 = \frac{\sum_{k=1}^{K} SE_{j,k}^2}{K}.$$

The elements in the diagonal represent the mean variance of $\hat{\gamma}$ for each SNP and all off-diagonal elements are zero. What is achieved through this is a summary of the uncertainty in the SNP-X associations that is not sensitive on the dimensions of the exposures. Instead of Eq. 4.1, we can then express the vector of the mean F-statistics for each exposure $F_{1-K} = [F_1, F_2, ..., F_K]$ as

$$\underset{K\times 1}{\mathrm{F}_{1-\mathrm{K}}} = \frac{1}{p} \times diag[\underset{K\times p}{\hat{\gamma}^T} \times \underset{p\times p}{\Sigma^{-1}} \times \underset{p\times K}{\hat{\gamma}}], \tag{A.1}$$

where $\hat{\gamma}$ is the *matrix* of the SNP-exposure associations. In a simulation study, we generate data under the mechanism in Figure A.15a. The strength of association is different in the three exposures. It is observed that the estimates with both methods (Eq. A.1 and Eq. 4.1) align well (Figure A.15b), supporting the equivalence of the two formulae.

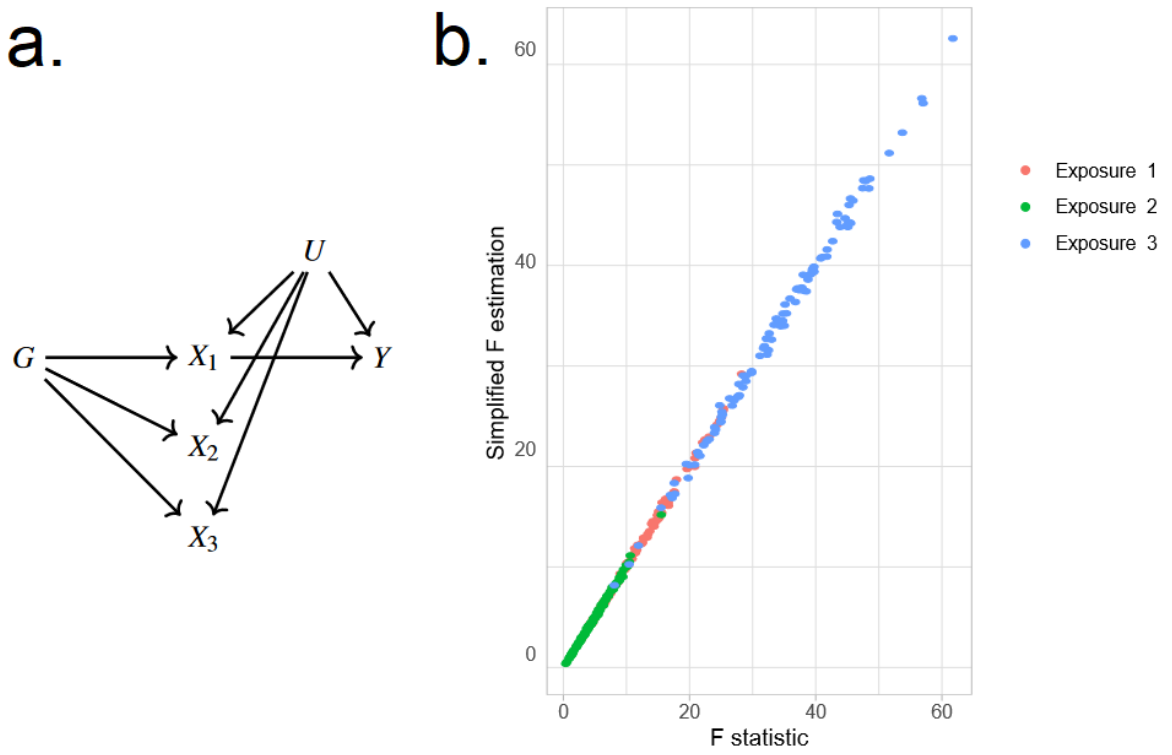


Figure A.15: a. Data generating mechanism. Three exposures with different degrees of strength of association with G are generated $\gamma_1 = 1, \gamma_2 = 0.5, \gamma_3 = 0.1$. b. F-statistic for the three exposures X_1, X_2, X_3 as estimated by the formulae in Eq. A.1 (horizontal axis) and Eq. 4.1 (vertical axis).

Our second aim is to use this matrix notation formula for the $F-$statistic to quantify the instrument strength of each PC with the respective $F-$statistic (F_{PC}). As presented above, we are not limited by the dimensions of point estimates and uncertainty matching exactly and we can use the formula in Eq. A.1 and substitute $\hat{\gamma}$ with the PCs. For the PCA approach, where we decompose $\hat{\gamma}$ as $\hat{\gamma} = UDV^T$ and carry forward $M << K$ non-trivial PCs, we use the matrix UD in place of $\hat{\gamma}$. Then, the mean F_{PC} can be estimated as follows.

$$\underset{M\times 1}{\mathrm{F}_{\mathrm{PC}_{1-\mathrm{M}}}} = \frac{1}{p} \times diag[\underset{M\times p}{UD^T} \times \underset{p\times p}{\Sigma^{-1}} \times \underset{p\times M}{UD}]. \tag{A.2}$$

The vector $F_{PC_{1-M}} = [F_{PC_1}, F_{PC_2}, ..., F_{PC_M}]$ contains the F_{PC} statistics for the M PCs. In a similar manner, we estimate F_{PC} for the sparse PCA methods but, instead of the scores matrix UD, we use the scores of the sparse methods. We illustrate the performance of this approach in a simulation study with an identical configuration for exposure generation as the one presented in Figure A.22. In a configuration with $b = 6$ blocks of $p = 30$ genetically correlated exposures (Figure 4.5), we vary the strength of association γ per block. This way, the first block has the highest strength of association and the last block the lowest, quantified by a lower mean $F-$statistic in the exposures of this block (red, Figure A.16). The instrument strength of the PCs and the sPCs follow closely the corresponding $F-$statistics of the individual exposures; in other word, in a PC of five exposures, F_{PC_1}, F_{SCA_1} and F_{1-5} align well (Figure A.16).

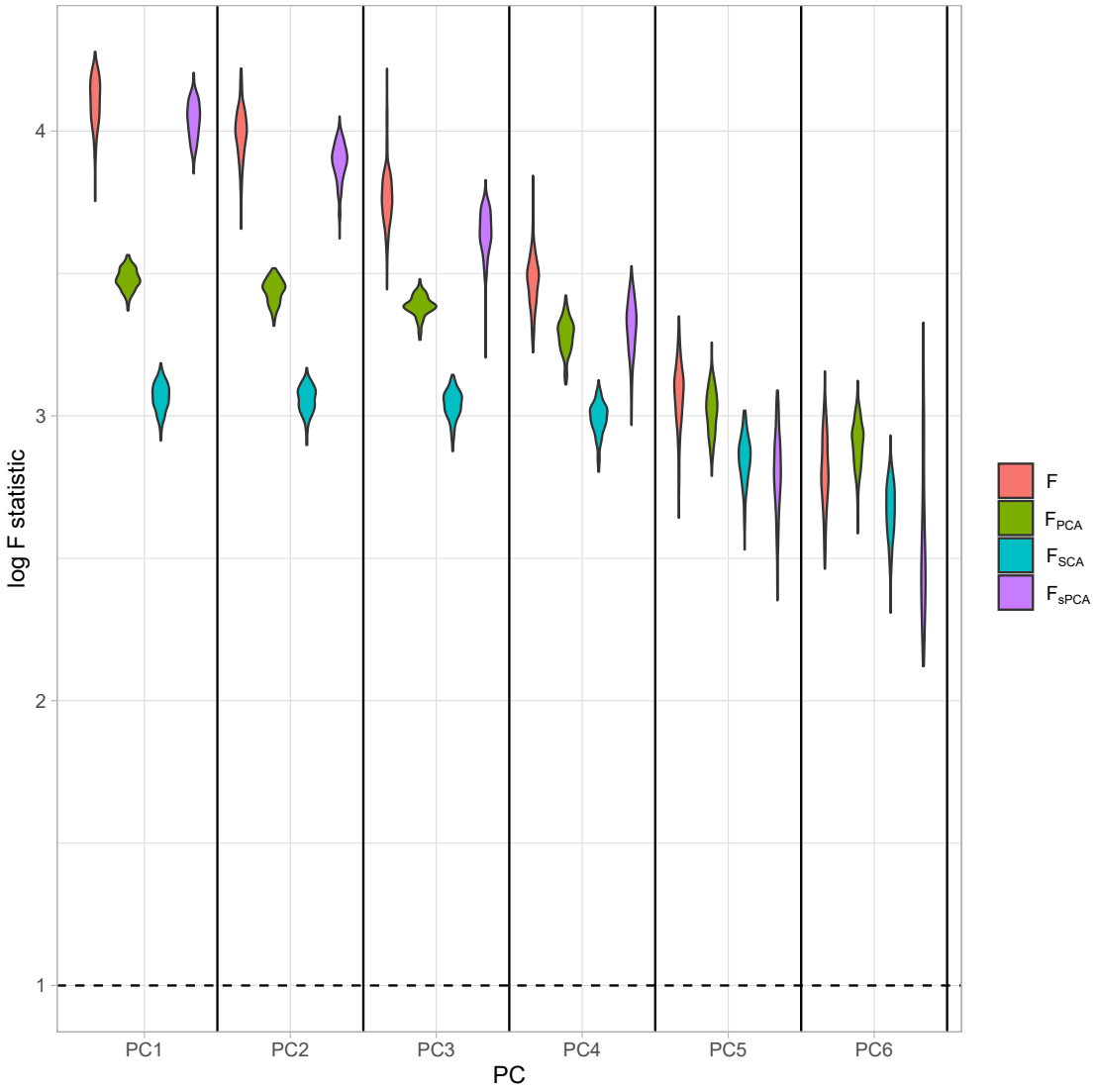


Figure A.16: Distributions of the $F-$statistics in PCA methods and individual (not transformed) exposures. Exposure data in different blocks are simulated with a decreasing strength of association and the correlated blocks map to PCs. Each distribution represents the $F-$statistics for each PC. In the case of the individual exposures (red), the distributions represent the $F-$statistics for the corresponding exposures. Individual: individual exposures without any transformation; PCA: $F-$statistics for PCA; SCA: sparse component analysis [112]; sPCA: sparse PCA as described by Zou et al. [107]

A.3.2 Multivariable IVW, MR GRAPPLE

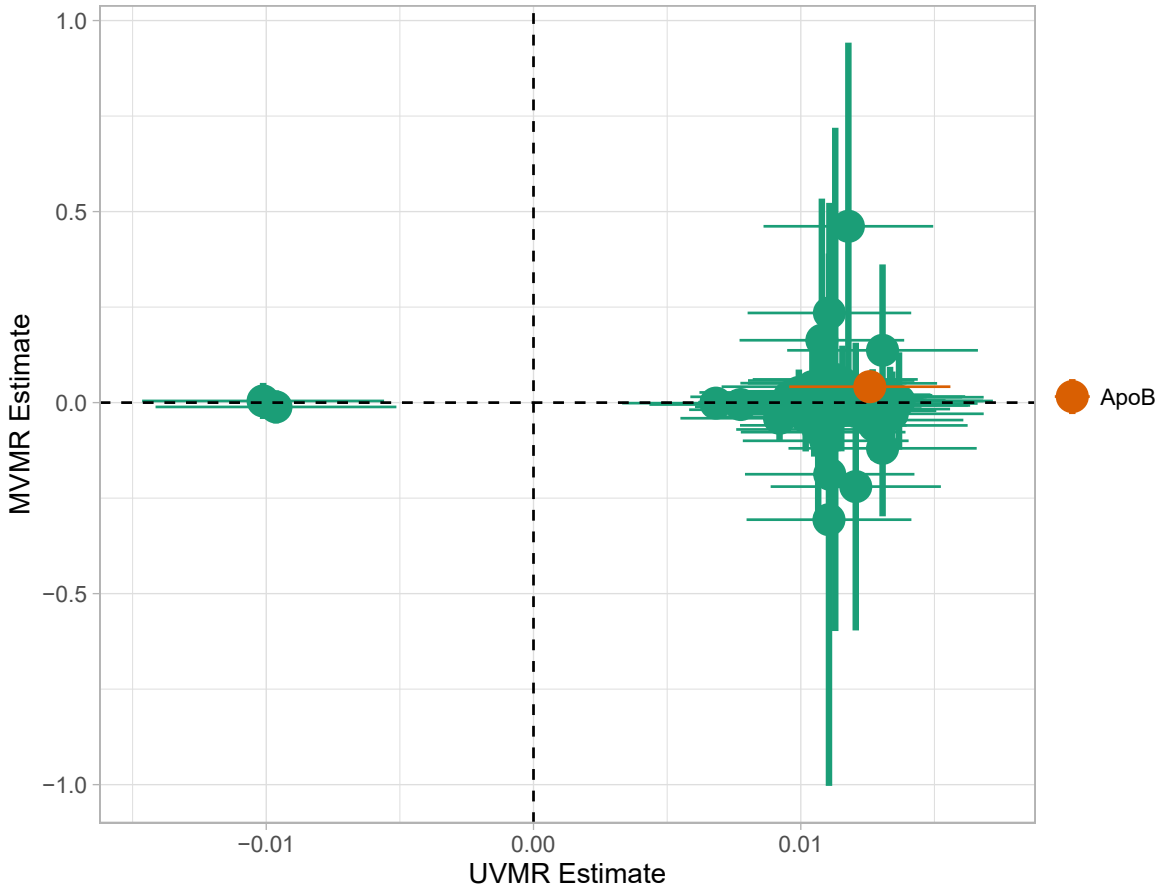


Figure A.17: MVMR and UVMR estimates. Only ApoB is strongly associated with CHD. All SEs are larger in the MVMR model (range of $\frac{SE_{MVMR}}{SE_{UVMR}}$ $2.7 - 225.96$).

A small negative effect for M.LDL.PL is noted as nominally statistically significant in Fig. A.18. This is not concordant with the UVMR direction of effect. In GRAPPLE, no traits surpass the nominal significance threshold.

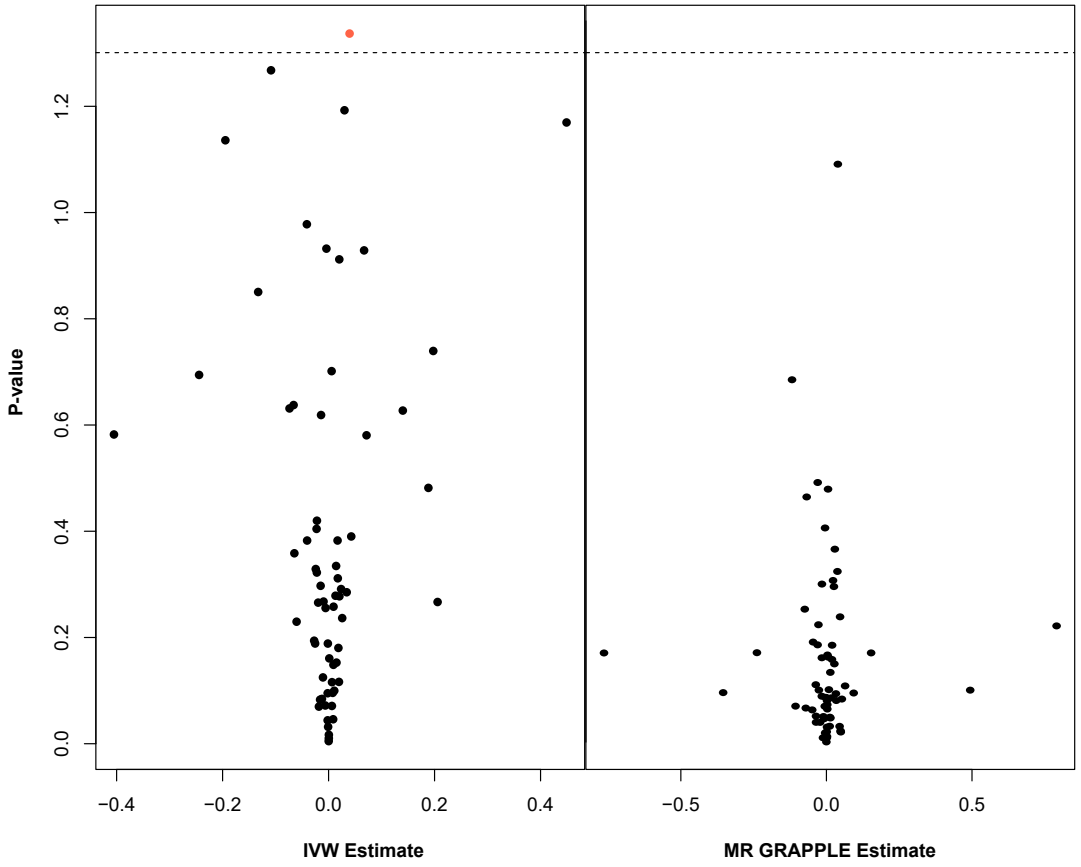


Figure A.18: MVMR with IVW (left) and MVMR with GRAPPLE [185] (right). Only the 66 exposures that are statistically significant in UVMR are put forward in these models. In IVW (left), ApoB shows nominal significance. In MR GRAPPLE (right), apolipoprotein B has the lowest p-value but no trait reaches nominal significance. On the vertical axis, the -$\log_{10}$P-value is shown.

Method	Package	Authors	Features	Choice	PCs
RSPCA	*pcaPP*	P. Filzmoser, H. Fritz, K. Kalcher [117]	Robust sPCA (RSPCA), different measure of dispersion (Q_n)	Permutation KSS	6
SFPCA	Code in publication, Supplementary Material	J. Guo, G. James, E. Levina, G. Michailidis, G. Zhu [111]	Fused penalties for block correlation	KSS	6
sPCA	*elasticnet*	H. Zou, T. Hastie [107]	Formulation of sPCA as a regression problem	KSS	6
SCA	*SCA*	F. Chen, K. Rohe [112]	Rotation of eigenvectors for approximate sparsity	Permutation KSS	6

Table A.9: Overview of sPCA methods used. KSS: Karlis-Saporta-Spinaki criterion. Package: *R* package implementation; Features: short description of the method; Choice: method of selection of the number of informative components in real data; PCs: number of informative PCs.

A.3.3 MVMR with PC Scores

PC	Method	OR	LCI	UCI
PC1	PCA	1.002	1.0015	1.0024
PC2	PCA	1.0002	0.9995	1.001
PC3	PCA	1.0013	1.0001	1.0024
PC4	PCA	0.9985	0.997	0.9999
PC5	PCA	0.9999	0.9978	1.002
PC6	PCA	0.9993	0.9976	1.0009
PC1	SCA	1.0027	1.0005	1.0049
PC2	SCA	1.0027	1.0004	1.005
PC3	SCA	0.9997	0.9976	1.0019
PC4	SCA	0.9965	0.9941	0.9989
PC5	SCA	1.0002	0.998	1.0024
PC6	SCA	1.0034	0.9989	1.0078
PC1	sPCA	1.0019	0.9999	1.0039
PC2	sPCA	1.0003	0.9986	1.002
PC3	sPCA	0.9988	0.997	1.0005
PC4	sPCA	0.9975	0.9955	0.9995
PC5	sPCA	0.998	0.9954	1.0006
PC6	sPCA	0.9998	0.9982	1.0014
PC1	RSPCA	1.0017	1.0006	1.0027
PC2	RSPCA	0.9998	0.9983	1.0013
PC3	RSPCA	0.9954	0.9918	0.999
PC4	RSPCA	0.9989	0.9969	1.0008
PC5	RSPCA	0.9944	0.9903	0.9986
PC6	RSPCA	1.01	1.0013	1.0188
PC1	SFPCA	1.002	1.0015	1.0025
PC2	SFPCA	0.9991	0.9979	1.0004
PC3	SFPCA	0.9998	0.9991	1.0006
PC4	SFPCA	0.9982	0.9967	0.9997
PC5	SFPCA	1.0001	0.9977	1.0025
PC6	SFPCA	1.0009	0.9985	1.0033

Table A.10: Estimated Causal effects of PCs on CHD risk. PCA: Principal Component Analysis; SCA: Sparse Component Analysis; sPCA: sparse PCA [107]; RSPCA: robust sparse PCA.

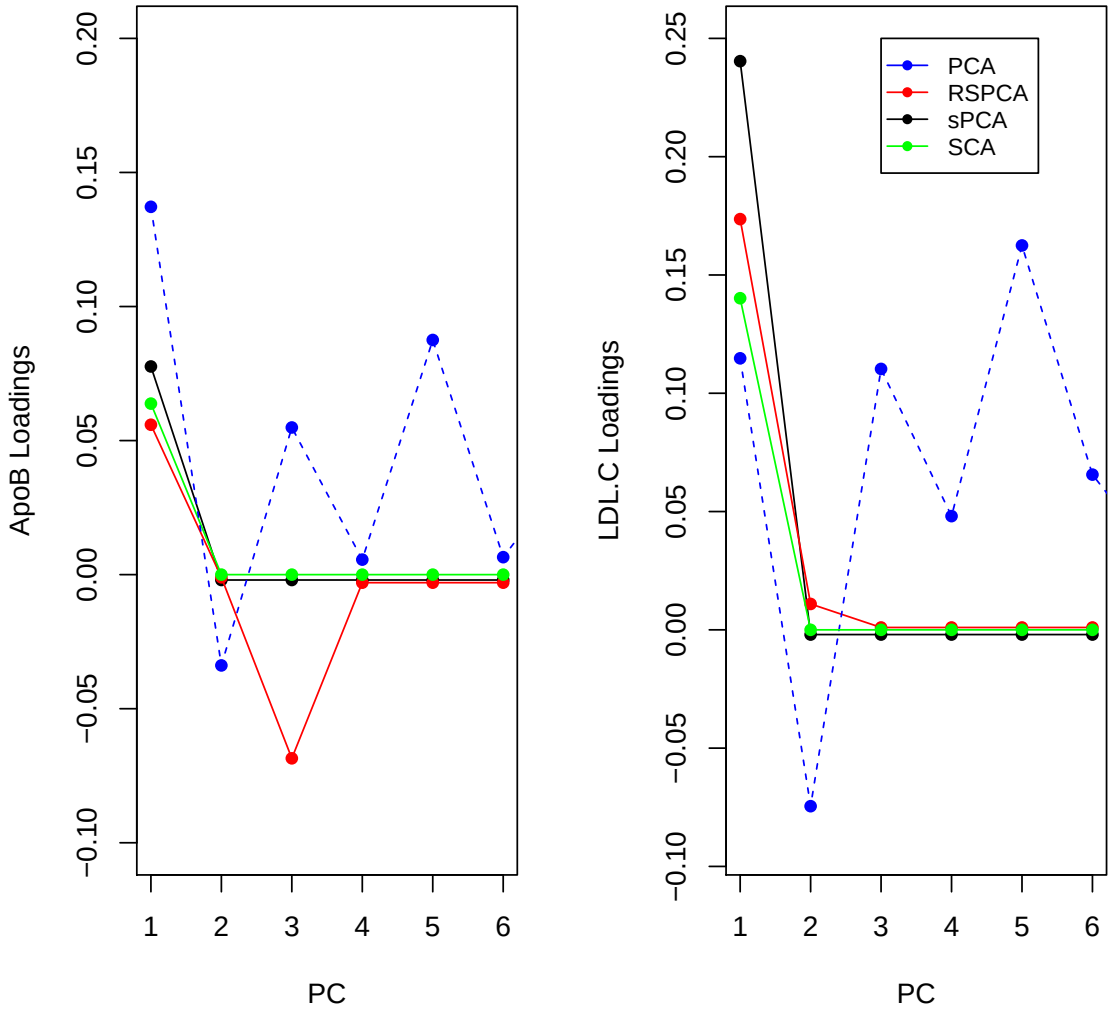


Figure A.19: Trajectories for the loadings of total cholesterol in LDL and ApoB in all methods. PCA loadings imply a contribution of LDL.c and ApoB to all PCs. In the sparse methods, this is limited to one PC (two for RSPCA).

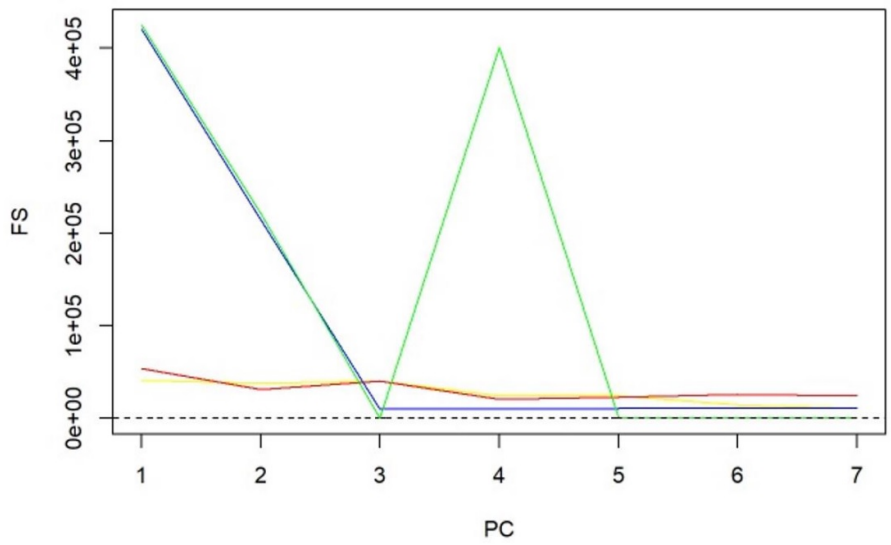


Figure A.20: : F-statistics for PCs and sparse PCs. The formula derived in Eq. A.1 is used. Black: PCA (no sparsity constraints); Yellow: SCA; Red: sparse PCA (Zou); Blue: Sparse robust PCA; Green: Sparse fused PCA. The dashed line represents the cutoff of 10 that is considered the minimum desired F-statistic for an exposure to be considered well instrumented. The green line diverges from the pattern of decreasing instrument strength but, when referring to the loadings heatmap (Figure 4.4), it can be observed that the 4th sparse PC in the fused sPCA receives negative loadings from multiple VLDL and LDL related traits. This may in turn cause the large F-statistic.

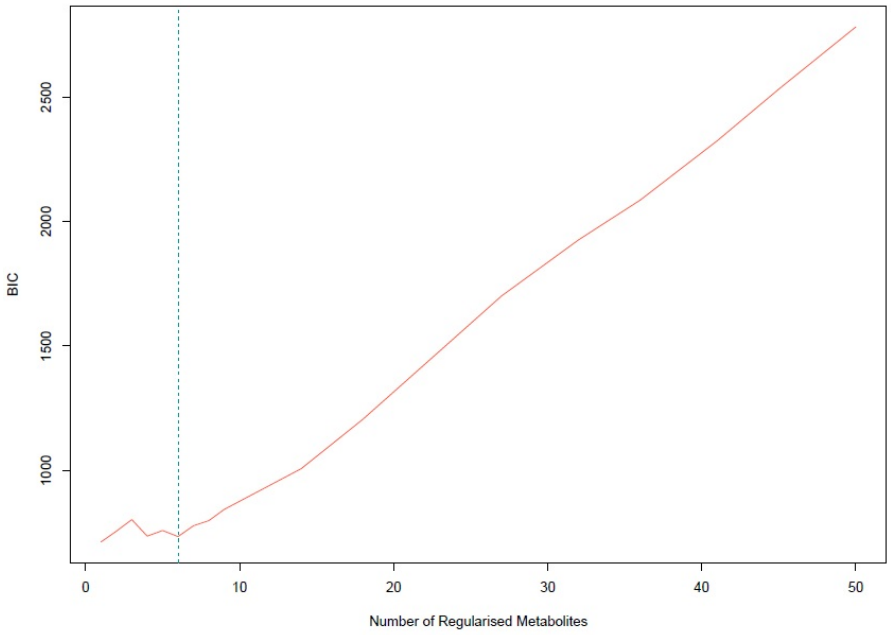


Figure A.21: Bayesian Information Criterion (BIC) for different numbers of metabolites regularized to 0. The lowest value is achieved for one non-zero exposure per component. However, six non-zero exposures per component also achieved a similar low BIC and this was selected.

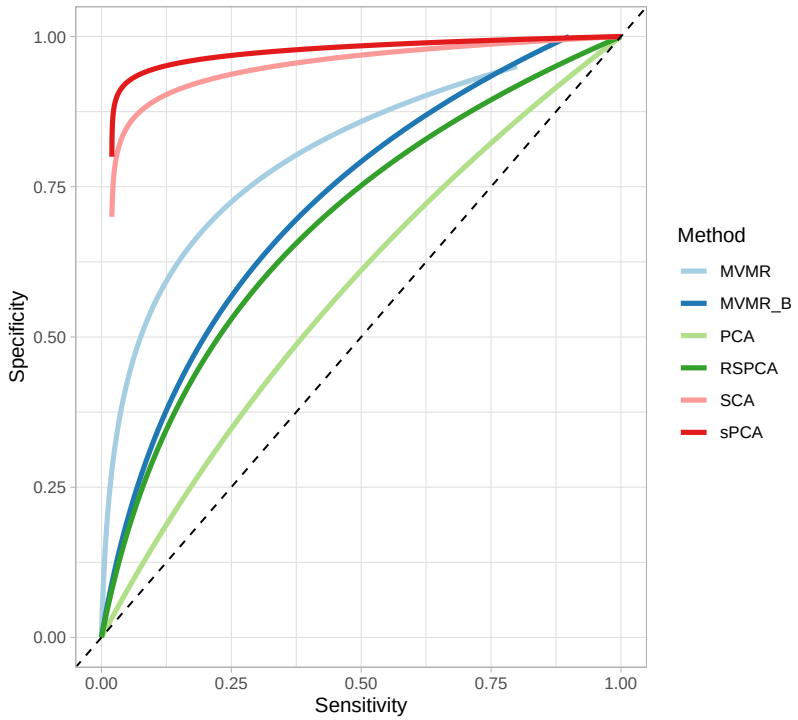


Figure A.22: Extrapolated ROC curves for all methods. SCA: Sparse Component Analysis [112]; sPCA: sparse PCA (Zou et al.) [107]; RSPCA: robust sparse PCA [117]; PCA: principal component analysis; MVMR: multivariable MR; MVMR_B: MVMR with Bonferroni correction.

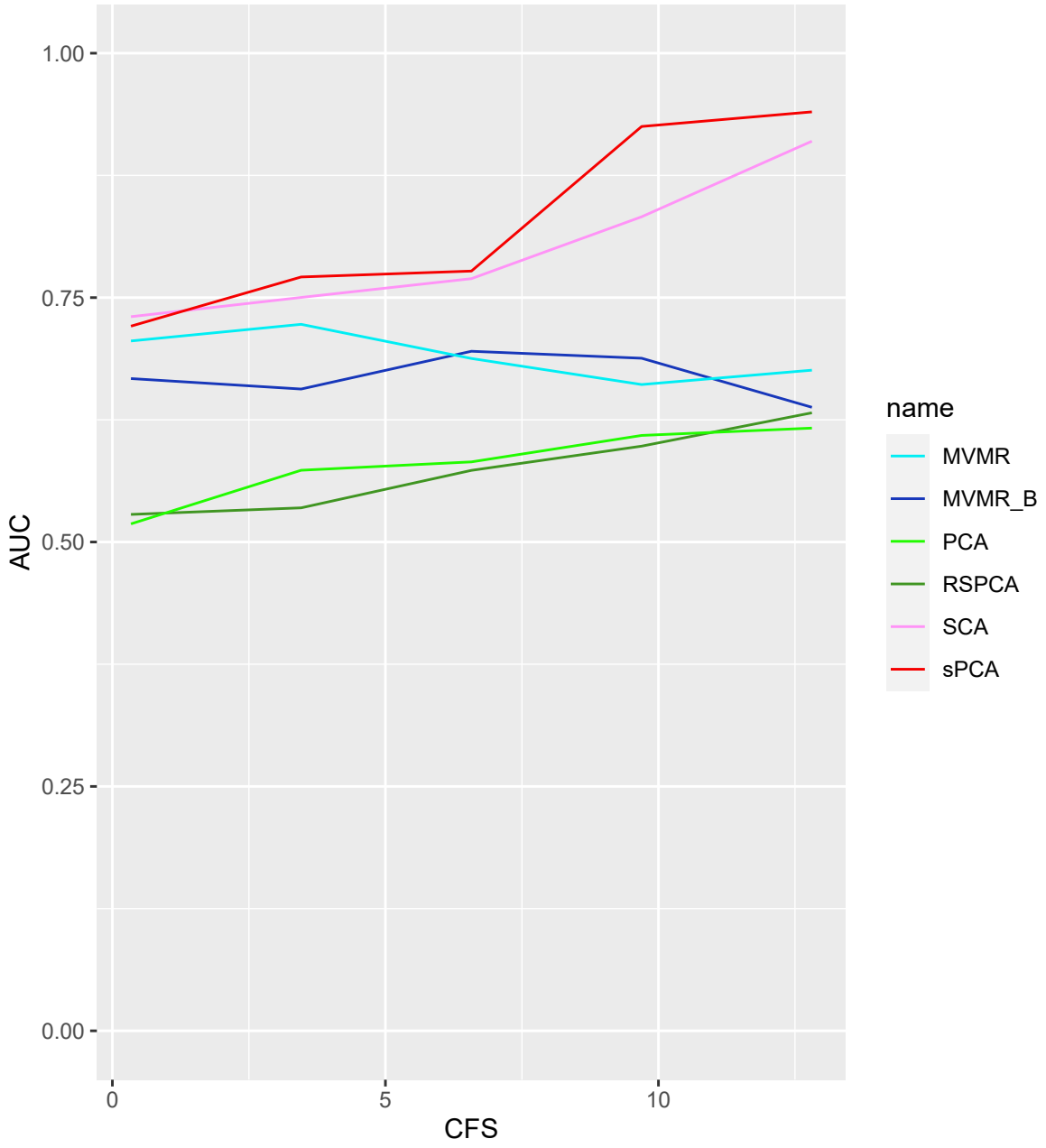


Figure A.23: AUC performance of MVMR and dimensionality reduction methods for increasing sample sizes. Two sparse methods (SCA, sPCA) perform better compared with PCA and MVMR, with improving performance as the sample size increases. CFS: Conditional F-statistic.

	PCA	SCA	sPCA	RSPCA	MVMR_B	MVMR
AUC	0.56	0.919	0.941	0.644	0.660	0.712
Sensitivity	1,0.1	1,0.21	1, 0.047	0.667, 0.251	0.222, 0.2	0, 0.076
Specificity	0,0.02	0.925,0.772	0.936, 0.097	0.192, 0.104	0.960, 0.048	1,0
Youden's J	0	0.584	0.778	-0.061	0.192	0.044

Table A.11: Sensitivity & Specificity presented as median and interquartile range across all simulations. Presented as median sensitivity/specificity and interquartile range across all simulations; AUC: area under the ROC curve.

	PCA	SCA	sPCA	RSPCA	MVMR	MVMR_B
AUC	0.799	0.714	0.859	0.492	0.511	0.675
SNS	1,0.03	0.75,0.25	1,0.17	0.5,0.25	0.25,0.25	0,0
SPC	0,0.2	0.76,0.46	0.66,0.18	0.37,0.15	0.94,0.07	1,0
Youdens J	0	0.428	0.625	-0.029	0.105	0.032

Table A.12: Simulation study on only four exposures (out of the total $K = 50$) contributing to the outcome Y. A drop in sensitivity and specificity is observed for SCA and sPCA compared with the simulation configuration in Table A.11.

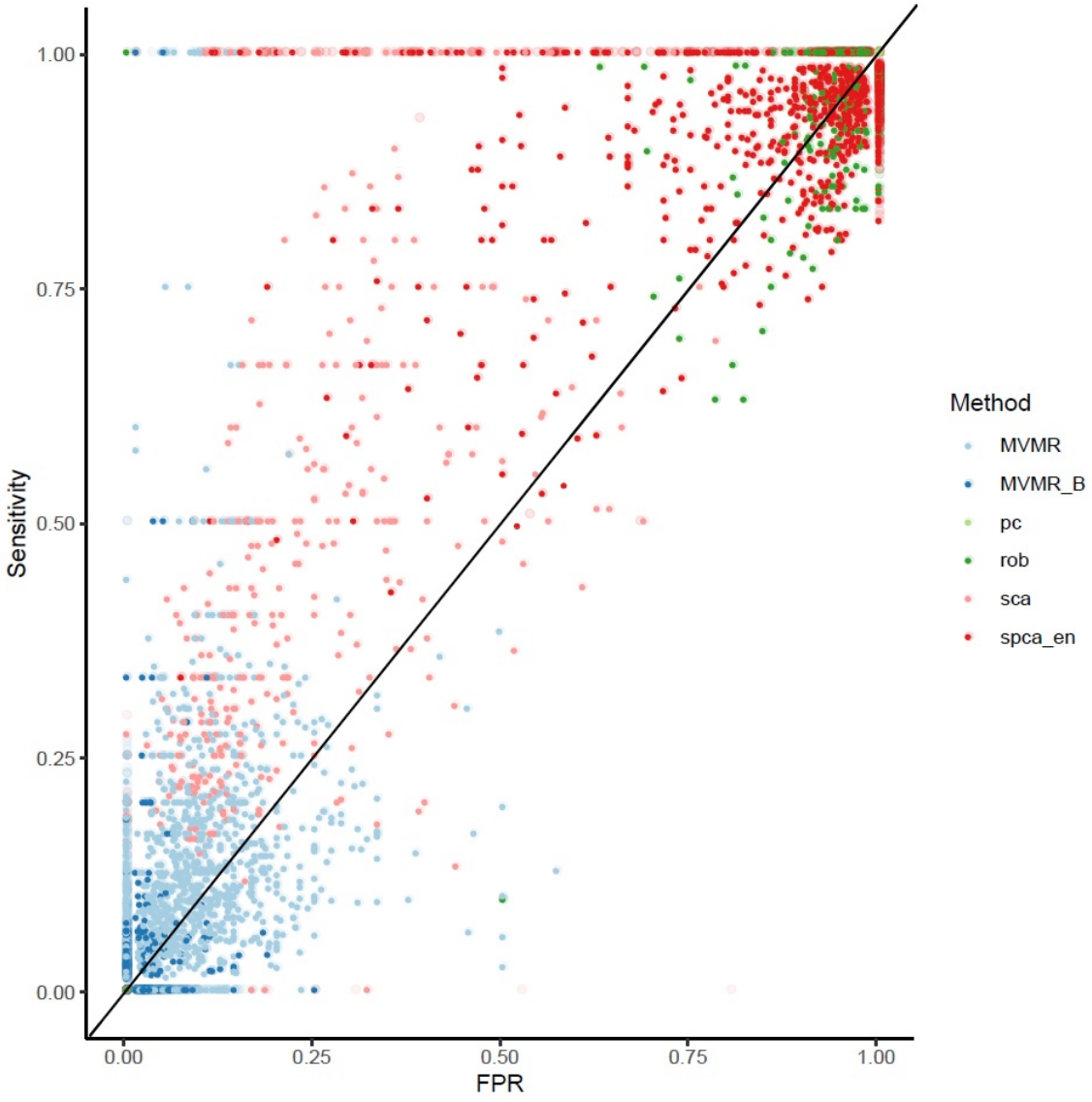


Figure A.24: Individual Results from $s = 1000$ simulations.

Inflammatory Mediator	SNPs	IVW β	IVW SE	IVW $p-$value	Sample Size
Interleukin-2 levels	rs13412535,rs61335305	-0.02823	0.00516	0.115122	3475
Macrophage colony stimulating factor levels	rs62294910,rs56367447	0.029188	0.000989	0.021565	840
Platelet-derived growth factor BB levels	rs55680718,rs72777070,rs13412535, rs116445074,rs2324229,rs12289510, rs4965869,rs9941733	0.005531	0.012795	0.678538	8293
CTACK levels	rs2731674,rs2070074,rs55764737, rs57338032,rs76395525,rs135555	-0.02374	0.008946	0.0452	3631
Eotaxin levels	rs187131,rs12075,rs2228467,rs112347425, rs2024050,rs9317045,rs80341932, rs2211994,rs11087905	0.009024	0.005999	0.176234	8153
Vascular endothelial growth factor levels	rs143479231,rs13209117,rs6921438, rs9472183,rs7030781,rs10967186, rs10761731,rs8045833	-0.0074	0.010442	0.505214	7118
Monocyte chemoattractant protein-3 levels	rs10892381	0.022388	0.020385	0.272092	843
Monocyte chemoattractant protein-1 levels	rs12075,rs7517040,rs56212190, rs2228467, rs12493471,rs112313229,rs62243917, rs10744620,rs146522229	-0.00331	0.00798	0.688802	8293
Interleukin-18 levels	rs385076,rs4482818,rs116383510, rs17229943, rs115267715,rs658805,rs78623212, rs71478720,rs1979967,rs10414578	0.004674	0.00634	0.482029	3636
beta-nerve growth factor levels	rs7970581,rs28637706	-0.00273	0.03838	0.954856	3531
Interleukin-5 levels	rs7767396,rs73040130	-0.00396	0.02839	0.889056	3364
Stem cell growth factor beta levels	rs4656185,rs17876031,rs139413256, rs117716477,rs116924815	0.002481	0.007541	0.758662	3682
Macrophage inflammatory protein 1b levels	rs116237296,rs76356863,rs113010081, rs79068918,rs72791296,rs76776296, rs113877493,rs17641689,rs2079664, rs34437725,rs141102180,rs117453826	-0.00197	0.006183	0.755441	8243
Interleukin-17 levels	rs17282552,rs1530455,rs184080173, rs17106604,rs11640734	-0.0033	0.015102	0.837833	7760
Growth-regulated protein alpha levels	rs12075,rs508977,rs185768063	-0.0058	0.006796	0.550495	3505
Hepatocyte growth factor levels	rs3748034,rs5745687,rs6077285	0.027562	0.037816	0.599044	8292
TRAIL levels	rs79287178,rs13278062,rs28521641, rs74778900,rs193112415,rs62093514, rs57396456,rs138987090	0.001527	0.006182	0.813077	8186
Interleukin-6 levels	rs13412535,rs72831623,rs73273528	-0.00239	0.049913	0.969499	8189
Tumor necrosis factor alpha levels	rs115669577,rs111332265	0.021804	0.015627	0.395875	3454
Stem cell factor levels	rs1557570,rs13412535,rs1568119, rs80271436,rs635634,rs4841899	0.013216	0.015349	0.428587	8290
Interleukin-7 levels	rs4320361,rs62006410,rs144701438, rs147747784	0.005734	0.011506	0.667639	3409
Fibroblast growth factor basic levels	rs13412535,rs145577605,rs9907295	0.068837	0.015502	0.047157	7565
Monokine induced by gamma interferon levels	rs6679677,rs55876513,rs41272086, rs1796086,rs62562991,rs816960, rs112861654	-0.01549	0.017527	0.417333	3685
Interleukin-2 receptor antagonist levels	rs115360066,rs12722497	-0.00223	0.012105	0.853863	3677
Macrophage inflammatory protein 1a levels	No hits surpassing the threshold	NA	NA	NA	3522
Interleukin-1-beta levels	No hits surpassing the threshold	NA	NA	NA	3309
Interleukin-16 levels	rs4253283,rs4513633,rs1801020, rs1255143,rs117916513,rs9706053, rs4778636,rs144691581	0.012466	0.008063	0.166	3483
Interleukin-1-receptor antagonist levels	No hits surpassing the threshold	NA	NA	NA	3638
Interleukin-8 levels	rs12075,rs2673604	0.011119	0.010636	0.485861	3526
Interleukin-4 levels	rs13106889,rs73023729,rs17713451, rs10512267,rs9941733	0.021532	0.012635	0.186892	8124
Stromal-cell-derived factor 1 alpha levels	rs10013755	-0.03772	0.030271	0.212714	5998
Interleukin-13 levels	rs9472168,rs142167313,rs117795020	-0.00241	0.004988	0.676875	3557
Interferon gamma-induced protein 10 levels	rs113831257,rs9450351,rs79848609	0.023366	0.009561	0.247261	3685
Interleukin-12p70 levels	rs13209117,rs9472183,rs4349809, rs2375980,rs10761731,rs72831623	-0.01399	0.016703	0.463723	8270
Macrophage Migration Inhibitory Factor levels	rs13142904,rs78098071,rs28994873, rs12594190,rs2330634	0.008635	0.021887	0.713321	3494
RANTES levels	rs112072646,rs7000423,rs74472919, rs2702950,rs147509526	0.020258	0.021908	0.407475	3421
Granulocyte-colony stimulating factor levels	rs11903143,rs2324653,rs115256310, rs145756094,rs76287671	0.020446	0.025704	0.470898	7904
Interleukin-9 levels	rs76963786	-0.01291	0.026298	0.62337	3634
Interferon gamma levels	rs78296352,rs10761731,rs113600793	-0.04601	0.017521	0.231618	7701
Tumor necrosis factor beta levels	rs78296352	-0.00982	0.009479	0.299994	1559
Interleukin-10 levels	rs10493718,rs282258,rs111913416, rs10457128,rs4349809,rs2375980, rs7088799				

Table A.8: Inflammatory Mediators [187] and Major depressive disorder [182].

Glossary

Phenotype: A characteristic of an individual that can be measured, such as height, weight, or diagnosis of a condition.

Causal Inference: The methodologies that allow for the extraction of conclusions about causes and effects in observed phenomena.

Genetic Variant: A variation in the genetic material of an individual. Commonly used as instruments in MR.

Instrumental variable (IV) analysis: A statistical method that uses instrumental variables as proxies to estimate causal effects of exposures on outcomes in the presence of confounding. The ideal instrumental variable directly affects only the exposure.

Pleiotropy: The phenomenon in genetics where a gene or genetic variant affects multiple phenotypes.

Genome-wide association study (GWAS): A study that identifies genetic variants associated with a particular phenotype or trait by analysing millions of genetic variants across the entire genome.

Mendelian randomisation (MR): A method of using genetic variants as instrumental variables to estimate causal effects of a risk factor on an outcome of interest.

Principal component analysis (PCA): A statistical technique used to reduce the

dimensionality of a dataset by identifying patterns and correlations among variables. PCA transforms the original variables into a new set of uncorrelated variables, called principal components, that explain the maximum amount of variance in the data.

Weak Instrument Bias: The situation where an instrumental variable is only weakly predictive of the target exposure and hence this leads to bias in the instrumental variable estimate.

Bibliography

[1] Steven A Szklo and F Javier Nieto. *Epidemiology Beyond the Basics*. 2nd. Jones and Bartlett Publishers, 2007, p. 27.

[2] D. L Sackett et al. “Evidence based medicine: what it is and what it isn’t”. In: *BMJ* 312.7023 (Jan. 1996), pp. 71–72. DOI: 10.1136/bmj.312.7023.71. URL: https://doi.org/10.1136/bmj.312.7023.71.

[3] Edward Shorter. *A history of psychiatry : from the era of the asylum to the age of Prozac*. John Wiley Sons, Inc, Nov. 1997. DOI: 10.1176/ps.49.9.1241. URL: https://doi.org/10.1176/ps.49.9.1241.

[4] Gordon Guyatt. “Evidence-Based Medicine”. In: *JAMA* 268.17 (Nov. 1992), p. 2420. DOI: 10.1001/jama.1992.03490170092032. URL: https://doi.org/10.1001/jama.1992.03490170092032.

[5] M Hassan Murad et al. “New evidence pyramid”. In: *Evidence Based Medicine* 21.4 (June 2016), pp. 125–127. DOI: 10.1136/ebmed-2016-110401. URL: https://doi.org/10.1136/ebmed-2016-110401.

[6] Neil M Davies, Michael V Holmes, and George Davey Smith. “Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians”. In: *BMJ* (July 2018), k601. DOI: 10.1136/bmj.k601. URL: https://doi.org/10.1136/bmj.k601.

[7] AJ Wakefield et al. “RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children”. In: *The Lancet* 351.9103 (Feb. 1998), pp. 637–

641. DOI: 10.1016/s0140-6736(97)11096-0. URL: https://doi.org/10.1016/s0140-6736(97)11096-0.

[8] R. Doll and A. B. Hill. “Smoking and Carcinoma of the Lung”. In: *BMJ* 2.4682 (Sept. 1950), pp. 739–748. DOI: 10.1136/bmj.2.4682.739. URL: https://doi.org/10.1136/bmj.2.4682.739.

[9] R Doll and R Peto. “Mortality in relation to smoking: 20 years’ observations on male British doctors.” In: *BMJ* 2.6051 (Dec. 1976), pp. 1525–1536. DOI: 10.1136/bmj.2.6051.1525. URL: https://doi.org/10.1136/bmj.2.6051.1525.

[10] UK Biobank. *UK Biobank*. https://www.ukbiobank.ac.uk/. Accessed on March 23, 2023. 2016.

[11] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), pp. 203–209.

[12] Anna Fry et al. “Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population”. In: *American Journal of Epidemiology* 186.9 (June 2017), pp. 1026–1034. DOI: 10.1093/aje/kwx246. URL: https://doi.org/10.1093/aje/kwx246.

[13] Sarah J. Lewis et al. “Cotinine levels and self-reported smoking status in patients attending a bronchoscopy clinic”. In: *Biomarkers* 8.3-4 (Jan. 2003), pp. 218–228. DOI: 10.1080/1354750031000120125. URL: https://doi.org/10.1080/1354750031000120125.

[14] Sindre M. Dyrstad et al. “Comparison of Self-reported versus Accelerometer-Measured Physical Activity”. In: *Medicine and Science in Sports and Exercise* 46.1 (Jan. 2014), pp. 99–106. DOI: 10.1249/mss.0b013e3182a0595f. URL: https://doi.org/10.1249/mss.0b013e3182a0595f.

[15] CR Rao, JP Miller, and DC Rao. *Handbook of Statistics 27*. Elsevier, June 2008. DOI: 10.1016/S0169-7161(07)27003-8. URL: https://doi.org/10.1016/S0169-7161(07)27003-8.

[16] Julian P.T. Higgins et al. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd. Chichester, UK: John Wiley & Sons, 2019. DOI: 10.1002/9781119536604.

[17] Matthias Egger et al. “Bias in meta-analysis detected by a simple, graphical test”. In: *BMJ (Clinical research ed.)* 315.7109 (1997), pp. 629–634. DOI: 10.1136/bmj.315.7109.629. URL: https://www.bmj.com/content/315/7109/629.

[18] Jonathan A C Sterne et al. “RoB 2: a revised tool for assessing risk of bias in randomised trials”. In: *BMJ* (Aug. 2019), p. l4898. DOI: 10.1136/bmj.l4898. URL: https://doi.org/10.1136/bmj.l4898.

[19] Robin Poole et al. “Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes”. In: *BMJ* 359 (Nov. 2017), j5024. DOI: 10.1136/bmj.j5024. URL: https://www.bmj.com/content/359/bmj.j5024.

[20] Jia-Yi Dong et al. “Depression and risk of stroke: a meta-analysis of prospective studies”. In: *Stroke* 43.1 (2012), pp. 32–37.

[21] Faith Matcham et al. “The Relationship Between Mental Health, Disease Severity, and Genetic Risk for Depression in Early Rheumatoid Arthritis”. In: *Psychosomatic Medicine* 79.6 (2017), pp. 638–645.

[22] Alan M Rathbun et al. “Dynamic Effects of Depressive Symptoms on Osteoarthritis Knee Pain”. In: *Arthritis Care Research* 70.1 (2018). PMCID: PMC5607075, NIHMSID: NIHMS860647, pp. 80–88. DOI: 10.1002/acr.23239. URL: https://doi.org/10.1002/acr.23239.

[23] Abebaw M Yohannes and George S Alexopoulos. “Depression and anxiety in patients with COPD”. In: *European Respiratory Review* 23 (2014), pp. 345–349. DOI: 10.1183/09059180.00007813.

[24] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[25] Joshua D. Angrist and Alan B. Krueger. “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments”. In: *Journal of Economic Perspectives* 15.4 (Sept. 2001), pp. 69–85.

[26] Stephen Soldz and George E. Vaillant. “The Big Five Personality Traits and the Life Course: A 45-Year Longitudinal Study”. In: *Journal of Research in Personality* 33.2 (June 1999), pp. 208–232. DOI: 10.1006/jrpe.1999.2243. URL: https://doi.org/10.1006/jrpe.1999.2243.

[27] Jack Bowden et al. “Connecting Instrumental Variable methods for causal inference to the Estimand Framework”. In: *Statistics in Medicine* 40.25 (July 2021), pp. 5605–5627. DOI: 10.1002/sim.9143. URL: https://doi.org/10.1002/sim.9143.

[28] Sander Greenland. “An introduction to instrumental variables for epidemiologists”. In: *International Journal of Epidemiology* 29.4 (Aug. 2000), pp. 722–729. DOI: 10.1093/ije/29.4.722. URL: https://doi.org/10.1093/ije/29.4.722.

[29] Patrick Bayer, Nathaniel O Keohane, and Christopher Timmins. “Migration and hedonic valuation: The case of air quality”. In: *Journal of Environmental Economics and Management* 58.1 (2009), pp. 1–14.

[30] Helena Holmlund, Mikael Lindahl, and Erik Plug. “The causal effect of parent’s schooling on children’s schooling: a comparison of estimation methods”. In: *IZA Discussion Papers* 3630 (2008).

[31] Jonathan Mellon. “Rain, Rain, Go Away: 192 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable”. In: *SSRN Electronic Journal* (2022).

[32] Anna G.C. Boef et al. “Physician’s prescribing preference as an instrumental variable”. In: *Epidemiology* (Nov. 2015), p. 1. DOI: 10.1097/ede.0000000000000425. URL: https://doi.org/10.1097/ede.0000000000000425.

[33] Gregor Mendel. “Versuche über Plflanzen-hybriden”. In: *Verhandlungen des naturforschenden Vereines in Brünn, Bd* (1866). URL: http://www.esp.org/foundations/genetics/classical/gm-65.pdf.

[34] Walter S. Sutton. “The Chromosomes In Heredity”. In: *The Biological Bulletin* 4.5 (Apr. 1903), pp. 231–250. DOI: 10.2307/1535741. URL: https://doi.org/10.2307/1535741.

[35] Peter S. Harper. *A Short History of Medical Genetics*. Oxford University Press, Nov. 2008. DOI: 10.1093/med/9780195187502.001.0001. URL: https://doi.org/10.1093/med/9780195187502.001.0001.

[36] H S Jennings. “The Numerical Results Of Diverse Systems Of Breeding, With Respect To Two Pairs Of Characters, Linked Or Independent, With Special Relation To The Effects Of

Linkage". In: *Genetics* 2.2 (Mar. 1917), pp. 97–154. DOI: 10.1093/genetics/2.2.97. URL: https://doi.org/10.1093/genetics/2.2.97.

[37] George Davey Smith and Shah Ebrahim. "'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*". In: *International Journal of Epidemiology* 32.1 (Feb. 2003), pp. 1–22. DOI: 10.1093/ije/dyg070. URL: https://doi.org/10.1093/ije/dyg070.

[38] Eleanor Sanderson et al. "Mendelian randomization". In: *Nature Reviews Methods Primers* 2.1 (Feb. 2022). ISSN: 2662-8449. DOI: 10.1038/s43586-021-00092-5. URL: http://dx.doi.org/10.1038/s43586-021-00092-5.

[39] Lina Chen et al. "Alcohol Intake and Blood Pressure: A Systematic Review Implementing a Mendelian Randomization Approach". In: *PLoS Medicine* 5.3 (Mar. 2008). Ed. by Cosetta Minelli, e52. DOI: 10.1371/journal.pmed.0050052. URL: https://doi.org/10.1371/journal.pmed.0050052.

[40] Ancel Keys et al. "Serum Cholesterol And Cancer Mortality In The Seven Countries Study". In: *American Journal of Epidemiology* 121.6 (June 1985), pp. 870–883. DOI: 10.1093/oxfordjournals.aje.a114057. URL: https://doi.org/10.1093/oxfordjournals.aje.a114057.

[41] Martijn B. Katan. "Apolipoprotein E Isoforms, Serum Cholesterol, And Cancer". In: *The Lancet* 327.8479 (Mar. 1986), pp. 507–508. DOI: 10.1016/s0140-6736(86)92972-7. URL: https://doi.org/10.1016/s0140-6736(86)92972-7.

[42] R Gray and K Wheatley. "How to avoid bias when comparing bone marrow transplantation with chemotherapy". In: *Bone Marrow Transplant* (1991). URL: pubmed.ncbi.nlm.nih.gov/1855097/.

[43] Jeffrey M. Wooldridge. "Instrumental Variables Estimation and Two-Stage Least Squares". In: *Introductory Econometrics: A Modern Approach*. 7th. Cengage Learning, 2019. Chap. 15, pp. 689–723.

[44] Vanessa Didelez, Sha Meng, and Nuala A. Sheehan. “Assumptions of IV Methods for Observational Epidemiology”. In: *Statistical Science* 25.1 (Feb. 2010). DOI: 10.1214/09-sts316. URL: https://doi.org/10.1214/09-sts316.

[45] P.-F. Verhulst. “Notice sur la loi que la population poursuit dans son accroissement”. In: *Correspondance Mathématique et Physique* 10 (1838), pp. 113–121.

[46] William H Greene. *Econometric Analysis*. Pearson Education, 2018.

[47] Kouichi Ozaki et al. “Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction”. In: *Nature Genetics* 32.4 (Nov. 2002), pp. 650–654. DOI: 10.1038/ng1047. URL: https://doi.org/10.1038/ng1047.

[48] Jonathan Marchini and Bryan Howie. “Genotype imputation for genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (June 2010), pp. 499–511. DOI: 10.1038/nrg2796. URL: https://doi.org/10.1038/nrg2796.

[49] John P. A. Ioannidis, Thomas A. Trikalinos, and Muin J. Khoury. “Implications of Small Effect Sizes of Individual Genetic Variants on the Design and Interpretation of Genetic Association Studies of Complex Diseases”. In: *American Journal of Epidemiology* 164.7 (Aug. 2006), pp. 609–614. DOI: 10.1093/aje/kwj259. URL: https://doi.org/10.1093/aje/kwj259.

[50] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. In: *Cell* 169.7 (June 2017), pp. 1177–1186. DOI: 10.1016/j.cell.2017.05.038. URL: https://doi.org/10.1016/j.cell.2017.05.038.

[51] European Union. *Recital 34 of the EU General Data Protection Regulation (GDPR)*. https://gdpr-info.eu/recitals/no-34/. Accessed on March 1, 2023. 2016.

[52] Stephen Burgess et al. “Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors”. In: *European Journal of Epidemiology* 30.7 (Mar. 2015), pp. 543–552. DOI: 10.1007/s10654-015-0011-z. URL: https://doi.org/10.1007/s10654-015-0011-z.

[53] Vanessa Didelez and Nuala Sheehan. “Mendelian randomization as an instrumental variable approach to causal inference”. In: *Statistical Methods in Medical Research* 16.4 (Aug.

2007), pp. 309–330. DOI: 10.1177/0962280206077743. URL: https://doi.org/10.1177/0962280206077743.

[54] Julian PT Higgins et al. “Chapter 9: Analysing data and undertaking meta-analyses”. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Ed. by Julian PT Higgins et al. 2nd ed. Chichester, UK: John Wiley & Sons, 2019, pp. 241–284. DOI: 10.1002/9781119536604.ch9. URL: https://handbook-5-1.cochrane.org/chapter_9/.

[55] Joanne E McKenzie et al. “Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis”. In: *Cochrane Handbook for Systematic Reviews of Interventions*. Ed. by Julian PT Higgins et al. John Wiley & Sons, 2019.

[56] Qingyuan Zhao et al. “Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples”. In: *Statist. Sci.* 34.2 (May 2019), pp. 317–333. DOI: 10.1214/18-STS692. URL: https://doi.org/10.1214/18-STS692.

[57] Ashish Patel et al. *Conditional inference in cis-Mendelian randomization using weak genetic factors*. 2020. DOI: 10.48550/ARXIV.2005.01765. URL: https://arxiv.org/abs/2005.01765.

[58] Olena O Yavorska and Stephen Burgess. “MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data”. In: *International Journal of Epidemiology* 46.6 (Apr. 2017), pp. 1734–1739. DOI: 10.1093/ije/dyx034. URL: https://doi.org/10.1093/ije/dyx034.

[59] Fatima Batool et al. “Disentangling the effects of traits with shared clustered genetic predictors using multivariable Mendelian randomization”. In: *Genetic Epidemiology* 46.7 (May 2022), pp. 415–429. DOI: 10.1002/gepi.22462. URL: https://doi.org/10.1002/gepi.22462.

[60] *LD-based result clumping procedure*. https://zzz.bwh.harvard.edu/plink/clump.shtml. Accessed: 23-04-2023.

[61] Florian Prive. *Why clumping should be preferred over pruning*. URL: https://privefl.github.io/bigsnpr/articles/pruning-vs-clumping.html.

[62] Tom M Palmer et al. “Using multiple genetic variants as instrumental variables for modifiable risk factors”. In: *Statistical Methods in Medical Research* 21.3 (Jan. 2011), pp. 223–242. DOI: 10.1177/0962280210394459. URL: https://doi.org/10.1177/0962280210394459.

[63] Stephen Burgess and Simon G Thompson and. “Avoiding bias from weak instruments in Mendelian randomization studies”. In: *International Journal of Epidemiology* 40.3 (Mar. 2011), pp. 755–764. DOI: 10.1093/ije/dyr036. URL: https://doi.org/10.1093/ije/dyr036.

[64] David H Richardson. “The Exact Distribution of a Structural Coefficient Estimator”. In: *Journal of the American Statistical Association* 63.324 (1968), pp. 549–556.

[65] Nadia Solovieff et al. “Pleiotropy in complex traits: challenges and strategies”. In: *Nature Reviews Genetics* 14.7 (June 2013), pp. 483–495. DOI: 10.1038/nrg3461. URL: https://doi.org/10.1038/nrg3461.

[66] Shanya Sivakumaran et al. “Abundant Pleiotropy in Human Complex Diseases and Traits”. In: *The American Journal of Human Genetics* 89.5 (Nov. 2011), pp. 607–618. DOI: 10.1016/j.ajhg.2011.10.004. URL: https://doi.org/10.1016/j.ajhg.2011.10.004.

[67] J. Bowden, G. Davey Smith, and S. Burgess. “Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression”. In: *International Journal of Epidemiology* 44.2 (Apr. 2015), pp. 512–525. DOI: 10.1093/ije/dyv080. URL: https://doi.org/10.1093/ije/dyv080.

[68] M. Egger et al. “Bias in meta-analysis detected by a simple, graphical test”. In: *BMJ* 315.7109 (Sept. 1997), pp. 629–634. DOI: 10.1136/bmj.315.7109.629. URL: https://doi.org/10.1136/bmj.315.7109.629.

[69] J. Bowden et al. “Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator”. In: *Genetic Epidemiology* 40 (4 2016), pp. 304–14. DOI: 10.1002/gepi.21965.

[70] S. Burgess and S. G. Thompson. “Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects”. In: *American Journal of Epidemiology* 181.4

(Jan. 2015), pp. 251–260. DOI: 10.1093/aje/kwu283. URL: https://doi.org/10.1093/aje/kwu283.

[71] Trang Quynh Nguyen, Ian Schmid, and Elizabeth A. Stuart. “Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn.” In: *Psychological Methods* 26.2 (Apr. 2021), pp. 255–271. DOI: 10.1037/met0000299. URL: https://doi.org/10.1037/met0000299.

[72] Alice R. Carter et al. “Mendelian randomisation for mediation analysis: current methods and challenges for implementation”. In: *European Journal of Epidemiology* 36.5 (May 2021), pp. 465–478. DOI: 10.1007/s10654-021-00757-1. URL: https://doi.org/10.1007/s10654-021-00757-1.

[73] Vasilios Karageorgiou et al. “Weak and pleiotropy robust sex-stratified Mendelian randomization in the one sample and two sample settings”. In: *Genetic Epidemiology* 47.2 (Jan. 2023), pp. 135–151. DOI: 10.1002/gepi.22512. URL: https://doi.org/10.1002/gepi.22512.

[74] Wes Spiller et al. “Interaction-based Mendelian randomization with measured and unmeasured gene-by-covariate interactions”. In: *PLOS ONE* 17.8 (Aug. 2022). Ed. by Momiao Xiong, e0271933. DOI: 10.1371/journal.pone.0271933. URL: https://doi.org/10.1371/journal.pone.0271933.

[75] Wes Spiller et al. “Detecting and correcting for bias in Mendelian randomization analyses using Gene-by-Environment interactions”. In: *International Journal of Epidemiology* (Nov. 2018). DOI: 10.1093/ije/dyy204. URL: https://doi.org/10.1093/ije/dyy204.

[76] Eric J. Tchetgen Tchetgen, BaoLuo Sun, and Stefan Walter. “The GENIUS Approach to Robust Mendelian Randomization Inference”. In: *arXiv.org* (2017), pp. 650–654. DOI: 10.48550/ARXIV.1709.07779. URL: https://arxiv.org/abs/1709.07779.

[77] Arthur Lewbel. “Using Heteroscedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models”. In: *Journal of Business &amp Economic Statistics* 30.1 (Jan. 2012), pp. 67–80. DOI: 10.1080/07350015.2012.643126. URL: https://doi.org/10.1080/07350015.2012.643126.

[78] Hudson Reddon, Jean-Louis Guéant, and David Meyre. “The importance of gene–environment interactions in human obesity”. In: *Clinical Science* 130.18 (Aug. 2016), pp. 1571–1597. DOI: 10.1042/cs20160221. URL: https://doi.org/10.1042/cs20160221.

[79] Stephen Burgess, Adam Butterworth, and Simon G. Thompson. “Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data”. In: *Genetic Epidemiology* 37.7 (Sept. 2013), pp. 658–665. DOI: 10.1002/gepi.21758. URL: https://doi.org/10.1002/gepi.21758.

[80] Jack Bowden, George Davey Smith, and Stephen Burgess. “Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression”. In: *International Journal of Epidemiology* 44.2 (June 2015), pp. 512–525. ISSN: 0300-5771. DOI: 10.1093/ije/dyv080. eprint: http://oup.prod.sis.lan/ije/article-pdf/44/2/512/2266954/dyv080.pdf. URL: https://doi.org/10.1093/ije/dyv080.

[81] Rebecca DerSimonian and Nan Laird. “Meta-analysis in clinical trials”. In: *Controlled Clinical Trials* 7.3 (Sept. 1986), pp. 177–188. DOI: 10.1016/0197-2456(86)90046-2. URL: https://doi.org/10.1016/0197-2456(86)90046-2.

[82] Fernando Pires Hartwig and Neil Martin Davies. “Why internal weights should be avoided (not only) in MR-Egger regression”. In: *International Journal of Epidemiology* 45.5 (Sept. 2016), pp. 1676–1678. ISSN: 0300-5771. DOI: 10.1093/ije/dyw240. eprint: https://academic.oup.com/ije/article-pdf/45/5/1676/7383008/dyw240.pdf. URL: https://doi.org/10.1093/ije/dyw240.

[83] Jack Bowden et al. “Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic”. In: *International Journal of Epidemiology* (Sept. 2016), dyw220. DOI: 10.1093/ije/dyw220. URL: https://doi.org/10.1093/ije/dyw220.

[84] Qingyuan Zhao et al. *Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score*. 2018. DOI: 10.48550/ARXIV.1801.09652. URL: https://arxiv.org/abs/1801.09652.

[85] Ciarrah Barry et al. “Exploiting collider bias to apply two-sample summary data Mendelian randomization methods to one-sample individual level data”. In: *PLOS Genetics* 17.8 (Aug. 2021). Ed. by Heather J Cordell, e1009703. DOI: 10.1371/journal.pgen.1009703. URL: https://doi.org/10.1371/journal.pgen.1009703.

[86] Qingyuan Zhao et al. “Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score”. In: *The Annals of Statistics* 48.3 (2020), pp. 1742–1769. DOI: 10.1214/19-AOS1866. URL: https://doi.org/10.1214/19-AOS1866.

[87] Cathie Sudlow et al. “UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age”. In: *PLOS Medicine* 12.3 (Mar. 2015), e1001779. DOI: 10.1371/journal.pmed.1001779. URL: https://doi.org/10.1371/journal.pmed.1001779.

[88] Dmitry Shungin et al. “New genetic loci link adipose and insulin biology to body fat distribution”. In: *Nature* 518.7538 (Feb. 2015), p. 187. DOI: 10.1038/nature14132.

[89] Sara L Pulit et al. “Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry”. In: *Human Molecular Genetics* 28.1 (Sept. 2018), pp. 166–174. DOI: 10.1093/hmg/ddy327. URL: https://doi.org/10.1093/hmg/ddy327.

[90] Jack Bowden et al. “Connecting Instrumental Variable methods for causal inference to the Estimand Framework”. In: *Statistics in Medicine* 40 (2021), pp. 5605–5627.

[91] Jack Bowden et al. “The Triangulation WIthin a STudy (TWIST) framework for causal inference within pharmacogenetic research”. In: *PLoS Genetics* 17 (9 2021), e1009783. ISSN: 1553-7390.

[92] Stijn Vansteelandt et al. “On Instrumental Variables Estimation of Causal Odds Ratios”. In: *Statistical Science* 26.3 (2011), pp. 403–422. ISSN: 08834237. URL: http://www.jstor.org/stable/23059139.

[93] Anders Huitfeldt, Mats J. Stensrud, and Etsuji Suzuki. “On the collapsibility of measures of effect in the counterfactual causal framework”. In: *Emerging Themes in Epidemiology* 16 (1 2019), pp. 1–5.

[94] Marie Verbanck et al. “Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases”. In: *Nature Genetics* 50.5 (Apr. 2018), pp. 693–698. DOI: 10.1038/s41588-018-0099-7. URL: https://doi.org/10.1038/s41588-018-0099-7.

[95] Daphne P Guh et al. “The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis”. In: *BMC Public Health* 9.1 (Mar. 2009). DOI: 10.1186/1471-2458-9-88. URL: https://doi.org/10.1186/1471-2458-9-88.

[96] Haris Riaz et al. “Association Between Obesity and Cardiovascular Outcomes”. In: *JAMA Network Open* 1.7 (Nov. 2018), e183788. DOI: 10.1001/jamanetworkopen.2018.3788. URL: https://doi.org/10.1001/jamanetworkopen.2018.3788.

[97] S. Burgess and S. G. Thompson. “Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects”. In: *American Journal of Epidemiology* 181.4 (Jan. 2015), pp. 251–260. DOI: 10.1093/aje/kwu283. URL: https://doi.org/10.1093/aje/kwu283.

[98] Stephen Burgess and Eric Harshfield. “Mendelian randomization to assess causal effects of blood lipids on coronary heart disease”. In: *Curr Opin Endocrinol Diabetes Obes* 23.2 (Apr. 2016), pp. 124–130. DOI: 10.1097/med.0000000000000230. URL: https://doi.org/10.1097/med.0000000000000230.

[99] Eleanor Sanderson et al. “An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings”. In: *International Journal of Epidemiology* 48.3 (Dec. 2018), pp. 713–727. ISSN: 0300-5771. DOI: 10.1093/ije/dyy262. eprint: https://academic.oup.com/ije/article-pdf/48/3/713/29958854/dyy262.pdf. URL: https://doi.org/10.1093/ije/dyy262.

[100] Eleanor Sanderson, Wes Spiller, and Jack Bowden. “Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization”. In: *Statistics in Medicine* 40.25 (2021), pp. 5434–5452. DOI: 10.1002/sim.9133. URL: https://doi.org/10.1002/sim.9133.

[101] Jingshu Wang et al. “Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments”. In: *PLOS Genetics* 17.6 (June 2021). Ed. by Xiaofeng Zhu, e1009575. DOI: 10.1371/journal.pgen.1009575. URL: https://doi.org/10.1371/journal.pgen.1009575.

[102] Johannes Kettunen et al. “Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA”. In: *Nature Communications* 7.1 (Mar. 2016). DOI: 10.1038/ncomms11122. URL: https://doi.org/10.1038%2Fncomms11122.

[103] Vasileios Karageorgiou et al. “Sparse dimensionality reduction approaches in Mendelian randomization with highly correlated exposures”. In: *eLife* 12 (Apr. 2023). DOI: 10.7554/elife.80063. URL: https://doi.org/10.7554/elife.80063.

[104] Vasilis Karageorgiou et al. “Selection Algorithms for Conditionally Weak Instruments in MVMR”. In: *Mendelian Randomization Conference*. University of Bristol. 2021.

[105] Vasilis Karageorgiou et al. “Dimensionality Reduction Approaches in MVMR”. In: *14th International Conference of the ERCIM WG on Computational and Methodological Statistics*. King’s College London. 2021.

[106] I.T. Jolliffe. *Principal Component Analysis*. 2nd. New York, NY: Springer, 1986. ISBN: 978-0-387-95442-4. DOI: 10.1007/978-1-4757-1904-8.

[107] Hui Zou, Trevor Hastie, and Robert Tibshirani. “Sparse Principal Component Analysis”. In: *Journal of Computational and Graphical Statistics* 15.2 (June 2006), pp. 265–286. DOI: 10.1198/106186006x113430. URL: https://doi.org/10.1198/106186006x113430.

[108] Hui Zou and Lingzhou Xue. “A Selective Overview of Sparse Principal Component Analysis”. In: *Proceedings of the IEEE* 106.8 (Aug. 2018), pp. 1311–1320. DOI: 10.1109/jproc.2018.2846588. URL: https://doi.org/10.1109/jproc.2018.2846588.

[109] Peter J. Rousseeuw and Christophe Croux. “Alternatives to the Median Absolute Deviation”. In: *Journal of the American Statistical Association* 88.424 (Dec. 1993), pp. 1273–1283. DOI: 10.1080/01621459.1993.10476408. URL: https://doi.org/10.1080/01621459.1993.10476408.

[110] Alan Heckert. *Qn Scale*. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/qn_scale.htm. [Online; accessed 19-July-2021]. 2003.

[111] Jian Guo et al. "Principal Component Analysis With Sparse Fused Loadings". In: *Journal of Computational and Graphical Statistics* 19.4 (Jan. 2010), pp. 930–946. DOI: 10.1198/jcgs.2010.08127. URL: https://doi.org/10.1198/jcgs.2010.08127.

[112] Fan Chen and Karl Rohe. *A New Basis for Sparse Principal Component Analysis*. 2021. arXiv: 2007.00596 [stat.ML].

[113] Kevin R. Coombes and Min Wang. *PCDimension: Finding the Number of Significant Principal Components*. https://CRAN.R-project.org/package=PCDimension. [Online; accessed 19-July-2021]. 2019.

[114] Dimitris Karlis, Gilbert Saporta, and Antonis Spinakis. "A Simple Rule for the Selection of Principal Components". In: *Communications in Statistics - Theory and Methods* 32.3 (Jan. 2003), pp. 643–666. DOI: 10.1081/sta-120018556. URL: https://doi.org/10.1081/sta-120018556.

[115] I. T. Jolliffe. *Principal component analysis*. New York: Springer, 2002. ISBN: 0387954422.

[116] Wayne F. Velicer. "Determining the number of components from the matrix of partial correlations". In: *Psychometrika* 41.3 (Sept. 1976), pp. 321–327. DOI: 10.1007/bf02293557. URL: https://doi.org/10.1007/bf02293557.

[117] Christophe Croux, Peter Filzmoser, and Heinrich Fritz. "Robust Sparse Principal Component Analysis". In: *Technometrics* 55.2 (May 2013), pp. 202–214. DOI: 10.1080/00401706.2012.727746. URL: https://doi.org/10.1080/00401706.2012.727746.

[118] Verena Zuber et al. "High-throughput multivariable Mendelian randomization analysis prioritizes apolipoprotein B as key lipid risk factor for coronary artery disease". In: *International Journal of Epidemiology* (Feb. 2020). DOI: 10.1101/2020.02.10.20021691. URL: https://doi.org/10.1101%2F2020.02.10.20021691.

[119] Donghoh Kim and Se-Kang Kim. "Comparing patterns of component loadings: Principal Component Analysis (PCA) versus Independent Component Analysis (ICA) in analyzing multivari-

ate non-normal data". In: *Behavior Research Methods* 44.4 (Feb. 2012), pp. 1239–1243. DOI: 10.3758/s13428-012-0193-1. URL: https://doi.org/10.3758/s13428-012-0193-1.

[120] Johannes B. Reitsma et al. "Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews". In: *Journal of Clinical Epidemiology* 58.10 (Oct. 2005), pp. 982–990. DOI: 10.1016/j.jclinepi.2005.02.022. URL: https://doi.org/10.1016/j.jclinepi.2005.02.022.

[121] UK Biobank. *Nightingale Health and UK Biobank announces major initiative to analyse half a million blood samples to facilitate global medical research*. https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/nightingale-health-and-uk-biobank-announces-major-initiative-to-analyse-half-a-million-blood-samples-to-facilitate-global-medical-research. [Online; accessed 19-July-2021]. 2018.

[122] Chobufo Ditah et al. "Small and medium sized HDL particles are protectively associated with coronary calcification in a cross-sectional population-based sample". In: *Atherosclerosis* 251 (Aug. 2016), pp. 124–131. DOI: 10.1016/j.atherosclerosis.2016.06.010. URL: https://doi.org/10.1016/j.atherosclerosis.2016.06.010.

[123] Xuedong Wang et al. "Small HDL subclass is associated with coronary plaque stability: An optical coherence tomography study in patients with coronary artery disease". In: *Journal of Clinical Lipidology* 13.2 (Mar. 2019), 326–334.e2. DOI: 10.1016/j.jacl.2018.12.002. URL: https://doi.org/10.1016/j.jacl.2018.12.002.

[124] Hugh A. Chipman and Hong Gu. "Interpretable dimension reduction". In: *Journal of Applied Statistics* 32.9 (Nov. 2005), pp. 969–987. DOI: 10.1080/02664760500168648. URL: https://doi.org/10.1080/02664760500168648.

[125] Jonathan Sulc et al. "Composite trait Mendelian Randomization reveals distinct metabolic and lifestyle consequences of differences in body shape". In: *Communications Biology* (Sept. 2020). DOI: 10.1101/2020.09.03.20187567. URL: https://doi.org/10.1101/2020.09.03.20187567.

[126] Helian Feng et al. "Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association

studies". In: *PLOS Genetics* 17.4 (Apr. 2021). Ed. by Michael P. Epstein, e1008973. DOI: 10.1371/journal.pgen.1008973. URL: https://doi.org/10.1371/journal.pgen.1008973.

[127] Qingyuan Zhao et al. "A Mendelian randomization study of the role of lipoprotein subfractions in coronary artery disease". In: *eLife* 10 (Apr. 2021). DOI: 10.7554/elife.58361. URL: https://doi.org/10.7554/elife.58361.

[128] Anatol Kontush. "HDL particle number and size as predictors of cardiovascular disease". In: *Frontiers in Pharmacology* 6 (Oct. 2015). DOI: 10.3389/fphar.2015.00218. URL: https://doi.org/10.3389/fphar.2015.00218.

[129] Andrew J. Grant and Stephen Burgess. "Pleiotropy robust methods for multivariable Mendelian randomization". In: *Statistics in Medicine* 40.26 (Aug. 2021), pp. 5813–5830. DOI: 10.1002/sim.9156. URL: https://doi.org/10.1002/sim.9156.

[130] Spencer L James, Degu Abate, and et al. "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017". In: 392.10159 (Nov. 2018), pp. 1789–1858. DOI: 10.1016/s0140-6736(18)32279-7. URL: https://doi.org/10.1016/s0140-6736(18)32279-7.

[131] Navneet Kapur et al. *Depression in adults: treatment and management. Consultation Draft*. 2018. URL: https://www.nice.org.uk/guidance/gid-cgwave0725/documents/full-guideline-updated.

[132] Anthony Cleare et al. "Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines". In: 29.5 (May 2015). Ed. by David J Nutt and Pierre Blier, pp. 459–525. DOI: 10.1177/0269881115581093. URL: https://doi.org/10.1177/0269881115581093.

[133] Andrea Cipriani et al. "Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis". In: 391.10128 (Apr. 2018), pp. 1357–1366. DOI: 10.1016/s0140-6736(17)32802-7. URL: https://doi.org/10.1016/s0140-6736(17)32802-7.

[134] Philip Brenner et al. “Treatment-resistant depression as risk factor for substance use disorders—a nation-wide register-based cohort study”. In: 114.7 (Apr. 2019), pp. 1274–1282. DOI: 10.1111/add.14596. URL: https://doi.org/10.1111/add.14596.

[135] Chiara Fabbri et al. “Genetic and clinical characteristics of treatment-resistant depression using primary care records in two UK cohorts”. In: 26.7 (Mar. 2021), pp. 3363–3373. DOI: 10.1038/s41380-021-01062-9. URL: https://doi.org/10.1038/s41380-021-01062-9.

[136] Floriana S. Luppino et al. “Overweight, Obesity, and Depression”. In: *Archives of General Psychiatry* 67.3 (Mar. 2010), p. 220. DOI: 10.1001/archgenpsychiatry.2010.2. URL: https://doi.org/10.1001/archgenpsychiatry.2010.2.

[137] Christine Emmer, Michael Bosnjak, and Jutta Mata. “The association between weight stigma and mental health: A meta-analysis”. In: *Obesity Reviews* 21.1 (Sept. 2019). DOI: 10.1111/obr.12935. URL: https://doi.org/10.1111/obr.12935.

[138] Hanieh Yaghootkar et al. “Genetic Evidence for a Link Between Favorable Adiposity and Lower Risk of Type 2 Diabetes, Hypertension, and Heart Disease”. In: *Diabetes* 65.8 (Apr. 2016), pp. 2448–2460. DOI: 10.2337/db15-1671. URL: https://doi.org/10.2337/db15-1671.

[139] Francesco Casanova et al. “Higher adiposity and mental health: causal inference using Mendelian randomization”. In: *Human Molecular Genetics* 30.24 (July 2021), pp. 2371–2382. DOI: 10.1093/hmg/ddab204. URL: https://doi.org/10.1093/hmg/ddab204.

[140] Olivia Patsalos et al. “Diet, Obesity, and Depression: A Systematic Review”. In: *Journal of Personalized Medicine* 11.3 (Mar. 2021), p. 176. DOI: 10.3390/jpm11030176. URL: https://doi.org/10.3390/jpm11030176.

[141] Naomi R. Wray and et al. “Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression”. In: 50.5 (Apr. 2018), pp. 668–681. DOI: 10.1038/s41588-018-0090-3. URL: https://doi.org/10.1038/s41588-018-0090-3.

[142] Rune Aabenhus et al. “Biomarkers as point-of-care tests to guide prescription of antibiotics in patients with acute respiratory infections in primary care”. In: (Nov. 2014). DOI: 10.1002/14651858.cd010130.pub2. URL: https://doi.org/10.1002/14651858.cd010130.pub2.

[143] Emerging Risk Factors Collaboration et al. “C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis”. In: *The Lancet* 375.9709 (Jan. 2010), pp. 132–140. DOI: 10.1016/s0140-6736(09)61717-7. URL: https://doi.org/10.1016/s0140-6736(09)61717-7.

[144] C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC). “Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data”. In: *BMJ* 342.feb15 2 (Feb. 2011), pp. d548–d548. DOI: 10.1136/bmj.d548. URL: https://doi.org/10.1136/bmj.d548.

[145] Samuel R. Chamberlain et al. “Treatment-resistant depression and peripheral C-reactive protein”. In: 214.1 (May 2018), pp. 11–19. DOI: 10.1192/bjp.2018.66. URL: https://doi.org/10.1192/bjp.2018.66.

[146] Paul R. Rosenbaum. “From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment”. In: *Journal of the American Statistical Association* 79.385 (Mar. 1984), pp. 41–48. DOI: 10.1080/01621459.1984.10477060. URL: https://doi.org/10.1080/01621459.1984.10477060.

[147] Stephen Burgess, Dylan S Small, and Simon G Thompson. “A review of instrumental variable estimators for Mendelian randomization”. In: *Statistical Methods in Medical Research* 26.5 (2017). PMID: 26282889, pp. 2333–2355. DOI: 10.1177/0962280215597579. eprint: https://doi.org/10.1177/0962280215597579. URL: https://doi.org/10.1177/0962280215597579.

[148] Nils Kappelmann et al. “Dissecting the Association Between Inflammation, Metabolic Dysregulation, and Specific Depressive Symptoms”. In: *JAMA Psychiatry* 78.2 (Feb. 2021), p. 161. DOI: 10.1001/jamapsychiatry.2020.3436. URL: https://doi.org/10.1001/jamapsychiatry.2020.3436.

[149] Fernando Pires Hartwig et al. “Body mass index and psychiatric disorders: a Mendelian randomization study”. In: *Scientific Reports* 6.1 (Sept. 2016). DOI: 10.1038/srep32730. URL: https://doi.org/10.1038/srep32730.

[150] Jessica Tyrrell et al. “Using genetics to understand the causal influence of higher BMI on depression”. In: *International Journal of Epidemiology* 48.3 (Nov. 2018), pp. 834–848. DOI: 10.1093/ije/dyy223. URL: https://doi.org/10.1093/ije/dyy223.

[151] Maria S. Speed et al. “Investigating the association between body fat and depression via Mendelian randomization”. In: *Translational Psychiatry* 9.1 (Aug. 2019). DOI: 10.1038/s41398-019-0516-4. URL: https://doi.org/10.1038/s41398-019-0516-4.

[152] Gibran Hemani et al. “The MR-Base platform supports systematic causal inference across the human phenome”. In: 7 (May 2018). DOI: 10.7554/elife.34408. URL: https://doi.org/10.7554/elife.34408.

[153] Eleanor Sanderson et al. “An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings”. In: *International Journal of Epidemiology* 48.3 (Dec. 2018), pp. 713–727. DOI: 10.1093/ije/dyy262. URL: https://doi.org/10.1093/ije/dyy262.

[154] T. M. Frayling et al. “A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity”. In: *Science* 316.5826 (May 2007), pp. 889–894. DOI: 10.1126/science.1141634. URL: https://doi.org/10.1126/science.1141634.

[155] “Six new loci associated with body mass index highlight a neuronal influence on body weight regulation”. In: *Nature Genetics* 41.1 (Dec. 2008), pp. 25–34. DOI: 10.1038/ng.287. URL: https://doi.org/10.1038/ng.287.

[156] Symen Ligthart et al. “Genome Analyses of 200, 000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders”. In: *The American Journal of Human Genetics* 103.5 (Nov. 2018), pp. 691–706. DOI: 10.1016/j.ajhg.2018.09.009. URL: https://doi.org/10.1016/j.ajhg.2018.09.009.

[157] Svetlana Puzhko et al. “Excess body weight as a predictor of response to treatment with antidepressants in patients with depressive disorder”. In: 267 (Apr. 2020), pp. 153–170. DOI: 10.1016/j.jad.2020.01.113. URL: https://doi.org/10.1016/j.jad.2020.01.113.

[158] Meg Fluharty et al. “The Association of Cigarette Smoking With Depression and Anxiety: A Systematic Review”. In: *Nicotine & Tobacco Research* 19.1 (May 2016), pp. 3–13. DOI: 10.1093/ntr/ntw140. URL: https://doi.org/10.1093/ntr/ntw140.

[159] Susan Martin et al. “Disease consequences of higher adiposity uncoupled from its adverse metabolic effects using Mendelian randomisation”. In: *eLife* 11 (Jan. 2022). DOI: 10.7554/elife.72452. URL: https://doi.org/10.7554/elife.72452.

[160] Katrina A. S. Davis et al. “Mental health in UK Biobank – development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis”. In: *BJPsych Open* 6.2 (Feb. 2020). DOI: 10.1192/bjo.2019.100. URL: https://doi.org/10.1192/bjo.2019.100.

[161] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. “The PHQ-9”. In: 16.9 (Sept. 2001), pp. 606–613. DOI: 10.1046/j.1525-1497.2001.016009606.x. URL: https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

[162] Ronald C. Kessler and T. Bedirhan Üstün. “The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI)”. In: 13.2 (June 2004), pp. 93–121. DOI: 10.1002/mpr.168. URL: https://doi.org/10.1002/mpr.168.

[163] Na Cai et al. “Minimal phenotyping yields genome-wide association signals of low specificity for major depression”. In: *Nature Genetics* 52.4 (Mar. 2020), pp. 437–447. DOI: 10.1038/s41588-020-0594-5. URL: https://doi.org/10.1038/s41588-020-0594-5.

[164] G. Taylor et al. “Change in mental health after smoking cessation: systematic review and meta-analysis”. In: *BMJ* 348.feb13 1 (Feb. 2014), g1151–g1151. DOI: 10.1136/bmj.g1151. URL: https://doi.org/10.1136/bmj.g1151.

[165] Harald H.H. Göring, Joseph D. Terwilliger, and John Blangero. “Large Upward Bias in Estimation of Locus-Specific Effects from Genomewide Scans”. In: *The American Journal of Human Genetics* 69.6 (Dec. 2001), pp. 1357–1369. DOI: 10.1086/324471. URL: https://doi.org/10.1086/324471.

[166] Symen Ligthart et al. “Pleiotropy among Common Genetic Loci Identified for Cardiometabolic Disorders and C-Reactive Protein”. In: *PLOS ONE* 10.3 (Mar. 2015). Ed. by Yun Li, e0118859. DOI: 10.1371/journal.pone.0118859. URL: https://doi.org/10.1371/journal.pone.0118859.

[167] Adam Locke et al. “Genetic studies of body mass index yield new insights for obesity biology”. In: *Nature* 518.7538 (Feb. 2015), pp. 197–206. DOI: 10.1038/nature14177.

[168] Daniel F. Levey et al. “Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in &gt1.2 million individuals highlight new therapeutic directions”. In: *Nature Neuroscience* 24.7 (May 2021), pp. 954–963. DOI: 10.1038/s41593-021-00860-2. URL: https://doi.org/10.1038/s41593-021-00860-2.

[169] Stephen Burgess et al. “Guidelines for performing Mendelian randomization investigations: update for summer 2023”. In: *Wellcome Open Research* 4 (Aug. 2023), p. 186. ISSN: 2398-502X. DOI: 10.12688/wellcomeopenres.15555.3. URL: http://dx.doi.org/10.12688/wellcomeopenres.15555.3.

[170] David Altshuler, Leonid Kruglyak, and Eric Lander. “Genetic Polymorphisms and Disease”. In: *New England Journal of Medicine* 338.22 (May 1998), pp. 1626–1626. DOI: 10.1056/nejm199805283382214. URL: https://doi.org/10.1056/nejm199805283382214.

[171] Alkes L. Price et al. “New approaches to population stratification in genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (June 2010), pp. 459–463. DOI: 10.1038/nrg2813. URL: https://doi.org/10.1038/nrg2813.

[172] J. D. Sargan. “The Estimation of Economic Relationships using Instrumental Variables”. In: *Econometrica* 26.3 (July 1958), p. 393. DOI: 10.2307/1907619. URL: https://doi.org/10.2307/1907619.

[173] Donald B. Rubin. “The Bayesian Bootstrap”. In: *The Annals of Statistics* 9.1 (Jan. 1981). DOI: 10.1214/aos/1176345338. URL: https://doi.org/10.1214/aos/1176345338.

[174] Ulrike E Maske et al. “Current major depressive syndrome measured with the Patient Health Questionnaire-9 (PHQ-9) and the Composite International Diagnostic Interview (CIDI): results

from a cross-sectional population-based study of adults in Germany". In: *BMC Psychiatry* 15.1 (Apr. 2015). DOI: 10.1186/s12888-015-0463-4. URL: https://doi.org/10.1186/s12888-015-0463-4.

[175] G. Rucker et al. "Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis". In: *Biostatistics* 12.1 (July 2010), pp. 122–142. DOI: 10.1093/biostatistics/kxq046. URL: https://doi.org/10.1093/biostatistics/kxq046.

[176] Jack Bowden et al. "Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression". In: *International Journal of Epidemiology* 47.4 (June 2018), pp. 1264–1278. DOI: 10.1093/ije/dyy101. URL: https://doi.org/10.1093/ije/dyy101.

[177] Genevieve M. Leyden et al. "Harnessing tissue-specific genetic variation to dissect putative causal pathways between body mass index and cardiometabolic phenotypes". In: *The American Journal of Human Genetics* 109.2 (Feb. 2022), pp. 240–252. DOI: 10.1016/j.ajhg.2021.12.013. URL: https://doi.org/10.1016/j.ajhg.2021.12.013.

[178] John Lonsdale, Jeffrey Thomas, and et al. "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6 (May 2013), pp. 580–585. DOI: 10.1038/ng.2653. URL: https://doi.org/10.1038/ng.2653.

[179] A Pan et al. "Bidirectional association between depression and obesity in middle-aged and older women". In: *International Journal of Obesity* 36.4 (June 2011), pp. 595–602. DOI: 10.1038/ijo.2011.111. URL: https://doi.org/10.1038/ijo.2011.111.

[180] Eléonore Beurel, Marisa Toups, and Charles B. Nemeroff. "The Bidirectional Relationship of Depression and Inflammation: Double Trouble". In: *Neuron* 107.2 (July 2020), pp. 234–256. DOI: 10.1016/j.neuron.2020.06.002. URL: https://doi.org/10.1016/j.neuron.2020.06.002.

[181] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)*. American Psychiatric Publishing, 2013.

[182] David M. Howard et al. “Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions”. In: *Nature Neuroscience* 22.3 (Feb. 2019), pp. 343–352. DOI: 10.1038/s41593-018-0326-7. URL: https://doi.org/10.1038/s41593-018-0326-7.

[183] Gibran Hemani, Kate Tilling, and George Davey Smith. “Orienting the causal relationship between imprecisely measured traits using GWAS summary data”. In: *PLOS Genetics* 13.11 (Nov. 2017). Ed. by Jun Li, e1007081. DOI: 10.1371/journal.pgen.1007081. URL: https://doi.org/10.1371/journal.pgen.1007081.

[184] Qingyuan Zhao et al. *Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score*. 2018. eprint: arXiv:1801.09652.

[185] Jingshu Wang et al. “Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments”. In: *PLOS Genetics* 17.6 (June 2021), pp. 1–24. DOI: 10.1371/journal.pgen.1009575. URL: https://doi.org/10.1371/journal.pgen.1009575.

[186] Kosuke Imai, Luke Keele, and Dustin Tingley. “A general approach to causal mediation analysis.” In: *Psychological Methods* 15.4 (2010), pp. 309–334. DOI: 10.1037/a0020761. URL: https://doi.org/10.1037/a0020761.

[187] Ari V. Ahola-Olli et al. “Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors”. In: *The American Journal of Human Genetics* 100.1 (Jan. 2017), pp. 40–50. DOI: 10.1016/j.ajhg.2016.11.007. URL: https://doi.org/10.1016/j.ajhg.2016.11.007.

[188] Ciarrah Barry et al. “Exploiting collider bias to apply two-sample summary data Mendelian randomization methods to one-sample individual level data”. In: (Oct. 2020). DOI: 10.1101/2020.10.20.20216358. URL: https://doi.org/10.1101/2020.10.20.20216358.

[189] Alish B. Palmos et al. “Assessing the Evidence for Causal Associations Between Body Mass Index, C-Reactive Protein, Depression, and Reported Trauma Using Mendelian Randomization”. In: *Biological Psychiatry Global Open Science* 3.1 (Jan. 2023), pp. 110–118. DOI: 10.1016/j.bpsgos.2022.01.003. URL: https://doi.org/10.1016/j.bpsgos.2022.01.003.

[190] Jonathan Flint. “The genetic basis of major depressive disorder”. In: *Molecular Psychiatry* (Jan. 2023). DOI: 10.1038/s41380-023-01957-9. URL: https://doi.org/10.1038/s41380-023-01957-9.

[191] Richard Border et al. “Cross-trait assortative mating is widespread and inflates genetic correlation estimates”. In: *Science* 378.6621 (Nov. 2022), pp. 754–761. DOI: 10.1126/science.abo2059. URL: https://doi.org/10.1126/science.abo2059.

[192] Eleonora Porcu et al. “Limited evidence for blood eQTLs in human sexual dimorphism”. In: *Genome Medicine* 14.1 (Aug. 2022). DOI: 10.1186/s13073-022-01088-w. URL: https://doi.org/10.1186/s13073-022-01088-w.

[193] Glenn A. Maston, Sara K. Evans, and Michael R. Green. “Transcriptional Regulatory Elements in the Human Genome”. In: *Annual Review of Genomics and Human Genetics* 7.1 (Sept. 2006), pp. 29–59. DOI: 10.1146/annurev.genom.7.080505.115623. URL: https://doi.org/10.1146/annurev.genom.7.080505.115623.

[194] Jonathan Sulc et al. “Composite trait Mendelian randomization reveals distinct metabolic and lifestyle consequences of differences in body shape”. In: *Communications Biology* 4.1 (Sept. 2021). DOI: 10.1038/s42003-021-02550-y. URL: https://doi.org/10.1038/s42003-021-02550-y.

[195] S Jackson. *Machine Learning — bookdown.org*. https://bookdown.org/ssjackson300/Machine-Learning-Lecture-Notes/. [Accessed 05-Jun-2023].

[196] Mark B. Pepys and Gideon M. Hirschfield. “C-reactive protein: a critical update”. In: *Journal of Clinical Investigation* 111.12 (June 2003), pp. 1805–1812. DOI: 10.1172/jci200318921. URL: https://doi.org/10.1172/jci200318921.

[197] Yangli Xie et al. “FGF/FGFR signaling in health and disease”. In: *Signal Transduction and Targeted Therapy* 5.1 (Sept. 2020). DOI: 10.1038/s41392-020-00222-7. URL: https://doi.org/10.1038/s41392-020-00222-7.

[198] Young Dong Yoo et al. “Fibroblast Growth Factor Regulates Human Neuroectoderm Specification Through ERK1/2-PARP-1 Pathway”. In: *Stem Cells* 29.12 (Nov. 2011), pp. 1975–1982. DOI: 10.1002/stem.758. URL: https://doi.org/10.1002/stem.758.

[199] Emanuele F. Osimo et al. “Inflammatory markers in depression: A meta-analysis of mean differences and variability in 5, 166 patients and 5, 083 controls”. In: *Brain, Behavior, and Immunity* 87 (July 2020), pp. 901–909. DOI: 10.1016/j.bbi.2020.02.010. URL: https://doi.org/10.1016/j.bbi.2020.02.010.

[200] Błażej Misiak et al. “Immune-inflammatory markers and psychosis risk: A systematic review and meta-analysis”. In: *Psychoneuroendocrinology* 127 (May 2021), p. 105200. DOI: 10.1016/j.psyneuen.2021.105200. URL: https://doi.org/10.1016/j.psyneuen.2021.105200.

[201] Nicholas G. Dowell et al. “Acute Changes in Striatal Microstructure Predict the Development of Interferon-Alpha Induced Fatigue”. In: *Biological Psychiatry* 79.4 (Feb. 2016), pp. 320–328. DOI: 10.1016/j.biopsych.2015.05.015. URL: https://doi.org/10.1016/j.biopsych.2015.05.015.

[202] Tom G Richardson et al. “Evaluating the direct effects of childhood adiposity on adult systemic metabolism: a multivariable Mendelian randomization analysis”. In: *International Journal of Epidemiology* 50.5 (Mar. 2021), pp. 1580–1592. DOI: 10.1093/ije/dyab051. URL: https://doi.org/10.1093/ije/dyab051.

[203] Tomáš Paus, Matcheri Keshavan, and Jay N. Giedd. “Why do many psychiatric disorders emerge during adolescence?” In: *Nature Reviews Neuroscience* 9.12 (Nov. 2008), pp. 947–957. DOI: 10.1038/nrn2513. URL: https://doi.org/10.1038/nrn2513.

[204] Cathy E. Elks et al. “Variability in the Heritability of Body Mass Index: A Systematic Review and Meta-Regression”. In: *Frontiers in Endocrinology* 3 (2012). DOI: 10.3389/fendo.2012.00029. URL: https://doi.org/10.3389/fendo.2012.00029.

[205] Robert Tibshirani. “Recent Advances in Post-Selection Statistical Inference”. In: *Conference on Neural Information Processing Systems*. Montreal, Dec. 2015.

[206] Guangsheng Pei et al. “deTS: tissue-specific enrichment analysis to decode tissue specificity”. In: *Bioinformatics* 35.19 (Mar. 2019). Ed. by Janet Kelso, pp. 3842–3845. DOI: 10.1093/bioinformatics/btz138. URL: https://doi.org/10.1093/bioinformatics/btz138.

[207] Claudia Giambartolomei et al. “Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics”. In: *PLoS Genetics* 10.5 (May 2014). Ed. by Scott M. Williams, e1004383. DOI: 10.1371/journal.pgen.1004383. URL: https://doi.org/10.1371/journal.pgen.1004383.