Sensitivity analyses for causal inference

By

Matthew James Tudball

MRC Integrative Epidemiology Unit
University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Department of Population Health Sciences.

September 2022

Word count: 46,521

Abstract

The validity of causal inference always rests on untestable assumptions. It is valuable to quantify the extent to which violations of different causal assumptions will alter the conclusions of a study. This is the motivation behind sensitivity analyses. The aim of this thesis is to develop sensitivity analyses for a variety of biases, including selection bias, coarsening bias in instrumental variable designs and familial and ancestral biases in genetic association studies.

The first chapter develops an approach to statistical inference in stochastic optimization problems when both the function to minimized, and the set over which it is minimized, must be estimated empirically. I apply this inference procedure to the problem of selection bias in large population cohorts such as UK Biobank. I propose a sensitivity analysis which is able to flexibly incorporate a wide variety of population-level information, while providing valid statistical inference.

The second chapter addresses the problem of coarsening bias in Mendelian randomization (MR) studies. In such studies, the exposure is often a coarsened approximation to some latent continuous trait. Genetically driven variation in the outcome can exist within categories of the exposure, violating the exclusion restriction. I derive a closed-form expression for the resulting bias and propose a simple correction that can be used with summary-level data to provide MR estimates with interpretable effect sizes.

The final chapter utilizes the increasing prevalence of within-family data to provide an "almost exact" approach to MR. I provide a formal justification for the validity of the MR design by building a causal model which includes features such as assortative mating, linkage disequilibrium, population stratification and transmission ratio distortion. I then propose an "almost exact" randomization test for MR based on explicitly modelling the distribution of crossovers. I apply this test to the Avon Longitudinal Study of Parents and Children (ALSPAC).

Dedication and acknowledgements

I am thankful for the support of my family and friends, and I am deeply grateful to my small army of supervisors Kate Tilling, Qingyuan Zhao, Rachael Hughes, Jack Bowden and George Davey Smith for their mentorship and advocacy.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Table of Contents

List of Tables

List of Figures

Figure **Page**

Chapter 1

Introduction

1.1 Foreword

Causal inference inevitably rests on untestable assumptions. A causal effect is characterized by a contrast of two or more possible interventions, only one of which can actually occur in a given individual at a given point in time (Holland, 1986). This fundamental problem means that a causal effect must typically be inferred by comparing sufficiently similar individuals who are differentiated by the intervention they experience. In practice, individuals often self-select or drop out of studies (Lu et al., 2022), measurements can be misreported or missing (Smeden, Lash, and Groenwold, 2020), and systematically different individuals may be more inclined to one intervention versus another (Hernán and Robins, 2020, p.25–37). These phenomena are ubiquitous in scientific research and can violate the assumptions on which causal inference relies, leading to bias.

It is crucial that scientific results are robust to these violations. The best way to mitigate bias is to prevent it at the outset. This can include ensuring comprehensive data collection, validating measurements and experimentally randomizing the receipt of interventions. This is not always possible, for logistical or financial reasons. The second best way to mitigate bias is to quantify the extent to which a study's conclusions would change if certain assumptions were violated. This is the principle behind sensitivity analysis.

Rosenbaum (2017, p.171) characterizes a sensitivity analysis as asking: "How would the results of a calculation, or the conclusion, change if the assumptions were changed by a limited amount? Would the conclusion barely change? If so, the conclusion is insensitive to a violation of the assumptions of that limited magnitude". In essence, a sensitivity analysis evaluates the stability of a study's conclusions to limited violations of the assumptions on which it is based.

The first example of a sensitivity analysis in an observational study is typically attributed to Cornfield et al. (1959). At the time of publication, there was an ongoing debate about the role of smoking in lung cancer incidence. The prominent statistician and geneticist Ronald A Fisher had argued against the evidence for a causal effect, instead suggesting that the association

could be driven by reverse causation, whereby the discomfort caused by a nascent tumour would cause sufferers to self-medicate via smoking (Stolley, 1991). Another explanation he put forward is the existence of a common genetic factor causing both smoking and lung cancer through independent pathways.

Fisher's alternative explanations are, if not plausible, difficult to refute. At the time, the relative risk of smoking versus non-smoking on lung cancer was estimated to be around 9. Cornfield et al. (1959)'s insight is that an implication of Fisher's hypothesis – that a common genetic factor could explain this association – is that the relative prevalence of this common factor must be at least 9 times greater among smokers than non-smokers. A relative prevalence of that magnitude for a genetic factor is highly implausible.

This influential argument helped to dismantle opposition to the hypothesized carcinogenicity of cigarettes. It is worth noting that this sensitivity analysis was augmented by a compelling body of evidence from other sources, including animal studies, ecological data, and studies of pathogenesis (Cornfield et al., 1959). This *triangulation* of evidence (Lawlor, Tilling, and Smith, 2016) is a pillar of robust scientific inquiry.

Sensitivity analyses have become more sophisticated and diverse in the decades since, but interrogating assumptions in a limited and quantifiable way remains an important scientific undertaking. In this thesis, I develop sensitivity analyses for three distinct biases: selection bias, coarsening bias in instrumental variable designs, and ancestral and familial biases in genetic association studies.

1.2 Thesis outline

In Chapter 3, I propose and validate an approach to statistical inference in a class of stochastic optimization problems characterized by the optimal value of a function over a set where both the function and set need to be estimated by empirical data. Despite some progress for convex problems, statistical inference in this general setting remains to be developed. To address this, I derive an asymptotically valid confidence interval for the optimal value through an appropriate relaxation of the estimated set. I then apply this general result to the problem of selection bias in population-based cohort studies. I show that existing sensitivity analyses, which are often conservative and difficult to implement, can be formulated in my framework and made significantly more informative via auxiliary information on the population. I conduct a simulation study to evaluate the finite sample performance of my inference procedure and conclude with two substantive motivating examples: the causal effect of education on income and the evaluation of risk factors for COVID-19, both in the highly-selected UK Biobank cohort. I demonstrate that my method can produce informative bounds using plausible population-level auxiliary constraints. I implement this method in the R package `selectioninterval`.

Chapter 4 is concerned with the assumption in Mendelian randomization studies that the

relationship between the genetic instruments and the outcome is fully mediated by the exposure, known as the exclusion restriction assumption. In epidemiological studies, the exposure is often a coarsened approximation to some latent continuous trait. For example, latent liability to schizophrenia can be thought of as underlying the binary diagnosis measure. Genetically-driven variation in the outcome can exist within categories of the exposure measurement, thus violating this assumption. I propose a framework to clarify this violation, deriving a simple expression for the resulting bias and showing that it may inflate or deflate effect estimates but will not reverse their sign. I then characterize a set of assumptions and a straightforward method for estimating the effect of standard deviation increases in the latent exposure. My method relies on a sensitivity parameter which can be interpreted as the genetic variance of the latent exposure. I show that this method can be applied in both the one-sample and two-sample settings. I conclude by demonstrating my method in an applied example and re-analyzing two papers which are likely to suffer from this type of bias, allowing meaningful interpretation of their effect sizes.

Chapter 5 is motivated by the incongruence between how Mendelian randomization (MR) is typically justified – as an observational design based on the random transmission of genes from parents to offspring – and how it is typically performed. In practice, this inferential basis is typically only implicit or used as an informal justification. As parent-offspring data becomes more widely available, I advocate a different approach to MR that is exactly based on this randomization, making explicit the common analogy between MR and a randomized controlled trial. I begin by developing a causal graphical framework for MR which formalizes several biological processes and phenomena, including population structure, gamete formation, fertilization, genetic linkage, and pleiotropy. This causal graph is then used to detect biases in the MR design and identify sufficient confounder adjustment sets to correct them. I then propose a randomization test for causal hypotheses in the MR design by using precisely the exogenous randomness in meiosis and fertilization. I term this "almost exact MR", because exactness of the inference depends on precisely knowing the distribution of offspring haplotypes resulting from meioses in one or both parents, which is widely studied in genetics. I demonstrate via simulation that propensity scores obtained from the underlying meiosis model can form powerful test statistics. Besides transparency and conceptual appeals, my approach also offers some practical advantages, including lack of commitment to a particular phenotype model, robustness to weak instruments, and eliminating bias that may arise from population structure, assortative mating, dynastic effects and 'pleiotropy by linkage' that is more prevalent in admixed samples. I conclude with a negative and positive control analysis in the Avon Longitudinal Study of Parents and Children using my R package `almostexactmr`.

Chapter 2

Background

2.1 Causal inference

2.1.1 Overview

Causal inference is an expansive field housing a number of intellectual traditions and languages, often developed independent of one another and serving different purposes. The two most influential languages are potential outcomes and causal graphical models. Throughout this thesis, I utilize both languages to communicate my findings and unify them when they express complementary ideas. This section provides an overview of potential outcomes and causal graphical models, focusing on aspects relevant to this thesis, then brings them together via the recently-developed single world intervention graphs. It concludes with a more detailed exposition of randomization inference, inverse probability weighting and instrumental variables, which are necessary for several of the later chapters.

2.1.2 Potential outcomes

The potential outcomes framework views causal inference fundamentally as a contrast of hypothetical outcomes under distinct interventions. It originated with the work of Neyman in the 1920s in the context of agricultural experiments (reprinted in Neyman (1990)), later popularized by Rubin (1974) and subsequent work. It is now a bedrock of causal reasoning in the social sciences (Imbens and Rubin, 2015) and expanding into more disparate fields such as epidemiology and population genetics (Hernán and Robins, 2020; Bates, Sesia, Sabatti, and Candes, 2020). This section draws heavily upon the reference texts Imbens and Rubin (2015) and Hernán and Robins (2020).

We start with a set of N individuals indexed by $i = 1, \ldots, N$. Each individual experiences a binary *exposure* $D_i \in \{0, 1\}$, where the exposure could be an experimental intervention or naturally-occurring phenomenon. I adhere to the common convention that $D_i = 1$ indicates an

experimental or active exposure and $D_i = 0$ indicates a placebo or control exposure. We are interested in the effect of this exposure on an *outcome* Y_i.

Potential outcomes are characterized by the realizations of Y_i under different exposures. Each individual has a set of potential outcomes $\{Y_i(d) \colon d = 0, 1\}$ that would occur if their exposure level was exogenously set to $D_i = d$. Individual i experiences a *causal effect* of their exposure on their outcome if $Y_i(1) \neq Y_i(0)$. This causal effect can be summarized by the scalar $\beta_i = Y_i(1) - Y_i(0)$, or some other contrast, such as $Y_i(1)/Y_i(0)$. Crucially, in the potential outcomes framework, a causal effect is a function of two or more potential outcomes.

My notation so far is not assumption-free. I have made an implicit *no interference* assumption which posits that the potential outcomes of each individual are unaffected by the exposures of other individuals (Rubin, 1980; Imbens and Rubin, 2015).

Assumption 2.1. (No interference) $Y_i(d) \perp\!\!\!\perp D_j$ for all $i \neq j$ and $d \in \{0, 1\}$.

I have also assumed *no hidden versions of the same treatment*. The exposure $D_i = 1$ could correspond to treatment with an experimental drug but, within that exposure category, different individuals could receive different doses. If the outcome Y_i is survival, it is possible that an individual would survive ($Y_i = 1$) if administered a high dose but die ($Y_i = 0$) if administered a low dose. The potential outcome $Y_i(1)$ is not well-defined since it could take two distinct values depending on the hidden sub-exposure. The previous two assumptions are sometimes jointly referred to as the *stable unit treatment value assumption* (Rubin, 1980).

The "fundamental problem of causal inference" (Holland, 1986) is that we cannot observe both potential outcomes of the same individual simultaneously, such that β_i cannot be identified from the observed data. We may only observe the potential outcome corresponding to the realized exposure. This assumed relationship between the observed data and potential data is called the *consistency* assumption (Hernán and Robins, 2020).

Assumption 2.2. (Consistency) $Y_i = Y_i(D_i) = D_i Y_i(1) + (1 - D_i) Y_i(0)$.

Table 2.1: Example data for the effect of aspirin (D_i) on self-reported headache severity (Y_i).

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$	C_i
1	1	5	5	?	8
2	1	7	7	?	7
3	1	3	?	3	5
4	0	5	5	?	6
5	0	9	?	9	9
6	0	7	?	7	7

From the perspective of Holland (1986), causal inference can be regarded as a missing data problem. Consider the simple hypothetical study in Table 2.1 consisting of $N = 6$ individuals

with headaches, 3 of whom take aspirin and 3 of whom do not. The outcome variable is self-reported headache severity on a 10-point scale (1 is low, 10 is high) an hour after ingestion. The potential outcome under the counterfactual exposure is unobserved for each individual. I also assume that there is a *confounder* C_i representing baseline headache severity. This confounder influences each individual's decision to take aspirin and also influences their headache severity an hour later.

There are several approaches for drawing causal inferences despite this fundamental missing data problem. One approach is to test hypotheses about the non-existence of individual-level causal effects, that is, $Y_i(1) = Y_i(0)$ for all $i = 1, \ldots, N$. I return to this approach in Section 2.1.5. The more common approach is to make inferences about *average causal effects*. An average causal effect is said to exist if $E\{Y_i(1)\} \neq E\{Y_i(0)\}$, where $E(\cdot)$ is typically defined over a hypothetical infinite *super-population* from which each individual i is randomly sampled. For a discussion on relaxing the need for this super-population assumption, see Ding, Li, and Miratrix (2017). The effect measure can be viewed as an average over the individual-level effects $\beta = E(\beta_i)$.

Identifying β from the observed data $\{(D_i, Y_i, C_i) \colon i = 1, \ldots, N\}$ requires an assumption on the distribution of the potential outcomes. If we assume that the confounder C_i is the only common cause of the exposure and outcome, then individuals who took aspirin and individuals who did not take aspirin are *exchangeable* with respect to the potential outcomes conditional on C_i.

Assumption 2.3. (Exchangeability) $D_i \perp\!\!\!\perp Y_i(d) \mid C_i$ for all $d \in \{0, 1\}$.

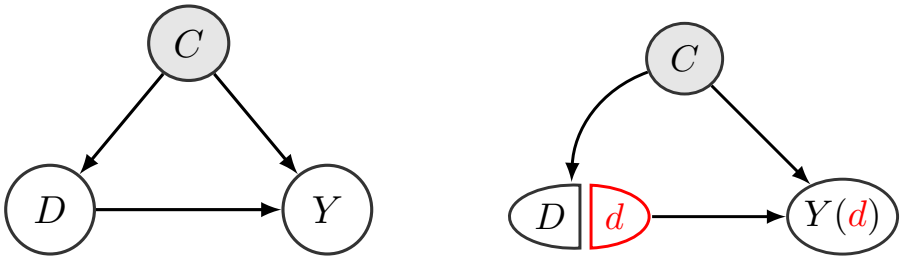
Colloquially, exchangeability posits that there are no systematic differences between individuals who took and did not take aspirin once we account for their baseline headache severity.

If C_i was observed, we could identify the expected potential outcome $E\{Y_i(d)\}$ by *standardizing* over C_i (Hernán and Robins, 2020), such that

$$
\begin{aligned}
& E\{E(Y_i \mid D_i = d, C_i)\} \\
&= E[E\{Y_i(d) \mid D_i = d, C_i\}] && \text{(Assumption 2.2)} \\
&= E[E\{Y_i(d) \mid C_i\}] && \text{(Assumption 2.3)} \\
&= E\{Y_i(d)\}
\end{aligned}
$$

where the last equality follows by the law of iterated expectations. This derivation relies on an implicit assumption that the expected value $E(Y_i \mid D_i = d, C_i)$ is always well-defined. It is straightforward to show that this is satisfied when the probability of taking (or not taking) aspirin is bounded away from zero and one within each strata of self-reported headache severity, known as *positivity* (Hernán and Robins, 2020, p. 24).

Assumption 2.4. (Positivity) $0 < \text{pr}(D_i = d \mid C_i) < 1$ for all $d \in \{0, 1\}$.



(a) Directed acyclic graph (b) Single world intervention graph

Figure 2.1: Causal diagram of the example in Section 2.1.2

Putting everything together, the average causal effect can be obtained from the observed data via $\beta = E\{E(Y_i \mid D_i = 1, C_i)\} - E\{E(Y_i \mid D_i = 0, C_i)\}$. Realistically, we observe a finite sample of individuals rather than a population distribution. I sidestep the discussion of estimation and statistical inference for now, but return to it in Section 2.1.5.

2.1.3 Causal graphical models

The causal inference problem in Section 2.1.2 can be recast using a causal graphical model. A graphical model summarizes the conditional independence relationships among a set of random variables. A causal graphical model introduces direction to the relationships, allowing us to view the directed edges as causal pathways among variables. This section draws heavily on Pearl (2000) and Pearl (2009). I begin by providing some general terminology for graphs.

Definition 2.1. A *directed graph* $\mathcal{G} = (V, E)$ is characterized by a set of *vertices* or *nodes* V and *edges* E. The set of edges is defined by $E \subseteq \{(i, j) \mid (i, j) \in V^2, v \neq v'\}$, where (i, j) is an ordered pair.

Definition 2.2. There is a *directed path* between vertices i and j if there exists a sequence of vertices $k_0 = i, k_1, k_2, \ldots, k_m = j$ such that $(k_{l-1}, k_l) \in E$ for all $l = 1, 2, \ldots, m$.

Definition 2.3. If $(i, j) \in E$ then we say that i is a *parent* of j and j is a *child* of i. The set of parents of a vertex i is denoted $pa_{\mathcal{G}}(i)$. Moreover, a vertex i is an *ancestor* of j and j is a *descendant* of i if there exists a directed path from i to j. The set of descendants of a vertex i is denoted $an_{\mathcal{G}}(i)$.

So far, the graph I have characterized is a deterministic set of vertices with edges connecting them. We say that the joint distribution of a set of random variables $X = (X_1, X_2, \ldots, X_p)$ *factorizes* with respect to a graph $\mathcal{G}$ with $V = (1, 2, \ldots, p)$ if each vertex is mapped to a random variable $i \to X_i$ and

$$X_i \perp\!\!\!\perp X_j \mid X_{pa_{\mathcal{G}}(i)}$$

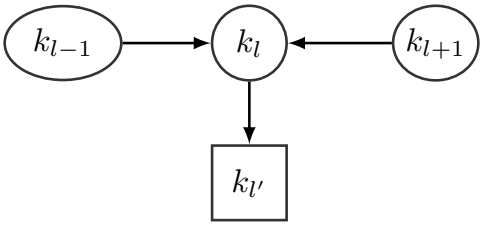


Figure 2.2: Conditioning on the descendant of a collider (Definition 2.4[4])

for all $i \in V$ and $j \in an_{\mathcal{G}}(i) \setminus pa_{\mathcal{G}}(i)$. This is sometimes called the *local Markov property*. For ease of notation, I will often omit the mapping $i \to X_i$ and simply work with the induced graph $\mathcal{G} = (V, E)$ where $V = (X_1, X_2, \ldots, X_p)$ and $E \subseteq \{(X_i, X_j) \colon (X_i, X_j) \in V^2,\, i \neq j\}$.

Figure 2.1a depicts the example in the previous section as a directed graph, such that $V = \{D, Y, C\}$ and $E = \{(D, Y), (C, D), (C, Y)\}$. A key feature of a causal graphical model is the absence of cycles among the vertices, which is to ensure temporal consistency, that is, recognizing that the future cannot cause the past. This type of graph is called a *directed acyclic graph* (DAG).

With the structure of the DAG established, we can now reason about implied conditional independencies. I use the following four rules to determine whether a path is blocked, implying independence, or not (Pearl, 1988).

Definition 2.4. A path $k_0 = i, k_1, \ldots, k_m = j$ is either *d-separated* (blocked) or *d-connected* (unblocked) according to the following criteria.

[1] A path is d-separated if it contains a *collider* $k_{l-1} \to k_l \leftarrow k_{l+1}$ for some $l = 1, \ldots, m$. For example, $D \to Y \leftarrow C$.

[2] A path is d-separated if it contains a non-collider that has been conditioned on, denoted by $k_{l-1} \leftarrow \boxed{k_l} \to k_{l+1}$ for some $l = 1, \ldots, m$. For example, $D \leftarrow \boxed{C} \to Y$.

[3] A path is d-connected if it contains a collider that has been conditioned on, $k_{l-1} \to \boxed{k_l} \leftarrow k_{l+1}$ for some $l = 1, \ldots, m$, and criteria [1] and [2] are not satisfied. For example, $D \to \boxed{Y} \leftarrow C$.

[4] A path is d-connected if it contains a descendant of a collider that has been conditioned on (see Figure 2.2) and criteria [1] and [2] are not satisfied.

We can now characterize causal effects in terms of counterfactuals on the DAG. A counterfactual of a random variable is induced by intervening on one or more of its ancestors. Counterfactuals can be formalized if we view the random variables as functions of their ancestors and an arbitrary disturbance term.

Definition 2.5. Given a DAG $\mathcal{G}$, the random variables $X = (X_1, X_2, \ldots, X_p)$ satisfy a *non-parametric structural equation model* (NPSEM) if, for all $i = 1, \ldots, p$,

$$X_i = f_i(X_{pa_{\mathcal{G}}(i)}, \epsilon_i)$$

for some function $f_i(\cdot)$ and random variable ϵ_i.

Example 2.1. A common example of the function $f_i(\cdot)$ in Definition 2.5 is the *linear structural equation model*. This model posits that, for all $i = 1, \ldots, p$,

$$X_i = \sum_{j \in pa_{\mathcal{G}}(i)} \beta_j X_j + \epsilon_i.$$

In a linear structural equation model, causal effects are characterized by the parameter β_j.

From the perspective of Definition 2.5, an intervention setting $X_{pa_{\mathcal{G}}(i)} = x_{pa_{\mathcal{G}}(i)}$ would give rise to the counterfactual

$$X_i(X_{pa_{\mathcal{G}}(i)} = x_{pa_{\mathcal{G}}(i)}) = f_i(x_{pa_{\mathcal{G}}(i)}, \epsilon_i).$$

This has obvious parallels with the potential outcomes of the previous section. For example, returning to Figure 2.1a, the counterfactual of Y following an intervention setting $D = 1$ would be $Y(D = 1) = f_Y(1, \epsilon_Y)$, where $f_Y(\cdot)$ and ϵ_Y are respectively the counterfactual function and disturbance term.

The effect of interventions on more distant ancestors can be defined recursively. Suppose $J \subseteq an_{\mathcal{G}}(i) \setminus pa_{\mathcal{G}}(i)$ indexes a set of interventions on ancestors of i, then the counterfactual can be defined as

$$X_i(x_J) = f_i\{X_{pa_{\mathcal{G}}(i)}(x_J), \epsilon_i\}.$$

2.1.4 Single world intervention graphs

The potential outcome and causal graph paradigms were developed independent of one another. While potential outcomes trace their origins to Neyman's agricultural experiments (Neyman, 1990), causal graphs emerged around the same time with Wright's path diagrams (Wright, 1920), later formalized by Pearl (1988). Despite their disparate origins, there have been recent attempts to unify the two frameworks and combine their unique strengths (Richardson and Robins, 2013a). While potential outcomes provide a clear definition of causation, graphs provide a coherent way of reasoning about networks of causal relationships.

This unifying framework is called *single world intervention graphs* (SWIGs). The SWIG representation of the motivating example in Section 2.1.2 is given in Figure 2.1b. Compared to Figure 2.1a, the SWIG splits the exposure node into two halves: D, representing the stochastic naturally-occurring exposure, and d, representing a fixed intervention value. The random half D inherits all incoming arrows in the original DAG and the fixed half d inherits all outgoing

arrows. Descendants of the intervention node (in this case Y) are replaced with the potential outcomes $Y(d)$ under the intervention value d. SWIGs are referred to as "single world" because they represent the causal graph under a particular realization of the exposure.

It has been shown that SWIGs define a graphical model for the potential outcomes (Richardson and Robins, 2013b), so we can apply d-separation to reason about conditional independence among counterfactuals. For example, Figure 2.1b implies that D and $Y(d)$ are d-separated conditional on C, which is equivalent to our definition of exchangeability in Assumption 2.3.

2.1.5 Randomization inference

The previous three sections have focused on defining causal quantities and outlining the conditions under which they can be obtained from observed data, commonly referred to as *identification* (Lewbel, 2019). This section is concerned with *inference*, which is the process of quantifying uncertainty in our causal statements. In particular, I focus on an approach to inference called *randomization inference* (Rosenbaum, 2002a; Zhang and Zhao, 2022). Randomization inference is applicable to problems where the exposure is randomized via a mechanism that is known to the analyst. In Chapter 5, I describe how this can be applied to the randomization that occurs during genetic inheritance.

Returning to the aspirin example from Section 2.1.2, suppose this data emerges from a randomized experiment where each individual i is assigned to take aspirin ($D_i = 1$) or a placebo ($D_i = 0$). The resulting vector of exposures is summarized by $\boldsymbol{D} = (D_1, D_2, \ldots, D_N)^T$. I will assume that the experiment is completely randomized, such that a fixed number of individuals N_t are assigned to take aspirin and $N_c = N - N_t$ are assigned to take a placebo. In the aspirin example, $N_t = N_c = 3$. Let $\Omega = \{(d_1, \ldots, d_N) \in \{0,1\}^N \colon \sum_{i=1}^N d_i = N_t\}$ denote the set of feasible assignment vectors. By assumption, all assignment vectors in Ω are realized with equal probability. Stated formally, the randomization distribution can be written as

$$\text{pr}(\boldsymbol{D} = \boldsymbol{d} \mid \mathcal{F}) = \begin{cases} \binom{N}{N_t}^{-1}, & \text{for all } \boldsymbol{d} \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

where $\mathcal{F} = \{(Y_i(1), Y_i(0)) \colon i = 1, \ldots . N\}$ denotes the collection of potential outcomes for all individuals in the sample, which we hold fixed.

Randomization inference is concerned with exact hypotheses of the form

$$H_0 : Y_i(d) = Y_i(0) + \beta_0, \text{ for all } d \in \{0,1\} \text{ and } i = 1, \ldots, N,$$

where $\beta_0 = 0$ corresponds to the *sharp null hypothesis* of no effect for any individual in the sample (Imbens and Rubin, 2015). This implies a constant additive treatment effect β_0 across individuals. Under this hypothesis and Assumption 2.2 (consistency), the baseline potential

outcome can be written in terms of the observable data $\{(D_i, Y_i) \colon i = 1, \ldots, N\}$ as

$$Y_i(0) = Y_i - \beta_0 D_i = \begin{cases} Y_i, & \text{if } D_i = 0, \\ Y_i - \beta_0, & \text{if } D_i = 1, \end{cases}$$

A key characteristic of exact hypotheses is that the potential outcomes are imputable under alternate treatment assignments. For example, individual 1 in Table 2.1 was randomized to take aspirin ($D_1 = 1$) and reported a headache severity of $Y_1 = Y_1(1) = 6$. If $\beta_0 = 0$ were true, then individual 1's headache severity would be identical regardless of their treatment assignment, implying that $Y_1(0) = 6$ as well. Table 2.2 extends this argument to all individuals in the sample.

Table 2.2: Imputed data for the effect of aspirin (D_i) on self-reported headache severity (Y_i) under the sharp null hypothesis.

i	D_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	1	5	5	5
2	1	7	7	7
3	1	3	3	3
4	0	5	5	5
5	0	9	9	9
6	0	7	7	7

When the null hypothesis is true, the complete randomization of D_i via Equation (2.1) implies unconditional exchangeability

$$D_i \perp\!\!\!\perp Y_i(0) \overset{H_0}{=} Y_i - \beta_0 D_i.$$

Consequently, testing the null hypothesis H_0 that the causal effect is a constant β_0 is equivalent to testing the independence of D_i and $Y_i - \beta_0 D_i$. To this end, a simple test statistic is the difference in outcomes between the two groups,

$$T(\boldsymbol{D} \mid \mathcal{F}) = \sum_{i=1}^{N} D_i(Y_i - \beta_0 D_i) - \sum_{i=1}^{N}(1 - D_i)(Y_i - \beta_0 D_i) \overset{H_0}{=} \sum_{i:D_i=1} Y_i(0) - \sum_{i:D_i=0} Y_i(0).$$

The test statistic for Table 2.2 is $T(\boldsymbol{D} \mid \mathcal{F}) = (6 + 7 + 3)/3 - (5 + 9 + 7)/3 = -2$. This tells us that individuals who were assigned to take aspirin reported a headache severity 2 points lower, on average, than individuals who were assigned to take placebo. However, even if aspirin had no effect whatsoever on headache severity ($\beta_0 = 0$), it is possible that individuals with less severe headaches happened to be randomized to take aspirin. This uncertainty can be quantified via the p-value

$$P(\boldsymbol{D} \mid \mathcal{F}) = \tilde{\text{pr}}\{T(\tilde{\boldsymbol{D}} \mid \mathcal{F}) \leq T(\boldsymbol{D} \mid \mathcal{F})\}. \tag{2.2}$$

Here $\tilde{\boldsymbol{D}}$ is an independent copy of $\boldsymbol{D}$ and p̃r means that the probability is taken over $\tilde{\boldsymbol{D}}$ according to the randomization distribution (2.1). Table 2.3 gives an example of $\tilde{\boldsymbol{D}}$. Since the potential outcomes are imputable under H_0, we know that $T(\tilde{\boldsymbol{D}} \mid \mathcal{F}) = 2$ in this instance. In plain terms, we are asking: if we re-ran the experiment many times under the null hypothesis (i.e., D_i and $Y_i - \beta_0 D_i$ are independent), how often would we observe a test statistic more extreme than our observed test statistic? If this probability is lower than some level α, then we have little confidence in the null hypothesis.

Table 2.3: Imputed data for the effect of aspirin (D_i) on self-reported headache severity (Y_i) under the sharp null hypothesis.

i	$\tilde{D}_i$	Y_i	$Y_i(1)$	$Y_i(0)$
1	0	5	5	5
2	1	7	7	7
3	0	3	3	3
4	1	5	5	5
5	1	9	9	9
6	0	7	7	7

We can characterize the size of this p-value in the following proposition.

Proposition 2.1. *The p-value* (2.2) *has size* α *in the sense that*

$$\text{pr}\{P(\boldsymbol{D} \mid \mathcal{F}) \leq \alpha \mid H_0\} \leq \alpha.$$

for any significance level $0 \leq \alpha \leq 1$ *and test statistic* $T(\cdot \mid \mathcal{F})$ *such that* $T(\tilde{\boldsymbol{Z}} \mid \mathcal{F}) \overset{d}{=} T(\boldsymbol{Z} \mid \mathcal{F})$ *under* H_0.

A sketch of the proof is as follows. Since $T(\tilde{\boldsymbol{Z}} \mid \mathcal{F}) \overset{d}{=} T(\boldsymbol{Z} \mid \mathcal{F})$ under H_0 and $P(\boldsymbol{D} \mid \mathcal{F}) = \tilde{\text{pr}}\{T(\tilde{\boldsymbol{D}} \mid \mathcal{F}) \leq T(\boldsymbol{D} \mid \mathcal{F})\}$, we can view $P(\boldsymbol{D} \mid \mathcal{F})$ as the distribution function of $T(\boldsymbol{Z} \mid \mathcal{F})$. Since distribution functions stochastically dominate the uniform distribution, the result follows. See Zhang and Zhao (2022) for further discussion.

2.1.6 Inverse probability weighting

Inverse probability weighting (IPW) is used when sub-groups of a sample are over- or under-represented relative to some desired target population (Hernán and Robins, 2020, p.20–24). IPW is a primary focus of Chapter 3 and also makes a brief appearance in Chapter 5 in the context of propensity score weighting. Suppose $S_i \in \{0, 1\}$ is a random variable denoting whether an individual is included in a sample ($S = 1$) or not ($S = 0$). The probability of selection is governed by a collection of random variables $X = (X_1, \ldots, X_k)$ such that $\text{pr}(S = 1 \mid X) = e(X)$ for some function $e(\cdot)$. One of the simplest statistics for which inverse probability weighting is

applicable is the population mean $\mu = E(Y)$ of a random variable Y. Horvitz and Thompson (1952) proposed one of the earliest IPW estimators for the population mean, given by

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} S_i \, Y_i \, e(X_i)^{-1}.$$

for an iid sample $i = 1, \ldots, N$.

Under the assumption that $Y \perp\!\!\!\perp S \mid X$ and $0 < e(X) \leq 1$ with probability one (Stuart et al., 2011), we can show that $\hat{\mu}$ is unbiased for μ by

$$\begin{aligned}
E(\hat{\mu}) &= E\left\{ \frac{1}{N} \sum_{i=1}^{N} S_i \, Y_i \, e(X_i)^{-1} \right\} \\
&= \frac{1}{N} \sum_{i=1}^{N} E[E\{S_i \, Y_i \, e(X_i)^{-1} \mid X_i\}] \text{ (law of iterated expectations)} \\
&= \frac{1}{N} \sum_{i=1}^{N} E[e(X_i)^{-1} \, E\{S_i \, Y_i \mid X_i\}] \\
&= \frac{1}{N} \sum_{i=1}^{N} E[e(X_i)^{-1} \, E\{S_i \mid X_i\} \, E\{Y_i \mid X_i\}] \text{ (by assumption)} \\
&= \frac{1}{N} \sum_{i=1}^{N} E\{E(Y_i \mid X_i)\} \\
&= E(Y_i) = \mu.
\end{aligned}$$

2.1.7 Propensity score weighting

In this section, I introduce a quantity called the *propensity score*. If X is a set of observed baseline variables, including all observed confounders, and $D \in \{0, 1\}$ is a binary exposure, then the propensity score can be written as

$$e(X) = \text{pr}(D = 1 \mid X), \tag{2.3}$$

that is, the probability of being exposed given a vector X. This bears obvious similarities with the probability of selection in Section 2.1.6, except that the selection indicator S is replaced by the exposure indicator D. Inverse probability weighting is also a valid approach for recovering expected potential outcomes and thereby average causal effects. An estimator for the average causal effect can be written as

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} D_i \, Y_i \, e(X_i)^{-1} - \frac{1}{N} \sum_{i=1}^{N} (1 - D_i) \, Y_i \, \{1 - e(X_i)\}^{-1}.$$

Under positivity ($0 < e(X) < 1$ with probability one) and exchangeability, the expected value of the first term in this estimator is

$$\begin{aligned}
E\left[\frac{1}{N}\sum_{i=1}^{N} D_i\, Y_i\, e(X_i)^{-1}]\right] &= \frac{1}{N_1}\sum_{i=1}^{N} E\{D_i\, Y_i\, e(X_i)^{-1}\} \\
&= \frac{1}{N}\sum_{i=1}^{N} E[E\{D_i\, Y_i\, e(X_i)^{-1}\} \mid X_i] \\
&= \frac{1}{N}\sum_{i=1}^{N} E\{e(X_i)^{-1} E(D_i Y_i \mid X_i)\} \\
&= \frac{1}{N}\sum_{i=1}^{N} E\{E(Y_i \mid D_i = 1, X_i)\} \\
&= \frac{1}{N}\sum_{i=1}^{N} E\{E(Y_i(1) \mid X_i)\} \\
&= E\{Y_i(1)\}.
\end{aligned}$$

Similarly, the second term is equal to $E\{Y_i(0)\}$ in expectation. This means that $E(\hat{\beta}) = E\{Y_i(1) - Y_i(0)\}$, the average causal effect.

Propensity score weighting has a connection to sensitivity analyses through the work of Rosenbaum and colleagues (Rosenbaum, 1987; Gastwirth, Krieger, and Rosenbaum, 1998; Rosenbaum, 2002b). The influential sensitivity analysis of Cornfield et al. (1959) introduced in Section 1.1 relies on a binary outcome, a single binary confounder and no adjustment for observed covariates. By placing restrictions on the maximal deviation between an estimated propensity score and the true propensity score containing unobserved confounders, Rosenbaum-type sensitivity analyses provide a flexible and interpretable way of assessing sensitivity to unobserved confounding.

2.1.8 Instrumental variables

It often occurs that some or all of the confounders C are unobserved, so that conditional exchangeability (see Assumption 2.3) is not satisfied. In this section, I describe an approach that can recover causal effects even in the presence of unobserved confounding.

The principles behind instrumental variable analysis emerged independently in several fields. In economics, an appendix of Wright (1928) derives instrumental variable estimators for simultaneously-determined supply and demand curves. Philip Wright's eldest son, Sewall, is suspected to have contributed to this appendix (Stock and Trebbi, 2003). Sewall is relevant to my thesis due to his method of path coefficients, a precursor to causal graphical models. In medicine, Zelen (1977) proposes a design for clinical trials, called an encouragement design, which randomize encouragement to a treatment rather than the treatment itself. As we will see, this has direct parallels with instrumental variable designs more generally. See Stock and

Trebbi (2003), Imbens (2014) and Imbens and Rubin (2015) for a more detailed survey of the history and development of instrumental variable analysis.

Instrumental variable analysis is characterized by the existence of an *instrument* or *instrumental variable* Z. An instrumental variable induces unconfounded variation in the exposure without otherwise affecting the outcome. We can then make inferences about the effect of the exposure through the variation induced by the instrument. Suppose the potential outcomes are given by $Y(z,d)$ and $D(z)$ for $z,d \in \{0,1\}$. Paraphrasing Condition 1 of Imbens and Angrist (1994), an instrument Z is valid for the causal effect of D on Y if the following three assumptions hold (omitting possible conditioning on observed confounders).

Assumption 2.5. (Random assignment) $Z \perp\!\!\!\perp \{D(z), Y(z,d)\}$ for all $z,d \in \{0,1\}$.

Assumption 2.6. (Exclusion restriction) $Y(z,d) = Y(z',d) = Y(d)$ for all $d,z,z' \in \{0,1\}$.

Assumption 2.7. (Relevance) $E(D \mid Z = z)$ is a non-trivial function of z.

Suppose for the moment that the outcome is binary $Y \in \{0,1\}$. Under Assumptions 2.5–2.7, Balke and Pearl (1997) derive sharp bounds for the average causal effect $\beta = \text{pr}\{Y(1) = 1\} - \text{pr}\{Y(0) = 1\}$. Consistent with their notation, let $p_{yd,z} = \text{pr}(Y = y, D = d \mid Z = z)$, then

$$\max \left\{ \begin{array}{c} p_{00,0} + p_{11,1} - 1 \\ p_{00,1} + p_{11,1} - 1 \\ p_{11,0} + p_{00,1} - 1 \\ p_{00,0} + p_{11,0} - 1 \\ 2p_{00,0} + p_{11,0} + p_{10,0} + p_{11,1} - 2 \\ p_{00,0} + 2p_{11,0} + p_{00,1} + p_{01,1} - 2 \\ p_{10,0} + p_{11,0} + 2p_{00,1} + p_{11,1} - 2 \\ p_{00,0} + p_{01,0} + p_{00,1} + 2p_{11,1} - 2 \end{array} \right\} \leq \beta$$

and

$$\min \left\{ \begin{array}{c} 1 - p_{10,0} - p_{01,1} \\ 1 - p_{01,0} - p_{10,1} \\ 1 - p_{01,0} - p_{10,0} \\ 1 - p_{01,1} - p_{10,1} \\ 2 - 2p_{01,1} - p_{10,0} - p_{10,1} - p_{11,1} \\ 2 - p_{10,0} - 2p_{10,0} - p_{00,1} - p_{01,1} \\ 2 - p_{10,0} - p_{11,0} - 2p_{01,1} - p_{10,1} \\ 2 - p_{00,0} - p_{01,0} - p_{01,1} - 2p_{10,1} \end{array} \right\} \geq \beta$$

The authors use the *do-calculus* notation (Pearl, 1995; Pearl, 2000), which I substitute for potential outcomes. As noted in Imbens (2020), for simple problems such as this, the distinction between do-calculus and potential outcomes is merely notational. There is another literature,

not explored in this thesis, on using related inequalities to test for instrument validity (Kitagawa, 2015; Mourifié and Wan, 2017).

The bounds on the average causal effect are often conservative in practice, limiting the utility of the inference (Swanson et al., 2015). If we are willing to place additional assumptions on the tuple $\{Z, D(0), D(1), Y(0), Y(1)\}$, it is possible to obtain more informative inference, although not necessarily for the average causal effect. Condition 2 of Imbens and Angrist (1994), which I rewrite below, is an additional assumption that allows point identification of the average causal effect within a subset of individuals.

Assumption 2.8. (Monotonicity) $D(1) \geq D(0)$ with probability one.

$$
\begin{aligned}
&E(Y \mid Z=1) - E(Y \mid Z=0) \\
&= E[D(1)Y(1) + \{1 - D(1)\}Y(0) \mid Z=1] - E[D(0)Y(1) + \{1 - D(0)\}Y(0) \mid Z=0] \\
&= E[D(1)Y(1) + \{1 - D(1)\}Y(0)] - E[D(0)Y(1) + \{1 - D(0)\}Y(0)] \\
&= E[\{D(1) - D(0)\}\{Y(1) - Y(0)\}] \\
&= \text{pr}\{D(1) - D(0) = 1\} E\{Y(1) - Y(0) \mid D(1) - D(0) = 1\}.
\end{aligned}
$$

The first line follows from consistency (Assumption 2.2), the second line follows from random assignment (Assumption 2.5) and the fourth line follows from monotonicity (Assumption 2.8). Since $E(D \mid Z=1) - E(D \mid Z=0) = \text{pr}\{D(1) - D(0) = 1\}$ by the same assumptions, it follows that

$$
\frac{E(Y \mid Z=1) - E(Y \mid Z=0)}{E(D \mid Z=1) - E(D \mid Z=0)} = E\{Y(1) - Y(0) \mid D(1) - D(0) = 1\}.
$$

The right-hand expression is called the *complier average causal effect* (CACE) or *local average treatment effect*. The subset of the population whose exposure level is shifted by the instrument ($D(1) - D(0) = 1$) is called the *complier* group. This is in contrast to *always-takers* ($D(1) = D(0) = 1$), *never-takers* ($D(1) = D(0) = 0$), and *defiers* ($D(1) - D(0) = -1$).

The CACE is not always of interest, for example, when compliers are an idiosyncratic subset of the population. Alongisde bounding approaches like Balke and Pearl (1997), it is possible to impose additional assumptions that allow the CACE to be interpreted as an average causal effect. Hernán and Robins (2006) describe so-called *homogeneity* assumptions which restrict the variability of individual-level effects. The most common homogeneity assumption is linearity, such that

$$
Y(D) = \beta D + Y(0) = \beta D + \epsilon,
$$

where $\epsilon = Y(0)$ is the error term. Under this model, $Y(1) - Y(0) = \beta$ for all individuals, so the CACE equals the average causal effect. Subsequently, weaker assumptions have been proposed which assume no additive interactions between confounders and either the instrument or the exposure (Wang and Tchetgen Tchetgen, 2018; Hartwig, Wang, et al., 2022).

Up until now, I have been working with a binary exposure, however, exposures are often multi-valued or continuous. This can introduce additional challenges, since the exclusion restriction can be violated if the exposure measure is coarser than the exposure itself. This is the motivation behind Chapter 4. This problem was first introduced in Section 3.1 of Angrist and Imbens (1995) in the context of incorrectly coded binary exposures (the paper refers to exposures as treatments, which is the convention in economics, but I retain the language I have been using). The authors consider a setting with a multi-valued exposure given by $S \in \{0, 1, \ldots, J\}$ and a binary instrument $Z \in \{0, 1\}$. The potential outcomes are given by $Y(j)$ for all $j = 1, \ldots, J$ and $S(z)$ for all $z = 0, 1$. In their Theorem 1, the authors show that, under monotonicity ($S(1) - S(0) \geq 0$ with probability one), the Wald estimator takes the form

$$\beta \equiv \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(S \mid Z = 1) - E(S \mid Z = 0)} = \sum_{j=1}^{J} \omega_j \, \beta_j,$$

where

$$\omega_j = \frac{\text{pr}\{S(1) \geq j > S(0)\}}{\sum_{i=1}^{J} \text{pr}(S(1) \geq i > S(0)} \text{ and } \beta_j = E\{Y(j) - Y(j-1) \mid S(1) \geq j > S(0)\},$$

so that β can be viewed as a weighted per-unit treatment effect. The weights can be viewed as the proportion of compliers within each level of the exposure, that is, the proportion of individuals whose exposure level is shifted from $j - 1$ or less to j or more due to the instrument.

Suppose S is miscoded as $D = I(S \geq l)$ for some $1 \leq l \leq J$. In their corollary of Theorem 1, Angrist and Imbens (1995) show that the Wald estimator takes the form

$$\frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)} = \phi \, \beta, \tag{2.4}$$

where

$$\phi = \frac{\sum_{j=1}^{J} \text{pr}\{S(1) \geq j > S(0)\}}{\text{pr}\{S(1) \geq l > S(0)\}}.$$

The insight from this expression is that $\phi = 1$ only when the instrument has no effect on the exposure except through switching individuals from $S = l - 1$ to $S = l$. Otherwise, $\phi > 1$ and the estimand is systematically larger than β.

Marshall (2016) is the first to refer to the phenomena where $\phi > 1$ as *coarsening bias*, which I use throughout Chapter 4. The author notes that (2.4) can equivalently be written as

$$\beta_l + \frac{\sum_{j \neq l} \text{pr}\{S(1) \geq j > S(0)\} \, \beta_j}{\text{pr}\{S(1) \geq l > S(0)\}}.$$

The author then provides some additional assumptions under which β_l is identified, namely,

1. $\text{pr}\{S(1) \geq j > S(0)\} = 0$ for all $j \neq l$ (this is equivalent to Angrist and Imbens (1995)'s identification condition).

2. $\beta_j = 0$ for all $j \neq l$.

3. $\text{pr}\{S(1) \geq j > S(0)\}\, \beta_j = 0$ for all $j \neq l$.

4. $\sum_{j \neq l} \text{pr}\{S(1) \geq j > S(0)\}\, \beta_j = 0$.

2.2 Genetics

This section introduces some definitions and concepts in genetics. My thesis does not require a comprehensive understanding of genetics, but a basic foundation is needed. This section amalgamates and condenses Davey Smith and Ebrahim, 2003, p.3–6 and Section 1.1 of Thompson (2000).

DNA is a molecule found in the nuclei of human cells that encodes instructions for development, function and reproduction. DNA consists of strands of *base pairs*, different combinations of which encode different instructions. The four distinct base pairs are abbreviated by A, C, G and T. In the nucleus, DNA is tightly wrapped into *chromosomes*, which are doubled strands of helical DNA. The number of base pairs across all chromosomes in the human genome is roughly 3×10^9.

A base pair which exhibits population-level variation in its nucleotides is called a *single nucleotide polymorphism* (SNP). DNA sequences are typically characterized by the detectable variant forms induced by different combinations of SNPs. These variant forms are called *alleles*. In this thesis, unless otherwise stated, I only consider variants with two alleles. The classification of alleles at a particular locus can be varied. The allele with the higher prevalence in a given population is typically called the *major allele*, in contrast to the *minor allele*. Alternatively, alleles can be classified according to a particular reference genome, where one will be called the *reference allele* and the other the *alternate allele*. I use both terms where appropriate.

Human cells can be broadly categorized into two types: *germ cells* and *somatic cells*. Germ cells, such as eggs and sperm, are involved in sexual reproduction. Somatic cells include all other cells in the body, except for undifferentiated stem cells. The two cell types are typically distinguished by the number of chromosomes they contain. Human somatic cells consist of 23 pairs of chromosomes, with one in each pair inherited from the mother and the other from the father. Germ cells consist of only one set of chromosomes.

Meiosis is a type of cell division that results in germ cells containing one copy of each chromosome. During this process, homologous chromosomes line up and exchange segments of DNA between themselves in a biochemical process called *crossing over*. The recombined chromosomes are then further divided and separated into germ cells. Since crossovers are infrequent (roughly one to four per chromosome in most eukaryotes) SNPs located nearby on the same parental chromosome are more likely to be transmitted together, which can result in *linkage disequilibrium*, which is defined as population-level correlation between SNPs.

Fertilization is the process by which germ cells in the father and mother join together to form a *zygote* (fertilized egg cell), which will normally develop into an embryo.

The following glossary summarizes some of the terminology described in this section, and other terminology that will arise in the thesis.

Table 2.4: Glossary of genetic terms

Name	Definition
Allele	Detectable variant forms of a DNA sequence.
Gene	A DNA sequence which codes for a particular biological function.
Genotype	An allele (or group of alleles) inherited at a specific site. If the alleles are the same, the genotype is *homozygous*. If the alleles are different, it is *heterozygous*.
Haplotype	A group of alleles on the same chromosome inherited together from the same parent. The corresponding genotype is the maternally-inherited and paternally-inherited haplotypes.
Linkage disequilibrium	Non-zero correlation between two SNPs at distinct sites.
Locus	A position in a DNA sequence. It can refer to a SNP or longer regions within the sequence.
Marker	A segment of DNA with an identifiable physical location on a chromosome which can be measured using a genetic sequencing assay. Markers are useful in genetic studies when they are correlated with unobserved, but functionally relevant, SNPs or DNA sequences.
Phenotype	An observable, measurable trait in an individual (e.g., height, eye colour).
Pleiotropy	A phenomenon where SNPs induce variation in more than one phenotype.
Single nucleotide polymorphism (SNP)	A position on a chromosome where base pairs differ among individuals in a population.

2.3 Mendelian randomization

2.3.1 A brief history of Mendelian randomization

Mendelian randomization (MR) is a causal inference approach that uses the random allocation of genes from parents to offspring as a foundation for causal inference (Sanderson et al., 2022). The ideas behind MR can be traced back to the intertwined beginning of modern statistics and genetics about a century ago. In one of the earliest examples, Wright (1920) used selective

inbreeding of guinea pigs to investigate the causes of colour variation and, in particular, the relative contribution of heredity and environment. In a later defence of this work, Wright (1923, p. 251) argued that his analysis of path coefficients, a precursor to modern causal graphical models, "rests on the validity of the premises, i.e.,, on the evidence for Mendelian heredity", and the "universality" of Mendelian laws justifies ascribing a causal interpretation to his findings.

At around the same time, Fisher (1926) started to contemplate the randomization principle in experimental design and used it to justify his analysis of variance (ANOVA) procedure, which was motivated by genetic problems. In fact, the term "variance" first appeared in Fisher's groundbreaking paper that bridged Darwin's theory of evolution and Mendel's theory of genetic inheritance (Fisher, 1918). Fisher (1935) described randomization as the "reasoned basis" (p. 12) for inference and "the physical basis of the validity of the test" (p. 17). Later, it was revealed that his factorial method of experimentation derives "its structure and its name from the simultaneous inheritance of Mendelian factors" (Fisher, 1951, p. 330). Indeed, Fisher viewed randomness in meiosis as uniquely shielding geneticists from the difficulties of establishing reliably controlled comparisons, remarking that "the different genotypes possible from the same mating have been beautifully randomized by the meiotic process" (Fisher, 1951, p. 332).

While this source of randomization was originally used for eliciting genetic causes of phenotypic variation, it was later identified as a possible avenue for understanding causation among modifiable phenotypes themselves (Davey Smith, 2006). Lower et al. (1979) used N-acetylation, a phenotype of known genetic regulation and a component of detoxification pathways for arylamine, to strengthen the inference that arylamine exposure causes bladder cancer. Katan (1986) proposed to address reverse causation in the hypothesized effect of low serum cholesterol on cancer risk via polymorphisms in the apolipoprotein E (*APOE*) gene. He argued that, if low cholesterol was indeed a risk factor for cancer, we would expect to see higher rates of cancer in individuals with the low cholesterol allele. Another pioneering application of this reasoning can be found in a proposed study of the effectiveness of bone marrow transplantation relative to chemotherapy (Gray and Wheatley, 1991), for example, in the treatment of acute myeloid leukaemia (Wheatley and Gray, 2004). Patients with a compatible donor sibling were more likely to receive transplantation than patients without. Since compatibility is a consequence of random genetic assortment, comparing survival outcomes between the two groups can be viewed as akin to an intention-to-treat analysis in a randomized controlled trial. This paper appears to be the first to use the term "Mendelian randomization".

It would be a dozen more years before an argument for the broader applicability of MR was put forward by Davey Smith and Ebrahim (2003). At the time, a number of criticisms had been levelled against the state of observational epidemiology and its methods of inquiry (Feinstein, 1988; Taubes, 1995; Davey Smith, 2001). Several high profile results failed to be corroborated by subsequent randomized controlled trials, such as the role of beta-carotene consumption in lowering risk of cardiovascular disease, with unobserved confounding identified as the likely

culprit (Davey Smith, 2001, p. 329-330). This string of failures motivated the development of a more rigorous observational design with an explicit source of unconfounded randomization in the exposures of interest (Davey Smith, Michael, et al., 2020).

Originally, Davey Smith and Ebrahim (2003) recognized that MR is best justified in a within-family design with parent-offspring trios. MR is commonly described as being analogous to a randomized controlled trial with non-compliance. This analogy is based on exact randomization in the transmission of alleles from parents to offspring which can be viewed as a form of treatment assignment. From its inception, it was recognized that data limitations would largely restrict MR to be performed in samples of unrelated individuals, which Davey Smith and Ebrahim (2003) termed "approximate MR". Such approximate MR has been the norm, seen in the majority of applied and methodological studies to date. However, MR in unrelated individuals lacks the explicit source of randomization offered by the within-family design, thereby suffering potential biases from dynastic effects, population structure and assortative mating (Davies, Howe, et al., 2019; Brumpton et al., 2020; Howe, Nivard, et al., 2022).

In addition to random assignment of exposure-modifying genetic variants, we must also assume that the effects of these genetic variants on the outcome are fully mediated by the exposure, known as the exclusion restriction. When this assumption holds, MR can be framed as a special case of instrumental variable analysis (Thomas and Conti, 2004; Didelez and Sheehan, 2007).

It is common practice in MR studies to select instruments based on the results of genome-wide association studies. These studies test the marginal association between each genetic variant and the exposure, with genetic variants satisfying some p-value threshold then selected as instruments. See Swerdlow et al. (2016) for a more detailed description of instrument selection in MR studies. Since instruments are typically selected due to their statistical association with the exposure, there is often minimal biological evidence that they satisfy instrument validity. Of particular concern is violations of the exclusion restriction. There is growing recognition that gene regulatory networks are sufficiently interconnected that almost all genetic variants will exhibit some influence on a given trait (Boyle, Li, and Pritchard, 2017). This has led to considerable recent methodological work to replace the exclusion restriction with more plausible assumptions, typically by placing structure on the sparsity (Kang, Zhang, et al., 2016a) or distribution of pleiotropic effects across individual genetic variants (Bowden, Davey Smith, and Burgess, 2015; Zhao, Wang, et al., 2020; Kolesár et al., 2015).

2.3.2 Summary data MR

An additional limitation of MR is the unavailability of individual-level genetic data. Typically, genetic data is released by large consortia in the form of *summary statistics*, which include the coefficients and standard errors from regressions of some trait on each variant. This has led to a sizeable literature on bespoke methods for summary data MR. In this section, I will

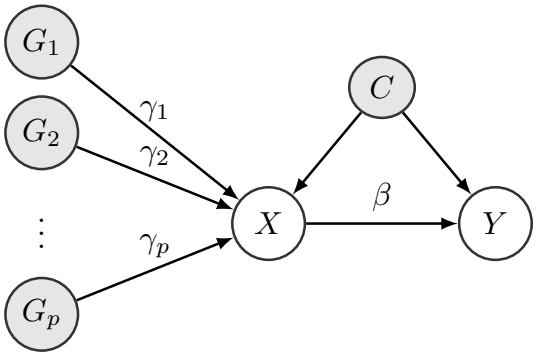


Figure 2.3: Graphical representation of the canonical MR model

briefly review the statistical foundation of summary data MR and some of the most prominent methods, signposting further reading where appropriate.

Summary data MR is typically motivated by linear structural equation models (Section 2.1.3) (Burgess, Butterworth, and Thompson, 2013; Bowden, Davey Smith, and Burgess, 2015; Zhao, Wang, et al., 2020). Suppose we have an outcome Y, exposure X and genetic instruments $G_j \in \{0, 1, 2\}$, $j = 1, \ldots, p$. $G_j = 2$ indicates that an individual has two minor or alternate alleles for SNP j, while $G_j = 0$ indicates they have two major or reference alleles. Each G_j regulates (or is a marker for a variant that regulates) the exposure X. The gene-exposure and exposure-outcome relationships are given by

$$X = \sum_{j=1}^{p} \gamma_j G_j + \varepsilon_X,$$
$$Y = \beta X + \varepsilon_Y,$$

where $\varepsilon_X \not\perp\!\!\!\perp \varepsilon_Y$ and $G_j \perp\!\!\!\perp (\varepsilon_X, \varepsilon_Y)$, $j = 1, \ldots, p$. Furthermore, it is commonly assumed that $G_j \perp\!\!\!\perp G_{j'}$ for any $j \neq j'$. This can be justified when instruments are located on different chromosomes or sufficiently far apart on the same chromosome. In MR, instruments are typically markers for their surrounding genomic regions. These markers are selected by a process called linkage disequilibrium (LD) clumping (Hemani, Zheng, et al., 2018). The model proposed in this paragraph can be summarized by Figure 2.3, where C is some unobserved confounder.

From the observed data tuple $(G_1, \ldots, G_p, X, Y)$ it is possible to identify the parameters γ_j and $\Gamma_j = \beta\gamma_j$. From there, the Wald estimator is simply the ratio of the latter over the former $\beta = \Gamma_j/\gamma_j$. This simple observation is the basis of summary data MR. Genome wide association studies will typically release summary estimates $\hat{\gamma}_j$ and $\hat{\Gamma}_j$ and their standard errors σ_{γ_j} and σ_{Γ_j}, in lieu of individual level data.

Burgess, Butterworth, and Thompson (2013) first propose an approach for summary data MR under the model

$$\hat{\Gamma}_j \sim \mathcal{N}(\Gamma_j, \sigma^2_{\Gamma_j})$$

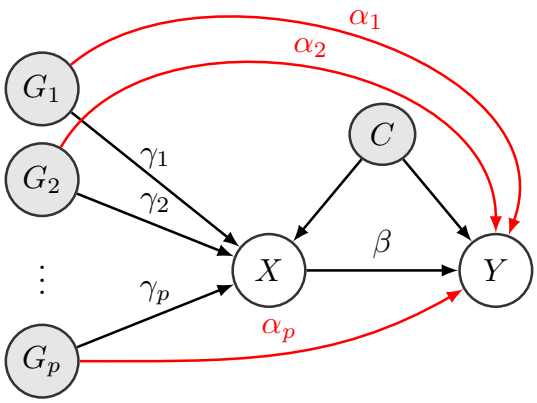


Figure 2.4: Graphical representation of the MR model with pleiotropy

which can be justified via the central limit theorem. The authors assume that $\hat{\gamma}_j$ is measured without error, so that it may be viewed as a fixed parameter (Bowden, Del Greco M, et al., 2016; Bowden, Del Greco M, et al., 2017).

Assumption 2.9. (No Measurement Error, NOME) $\hat{\gamma}_j = \gamma_j$ with probability one.

They then solve the log-likelihood

$$\ell(\beta) = -\frac{1}{2}\sum_{j=1}^{p}\frac{(\hat{\Gamma}_j - \beta\gamma_j)^2}{\sigma^2_{\Gamma_j}}, \quad \hat{\beta} = \frac{\sum_{j=1}^{p}\hat{\Gamma}_j\gamma_j\sigma^{-2}_{\Gamma_j}}{\sum_{j=1}^{p}\gamma_j^2\sigma^{-2}_{\Gamma_j}}$$

This gives rise to the so-called *inverse variance weighted* (IVW) estimator for β. The standard error is simply given by

$$\sigma_\beta = \left(\sum_{j=1}^{p}\hat{\gamma}_j^2\sigma^{-2}_{\Gamma_j}\right)^{-1/2}$$

As discussed in Section 2.3.1, MR studies often utilize many potentially invalid instruments. This is represented with the following model (Bowden, Davey Smith, and Burgess, 2015):

$$X = \sum_{j=1}^{p}\gamma_j G_j + \varepsilon_X$$

$$Y = \beta X + \sum_{j=1}^{p}\alpha_j G_j + \varepsilon_Y$$

In this model, each genetic variant G_j can exert a direct *pleiotropic* effect α_j on Y not mediated through X. This violates the exclusion restriction. Figure 2.4 summarizes this more general model.

To regain identification, Bowden, Davey Smith, and Burgess (2015) propose an assumption which ensures that each variant's effect on the exposure γ_j is independent of its direct effect on the outcome α_j. This assumption views the parameters α_j and γ_j as random effects.

Assumption 2.10. (Instrument Strength Independent of Direct Effect, InSIDE) $\text{cov}(\gamma_j, \alpha_j) = 0$.

The motivation for this assumption is that

$$\frac{\text{cov}(\gamma_j, \Gamma_j)}{\text{var}(\gamma_j)} = \beta + \frac{\text{cov}(\gamma_j, \alpha_j)}{\text{var}(\gamma_j)} \overset{\text{InSIDE}}{=} \beta. \tag{2.5}$$

Therefore, a weighted linear regression of $\hat{\Gamma}_j$ on $\hat{\gamma}_j$ with weights $\sigma_{\gamma_j}^{-2}$ leads to a consistent estimator for β under InSIDE. Bowden, Davey Smith, and Burgess (2015) refer to this as the *MR Egger* estimator, in reference to the author of a similar approach for addressing small study bias in meta-analysis (Egger, Davey Smith, and Minder, 1997).

Alongside MR Egger, there is an expansive methodological literature in MR concerned with relaxing the exclusion restriction with different assumptions, typically related to the sparsity or distribution of pleiotropic effects (Bowden, Davey Smith, et al., 2016; Kang, Zhang, et al., 2016b; Hartwig, Davey Smith, and Bowden, 2017; Guo et al., 2018; Tchetgen, Sun, and Walter, 2021; Liu et al., 2022).

The NOME assumption that $\hat{\gamma}_j = \gamma_j$ will lead to erroneous standard errors, particularly in the presence of *weak instruments*, which are only marginally associated with the exposure (Stock and Trebbi, 2003). Zhao, Wang, et al. (2020) propose a robust profile likelihood estimator which considers sampling variation in both $\hat{\gamma}_j$ and $\hat{\Gamma}_j$. This estimator is based on the model

$$\begin{pmatrix} \hat{\gamma}_j \\ \hat{\Gamma}_j \end{pmatrix} \sim \mathcal{N}_2 \left\{ \begin{pmatrix} \gamma_j \\ \Gamma_j \end{pmatrix}, \begin{pmatrix} \sigma_{\gamma_j}^2 & \rho\sigma_{\gamma_j}\sigma_{\Gamma_j} \\ \rho\sigma_{\gamma_j}\sigma_{\Gamma_j} & \sigma_{\Gamma_j} \end{pmatrix} \right\},$$

where $\mathcal{N}_2$ denotes the bivariate normal distribution. When the gene-exposure parameter $\hat{\gamma}_j$ and gene-outcome parameter $\hat{\Gamma}_j$ are estimated in non-overlapping samples, $\rho = 0$. This is common in genetic association studies, where different traits are often measured in different samples. The case where $\rho \neq 0$ is discussed in Wang, Zhao, et al. (2021).

The log-likelihood takes the form

$$\ell(\beta, \gamma_1, \ldots, \gamma_p) = -\frac{1}{2}\left[\sum_{j=1}^{p} \frac{(\hat{\gamma}_j - \gamma_j)^2}{\sigma_{\gamma_j}^2} + \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\gamma_j)^2}{\sigma_{\Gamma_j}^2}\right].$$

Zhao, Wang, et al. (2020) view the parameters $\gamma_1, \ldots, \gamma_p$ as nuisance parameters. These parameters can be profiled out of the log-likelihood, such that

$$\ell(\beta) = \max_{\gamma_1,\ldots,\gamma_p} \ell(\beta, \gamma_1, \ldots, \gamma_p) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\gamma_j)^2}{\sigma_{\gamma_j}^2\beta^2 + \sigma_{\Gamma_j}^2}.$$

This estimator has good efficiency and consistency properties. Within this framework, Zhao, Wang, et al. (2020) consider a random effects model for pleiotropy where $\alpha_j \sim \mathcal{N}(0, \tau_0^2)$ and τ_0^2

is not too large. This leads to the profile log-likelihood

$$\ell(\beta, \tau^2) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\gamma_j)^2}{\sigma^2_{\gamma_j}\beta^2 + \sigma^2_{\Gamma_j} + \tau^2} + \log(\sigma^2_{\Gamma_j} + \tau^2).$$

The authors generalize this to a contaminated normal distribution, where some small number of $|\alpha_j|$ can be large.

Chapter 3

Sample-constrained partial identification

Publications arising from this chapter

This chapter has been published in the journal *Biometrika* under the title "Sample-constrained partial identification with application to selection bias" (Tudball, Hughes, et al., 2022). My contribution was: conceptualization; stating and proving theorems, lemmas and propositions; coding and validating the accompanying R package; designing and analyzing the simulation study and applied example; writing the manuscript and revising in response to co-author and reviewer comments. Co-authors' contribution was: assistance with proving the main theorem; suggestions for the design, analysis and interpretation of the applied example; providing feedback on the manuscript and reviewer response.

SIGNED: Matthew Tudball (First Author) DATE: 24 July, 2022
SIGNED: Qingyuan Zhao (Senior Author) DATE: 24 July, 2022

Section 3.7 is an extension of the online appendix of an article published in *Nature Communications* titled "Collider bias undermines our understanding of COVID-19 disease risk and severity" (Griffith et al., 2020). My contribution to the article was: assisting with writing the manuscript; writing and data analysis for the online appendix.

SIGNED: Gareth Griffith (First Author) DATE: 24 July, 2022
SIGNED: Gibran Hemani (Senior Author) DATE: 24 July, 2022

Software

I prepared an open source R package to accompany this publication. It implements the sensitivity analysis for selection bias described in Section 3.3. See `https://github.com/matt-tudball/selectioninterval` for installation instructions.

3.1 Introduction

3.1.1 General problem

Partial identification problems arise when the observable data are only sufficient to identify a set or interval in which a parameter of interest is contained. A classical example from Manski (2003) is the missing data problem, where Y is a discrete random variable and S is a binary random variable indicating whether Y is observed ($S = 1$) or not ($S = 0$). The distribution of Y can be decomposed into

$$\text{pr}(Y = y) = \text{pr}(Y = y \mid S = 1)\,\text{pr}(S = 1) + \text{pr}(Y = y \mid S = 0)\,\text{pr}(S = 0) \tag{3.1}$$

for any y in the support of Y. Given that $\text{pr}(Y = y \mid S = 0)$ is unobserved, the smallest value that $\text{pr}(Y = y)$ could take is $\text{pr}(Y = y \mid S = 1)\,\text{pr}(S = 1)$ and the largest value is $\text{pr}(Y = y \mid S = 1)\,\text{pr}(S = 1) + \text{pr}(S = 0)$. Therefore, although $\text{pr}(Y = y)$ itself cannot be point identified, it can be partially identified via the interval corresponding to the smallest and largest possible values.

Many partial identification problems can be formulated as the optimal value of a population objective function, which we write as

$$\nu = \inf\{Q(\theta) \colon \theta \in \Theta\}, \tag{3.2}$$

where $Q : \mathbb{R}^p \to \mathbb{R}$ and $\Theta \subseteq \mathbb{R}^p$. In the missing data example, $Q(\theta) = \text{pr}(Y = y \mid S = 1)\,\text{pr}(S = 1) + \theta\,\text{pr}(S = 0)$ and $\Theta = [0, 1]$.

The field of stochastic optimization also considers problems of the form in (3.2) and has built a large literature on estimation of, and inference for, ν when a sample analogue Q_n is observed instead of Q (where n denotes the sample size). We demonstrate that framing the partial identification problem as a stochastic optimization problem will allow us to draw upon these existing results.

Specifically, in this article we are concerned with the difficult setting where Θ must also be estimated empirically. We consider a setting where Θ is characterized by inequality constraints of the form $\Theta = \{\theta \colon h_j(\theta) \leq 0, j = 1, \ldots, J\}$, where we may only observe corresponding estimators $h_{nj}(\theta)$. Within this setting, our goal is to find a lower confidence bound C_n for any $0 < \alpha < 1$ such that

$$\lim_{n\to\infty} \text{pr}(C_n \leq \nu) \geq 1 - \alpha, \tag{3.3}$$

which will suffice to provide useful statistical inference in a wide set of applications.

3.1.2 Motivating application

Our investigation is motivated by an applied question: how will selection bias affect the conclusions of population-based cohort studies? Many statistical analyses begin by selecting a study sample from some population of interest. When the sample is drawn non-randomly, then valid inference for the population is no longer guaranteed; see Heckman (1979) and Bareinboim, Tian, and Pearl (2014) for references. Inverse probability weighting could be used to correct for this selection bias (Horvitz and Thompson, 1952; Stuart et al., 2011), but data on non-selected observations may be limited or unavailable altogether, such that the weights cannot be estimated. In such settings, there exist approaches to assess the sensitivity of estimates to a range of plausible inverse probability weights (Aronow and Lee, 2013; Thompson and Arah, 2014). However, these approaches could be made more informative via a principled procedure for conducting statistical inference and the inclusion of relevant auxiliary information about the population. We demonstrate that such improvements can be made by casting these sensitivity analyses within the general framework described in Section 3.1.1.

We are specifically motivated by studies conducted in UK Biobank, which is a large population-based cohort study widely analyzed by health researchers. Studies of this cohort are potentially biased since recruited participants are known to differ systematically from the rest of the UK population on measures such as education, health status, age and geographical location (Fry et al., 2017; Hughes et al., 2019).

Another application for which our statistical framework could be applied is partial identification of causal effects with discrete instrumental variables. The partial identification results in Richardson, Evans, and Robins (2011) can be formulated as an optimization problem subject to constraints, which is amenable to our framework.

3.1.3 Existing literature

Statistical inference procedures have been developed for some special cases of our general problem in eq. (3.3). An area of particular focus is the so-called "sample average approximation" (Shapiro, Dentcheva, and Ruszczyński, 2009). In this case, Q is the expected value $Q(\theta) = E\{f(\theta, X)\}$ of some function f and Q_n is a sample average $Q_n(\theta) = n^{-1} \sum_{i=1}^{n} f(\theta, X_i)$, where X is some random variable and $X_1, \ldots, X_n$ are independent draws of X.

Statistical inference in the presence of Θ_n has been developed for convex sample average approximations, such that f is convex in θ and $\Theta = \{\theta \colon h_j(\theta) \leq 0, j = 1, \ldots, J\}$ where $h_j(\theta) = E\{g_j(\theta, X)\}$ and $g_j(\theta, X)$ is convex in θ for all j. Shapiro (1991) shows that the plug-in estimator

$$\nu_n^p = \inf\{Q_n(\theta) \colon \theta \in \Theta_n\}. \tag{3.4}$$

satisfies a central limit theorem under these convexity assumptions (and some additional regularity conditions), where $\Theta_n = \{\theta \colon \sum_{i=1}^{n} g_j(\theta, X_i) \leq 0, j = 1, \ldots, J\}$.

Moving away from convex problems, Wang and Ahmed (2008) consider the special case of minimizing a known function Q subject to a single expected value constraint $\Theta = \{\theta \colon E\{g(\theta, X)\} \leq 0\}$. They propose an approach for calculating a sample size n so that Θ_n is feasible to some small relaxation of the true problem with high probability.

Our work also overlaps with the partial identification literature in econometrics, much of which considers inference for identified sets characterized by conditional or unconditional moment inequalities, commonly interpreted as the set of minimizers of some criterion function (Chernozhukov, Hong, and Tamer, 2007; Andrews and Soares, 2010; Andrews and Shi, 2013). A related literature provides inference for parameters lying within partially identified sets, as opposed to inference for the set itself (Imbens and Manski, 2004; Stoye, 2009). For a more comprehensive review of the partial identification literature, see Molinari (2020).

A partially identified set can often be formulated as a region over which a likelihood is maximized, with point identification achieved when this region is a singleton (Giacomini and Kitagawa, 2021). Our framework is consistent with this interpretation, where Θ can be viewed as a region corresponding to a flat likelihood and $Q(\theta)$ can be viewed as some function defined over this region. In the missing data problem in Section 3.1, the likelihood is flat with respect to $\text{pr}(Y = y \mid S = 0)$ but we want to minimize another function, $\text{pr}(Y = y)$, over this flat region.

3.2 Confidence intervals for sample-constrained partial identification

3.2.1 Confidence intervals under known constraints

In this section, we briefly summarize existing results on statistical inference for stochastic optimization when the set Θ is observed, which forms the basis of our generalization to situations where an estimate Θ_n of Θ is observed instead. Suppose the parameter space is defined by a set of inequality constraints

$$\Theta = \big\{\theta \colon h_j(\theta) \leq 0,\, j = 1, \ldots, J\big\} \tag{3.5}$$

where an equality constraint for some $h_j(\theta)$ can be introduced by taking the inequality constraints of both $h_j(\theta)$ and $-h_j(\theta)$. Recall that our goal is to provide inference about the infimum $\nu = \inf\{Q(\theta) \colon \theta \in \Theta\}$.

Much of the literature in stochastic optimization is centered on the statistical properties of the estimator

$$\nu_n = \inf\{Q_n(\theta) \colon \theta \in \Theta\}. \tag{3.6}$$

Consistency of optimal values and optimal solutions to such stochastic optimization problems is typically achieved by imposing uniform convergence of $Q_n(\theta)$ to $Q(\theta)$. First order asymptotic

properties are obtained via the functional delta method. The key conditions are that the infimum, viewed as a function of Q, satisfies some notion of differentiability at Q and that $n^{-1/2}(Q - Q_n)$ converges to a Gaussian process. See Shapiro (1991) for further details.

To make the previous discussion more concrete, consider the following four assumptions commonly placed on the stochastic optimization problem described above.

Assumption 3.1. The set of solutions to (3.2) is a singleton $\{\theta \in \Theta \colon Q(\theta) = \nu\} = \{\vartheta\}$.

Assumption 3.2. Let $B \subseteq \mathbb{R}^p$ denote a compact set and $C(B)$ denote the space of continuous functions on domain B. Then $\Theta \subseteq B$, $Q \in C(B)$ and $Q_n \in C(B)$ with probability one.

Assumption 3.3. $Q_n(\theta)$ converges to $Q(\theta)$ with probability one as $n \to \infty$ uniformly on B.

Assumption 3.4. As $n \to \infty$, the sequence $V_n(\theta) = n^{1/2}\{Q(\theta) - Q_n(\theta)\}$ converges in distribution to a random element $V(\theta) \in C(B)$, where $V(\theta)$ is Gaussian process with mean 0 and variance $\sigma^2(\theta) \in C(B)$.

These assumptions are jointly sufficient to achieve consistency and asymptotic normality of ν_n, which we state formally in the following two propositions.

Proposition 3.1. *Let $\vartheta_n \in \arg\min\{Q_n(\theta) \colon \theta \in \Theta\}$ be a sample solution and ν_n be defined as in eq.* (3.6)*. Under Assumptions 3.1, 3.2 and 3.3, $\nu_n \to \nu$ and $\vartheta_n \to \vartheta$ with probability one.*

Proposition 3.1 is identical to Theorem 5.3 in Shapiro, Dentcheva, and Ruszczyński (2009) under the condition that ϑ is unique.

Proposition 3.2. *Under Assumptions 3.1, 3.2 and 3.4,*

$$n^{1/2}(\nu_n - \nu) \to \mathcal{N}\big\{0, \sigma^2(\vartheta)\big\} \tag{3.7}$$

in distribution, where $\sigma^2(\vartheta)$ is the asymptotic variance of ν_n defined in Assumption 3.4.

Proposition 3.2 is an immediate consequence of Theorem 3.2 in Shapiro (1991). Although we do not restate the proof here, the intuition is that Assumptions 3.1 and 3.2 allow a notion of differentiability of the infimum and Assumption 3.4 provides weak convergence of $n^{1/2}(Q - Q_n)$ to a Gaussian process, thus providing the conditions needed for an application of the delta method.

To use Proposition 3.2 to construct a valid confidence interval, we must take into consideration that both σ^2 and ϑ are unknown. To this end, we state an additional assumption followed by a proposition.

Assumption 3.5. There exists a uniformly strongly consistent estimator $\sigma_n^2(\theta) \in C(B)$ for $\sigma^2(\theta)$ such that $\sup_{\theta\in\Theta} |\sigma_n^2(\theta) - \sigma^2(\theta)| \to 0$ with probability one.

Proposition 3.3. *Under Assumptions 3.1, 3.2, 3.3 and 3.5, $\sigma_n^2(\vartheta_n) \to \sigma^2(\vartheta)$ with probability one.*

Assumption 3.5 applies uniform convergence to an estimator for the asymptotic variance of $Q_n(\theta)$. This strong notion of convergence for $\sigma_n^2(\vartheta_n)$ allows us to construct a confidence bound of the form

$$C_n = \nu_n - Z_\alpha \, \sigma_n(\vartheta_n) \, n^{-1/2} \tag{3.8}$$

where Z_α is the upper α-quantile of the standard normal distribution. This choice of C_n has asymptotically exact nominal coverage $1 - \alpha$ by Proposition 3.2, Proposition 3.3 and Slutsky's theorem.

3.2.2 Confidence intervals under sample constraints

We now consider the more difficult setting where the constraint functions $h_j(\theta)$ need to be estimated as well. We instead observe an estimator $\Theta_n = \{\theta \colon h_{nj}(\theta) \leq 0, j = 1, \ldots, J\}$ comprised of estimators of the constraint functions $h_{nj}(\theta)$. We will discuss what properties Θ_n must have to allow valid statistical inference for ν.

It is tempting to follow the approach of the previous section and construct a plug-in estimator for ν by simply replacing Q with Q_n and Θ with Θ_n and finding the corresponding infimum. This is the approach taken in Shapiro, Dentcheva, and Ruszczyński (2009), given by ν_n^p in eq. (3.4). A problem with this approach is that it is possible that $\Theta \cap \Theta_n = 0$ with probability one even as n becomes large. This means that the true solution ϑ will almost never lie inside Θ_n, prohibiting the construction of a valid confidence interval for ν as illustrated by the contrived example below.

Example 3.1. Consider a problem of the form $Q(\theta) = \theta^2 + E(X)$ and $\Theta = \big\{\theta \colon \theta = E(X)\big\}$ where $X \sim \mathcal{N}(1, 1)$ is a normally-distributed random variable. The plug-in estimators are $Q_n(\theta) = \theta^2 + \bar{X}_n$ and $\Theta_n = \{\theta \colon \theta = \bar{X}_n\}$, where $\bar{X}_n$ is the mean of n independent and identically distributed draws of X. It follows that $\nu = 2$ and $\nu_n^p = \bar{X}_n^2 + \bar{X}_n$, where ν_n^p is the plug-in estimator in eq. (3.4). The asymptotic variance of $Q_n(\theta)$ is $\sigma^2(\theta) = 1$, which we assume is known. The confidence bound in eq. (3.3) is $C_n = \bar{X}_n^2 + \bar{X}_n - Z_\alpha n^{-1/2}$. A simple Monte Carlo simulation demonstrates that the corresponding 95% confidence interval for $n = 100$ exhibits sub-nominal coverage of around 70%.

Existing approaches in stochastic optimization address the problem in Example 3.1 by restricting to sample average approximations and imposing convexity of both Q and h. To allow inference for a broader class of problems, we propose an intuitive but conservative approach which replaces Θ_n with an appropriate relaxation. In particular, we propose to use the relaxed set

$$\Theta_n^r = \big\{\theta \colon h_{nj}(\theta) \leq \epsilon_{nj}(\theta), \, j = 1, \ldots, J\big\} \tag{3.9}$$

where $\epsilon_n(\theta) = (\epsilon_{n1}(\theta), \ldots, \epsilon_{nJ}(\theta))^T$ is some J-dimensional sequence such that $\epsilon_{nj}(\theta) \geq 0$ for all $\theta \in B$, chosen so that

$$\lim_{n\to\infty} \text{pr}(\Theta \subseteq \Theta_n^r) \geq 1 - \alpha_1 \tag{3.10}$$

for some $0 < \alpha_1 < 1$. The exact forms of Θ_n^r and $\epsilon_n(\theta)$ are not crucial for our main results, provided (3.10) holds, which we discuss in more detail toward the end of this section.

Our proposed confidence bound is of the form $C_n(\theta) = Q_n(\theta) - Z_{\alpha_2}\sigma_n(\theta)n^{-1/2}$ for some $0 < \alpha_2 < 1$, where Z_{α_2} is the upper α_2-quantile of the standard normal distribution. We need to select a θ so that eq. (3.3) is satisfied. This is accomplished by finding the optimal value and solution over the relaxed constraint set,

$$\nu_n^r = \inf\{Q_n(\theta) \colon \theta \in \Theta_n^r\} \text{ and } \vartheta_n^r \in \arg\min\{Q_n(\theta) \colon \theta \in \Theta_n^r\}, \tag{3.11}$$

and constructing a confidence bound of the form

$$C_n = C_n(\vartheta_n^r) = \nu_n^r - Z_{\alpha_2}\, \sigma_n(\vartheta_n^r)\, n^{-1/2}. \tag{3.12}$$

We now need to demonstrate that C_n covers ν with known probability in the limit. To this end, we need an additional technical assumption to hold.

Assumption 3.6. Let $\zeta_n^r \in \arg\min\{C_n(\theta) \colon \theta \in \Theta_n^r\}$ be the optimal solution of $C_n(\theta)$ over Θ_n^r, then $|\zeta_n^r - \vartheta_n^r|$ converges to 0 in probability.

Assumption 3.6 is imposed so that two important quantities become asymptotically close. The first quantity is $C_n(\zeta_n^r)$, which is the infimum over all confidence bounds in Θ_n^r. This confidence bound is important because it provides a lower bound for other quantities with known coverage probabilities, which is a fact we utilize in our main result in Theorem 3.1. The second quantity is $C_n(\vartheta_n^r)$, which is our main confidence bound proposed in (3.12).

We argue that Assumption 3.6 is reasonable in the sense that ζ_n^r and ϑ_n^r are solutions over two objective functions which converge uniformly to the same limit. To make this intuition more concrete, we provide some sufficient conditions for Assumption 3.6 in the Supplementary Material. Essentially, if Assumption 3.3 is satisfied and $h_{nj}(\theta)$ and $\epsilon_{nj}(\theta)$ converge to $h_j(\theta)$ and 0 for all $j = 1, \ldots, J$ uniformly on B with probability one, then we can show that both ϑ_n^r and ζ_n^r converge to ϑ with probability one.

We claim that C_n provides an asymptotically valid lower confidence bound.

Theorem 3.1. *Suppose we select a relaxed constraint set Θ_n^r as in eq. (3.9) and significance level $0 < \alpha_1 < 1$ such that $\lim_{n\to\infty} \text{pr}(\Theta \subseteq \Theta_n^r) \geq 1 - \alpha_1$. Suppose we also select a significance level $0 < \alpha_2 < 1$. Then, under Assumptions 3.1 - 3.6,*

$$\lim_{n\to\infty} \text{pr}(C_n \leq \nu) \;\geq\; 1 \;-\; \alpha_1 \;-\; \alpha_2.$$

Here we outline the key steps in the proof. We begin by defining a deterministic sequence $\delta_n = Z_{\alpha_2}\epsilon n^{-1/2}$ where $\epsilon > 0$ is some small constant. We then show that $\text{pr}(C_n \leq \nu)$ is bounded from below by the sum of two quantities: $\text{pr}\{C_n(\zeta_n^r) \leq \nu - \delta_n\}$ and $\text{pr}\{|\sigma_n(\zeta_n^r) - \sigma_n(\vartheta_n^r)| \leq \epsilon\} - 1$. The second quantity converges to 0 under Assumption 3.6. The remainder of the proof follows a similar argument to the main lemma of Berger and Boos (1994). Whenever $\Theta \subseteq \Theta_n^r$, we know that $C_n(\zeta_n^r)$, which is the infimum over all confidence bounds in Θ_n^r, will cover ν at least as often as $C_n(\vartheta_n)$, which is the confidence bound (3.8). Therefore, $\text{pr}\{C_n(\zeta_n^r) \leq \nu, \Theta \subseteq \Theta_n^r\} \geq \text{pr}\{C_n(\vartheta_n) \leq \nu, \Theta \subseteq \Theta_n^r\}$. We also know that $\text{pr}\{C_n(\vartheta_n) \leq \nu, \Theta \subseteq \Theta_n^r\} = \text{pr}\{C_n(\vartheta_n) \leq \nu\} - \text{pr}\{C_n(\vartheta_n) \leq \nu, \Theta \not\subseteq \Theta_n^r\}$ by the law of total probability. In the limit, the first probability on the right-hand side is equal to $1 - \alpha_2$ by Proposition 3.2 and the second probability is at most α_1 by assumption. This allows us to arrive at our main result.

The proof sketch also provides some insight into why the naive plug-in estimator ν_n^p defined in eq. (3.4) may fail to yield a valid confidence interval. A crucial quantity is $\text{pr}(\Theta \subseteq \Theta_n^r)$, which is known under an appropriate choice of $\epsilon_n(\theta)$. The corresponding quantity for the plug-in estimator is $\text{pr}(\Theta \subseteq \Theta_n)$, which could be arbitrarily small. In Example 3.1, this probability is zero.

It remains to discuss how to construct a relaxed set Θ_n^r. Whenever Θ can be characterized by a set of moment inequalities, such that $h_j(\theta) = E\{m_j(\theta)\}$, the moment inequalities literature summarized in Section 3.1.3 could be used to construct Θ_n^r. A more conservative relaxed set could be constructed via an application of the intersection bound. Suppose the following assumption holds on the constraint functions:

Assumption 3.7. For all $\theta \in \Theta$ and $j = 1, \ldots, J$, $n^{1/2}\big\{h_{nj}(\theta) - h_j(\theta)\big\} \to \mathcal{N}\big\{0, \sigma_j^2(\theta)\big\}$ in distribution and $\sigma_{nj}^2(\theta)$ is a consistent estimator for $\sigma_j^2(\theta)$.

This fairly weak assumption means that $h_{nj}(\theta)$ is pointwise asymptotically normally distributed and that there is a consistent estimator for the variance. This assumption allows us to select

$$\epsilon_n(\theta) = Z_{\alpha_{1j}}\, \sigma_{nj}(\theta)\, n^{-1/2}$$

where $\alpha_1 = \alpha_{11} + \alpha_{12} + \ldots + \alpha_{1J}$. It is straightforward to show that this choice of $\epsilon_n(\theta)$ satisfies (3.10). We could shrink the size of Θ_n^r by assuming that $h_n(\theta) = \{h_{1,n}(\theta), \ldots, h_{nj}(\theta)\}^T$ converges pointwise to a multivariate Gaussian with covariance matrix Σ and consistent estimator Σ_n. This would allow us to construct Θ_n^r as an ellipsoid confidence region.

Remark 3.1. It remains to discuss how one would select α_1 and α_2. As a rule-of-thumb, we typically choose the midpoint $\alpha_1 = \alpha_2 = \alpha/2$. It is tempting to select $\alpha = \alpha_1 + \alpha_2$ and choose C_n as the largest confidence bound over all α_1 and α_2 satisfying this equality. This would mean that α_1 and α_2 are sample-dependent quantities and so Theorem 3.1 will not directly apply. However, we can reason heuristically that the best choice of α_1 and α_2 should lie at an interior point $\alpha_1 > 0$ and $\alpha_2 > 0$. For a fixed sample, as $\alpha_1 \to 0$ and $\alpha_2 \to \alpha$, $\Theta_n^r \to B$ and thus C_n

approaches the $100(1-\alpha)\%$ confidence interval over the unconstrained problem. As $\alpha_1 \to \alpha$ and $\alpha_2 \to 0$, $C_n \to -\infty$ and thus the confidence interval becomes arbitrarily wide.

Remark 3.2. So far we have focused on inference for the infimum, however, partial identification problems are often characterized by an identified set of the form $I = [\nu^l, \nu^u]$, where $\nu^l = \inf\{Q(\theta) \colon \theta \in \Theta\}$ and $\nu^u = \sup\{Q(\theta) \colon \theta \in \Theta\}$ Imbens and Manski (2004) and Chernozhukov, Lee, and Rosen (2013). Suppose Θ_n^r is chosen so that $\text{pr}(\Theta \subseteq \Theta_n^r) \geq 1 - \alpha_1/2$. Moreover, let $\nu_n^{r,l} = \inf\{Q_n(\theta) \colon \theta \in \Theta_n^r\}$ and $\nu_n^{r,u} = \sup\{Q_n(\theta) \colon \theta \in \Theta_n^r\}$ denote the optimal values and $\vartheta_n^{r,l}$ and $\vartheta_n^{r,u}$ denote the corresponding optimal solutions. The estimated interval can be written as $[\nu_n^{r,l}, \nu_n^{r,u}]$ and we can construct a confidence interval by combining the lower confidence bound for ν^l and the upper confidence bound for ν^u, so that

$$[\nu_n^{r,l} - Z_{\alpha_2/2}\sigma_n(\vartheta_n^{r,l})n^{-1/2}, \nu_n^{r,u} + Z_{\alpha_2/2}\sigma_n(\vartheta_n^{r,u})n^{-1/2}]$$

will cover I with probability at least $1 - \alpha_1 - \alpha_2$. This is the two-sided analogue of the one-sided confidence interval proposed in eq. (3.12) and Theorem 3.1.

3.3 Sensitivity analysis via a logistic model

3.3.1 Set-up

We now return to the motivating example of selection bias in population-based cohort studies briefly described in Section 3.1.2. Specifically, we generalize the sensitivity analysis proposed in Thompson and Arah (2014), who define a logistic model for the probability of sample selection and propose to select parameters based on domain knowledge, or enumerate a large number of possible parameters. This approach is challenging to implement in the presence of complicated selection mechanisms with many parameters. Plausible sets of parameters that introduce bias in estimates of interest may be overlooked. Therefore, we begin by framing Thompson and Arah (2014) as an optimization problem over a space of plausible parameters, and describe how relevant auxiliary information could be introduced to further restrict the parameter space and provide more informative bounds, such as survey response rates, population means and negative controls. An additional sensitivity analysis, Aronow and Lee (2013), is generalized in the Supplementary Material.

Consider an independent and identically distributed draw of size N from an infinite population. For concreteness, we can think of this finite draw as the set of individuals who are eligible to enter the sample. Let $S_i \in \{0, 1\}$ be a selection indicator for whether individual i enrols in the sample, where $S_i = 1$ indicates sample participation, and let the observed sample size be denoted by $n = \sum_{i=1}^{N} S_i$. For notational convenience, we assume $S_1 = \ldots = S_n = 1$ and $S_{n+1} = \ldots = S_N = 0$.

Within the observed sample, we observe a vector of variables related to sample selection $W_i \in \mathbb{R}^K$. As in Thompson and Arah (2014), we assume that the probability of sample selection

admits a logistic form,

$$e(W_i;\theta) = \text{pr}(S_i = 1 \mid W_i) = \frac{\exp(\theta_0 + \theta_1 W_{i1} + \ldots + \theta_k W_{iK})}{1 + \exp(\theta_0 + \theta_1 W_{i1} + \ldots + \theta_k W_{iK})}, \tag{3.13}$$

where $\theta = (\theta_0, \theta_1, \theta_2 \ldots, \theta_K)^T$. We further assume that the sample is generated by some true selection probabilities $e(W_i;\theta^*)$ parametrized by θ^*.

For illustration, suppose that our object of interest is the population mean of a random variable X_i. We can write the population mean, and corresponding sample mean, in terms of θ^* as

$$\beta(\theta^*) = E(X_i) = \frac{E\{X_i/e(W_i;\theta^*) \mid S_i = 1\}}{E\{1/e(W_i;\theta^*) \mid S_i = 1\}}, \quad \beta_n(\theta^*) = \frac{\sum_{i=1}^n X_i/e(W_i;\theta^*)}{\sum_{i=1}^n 1/e(W_i;\theta^*)}. \tag{3.14}$$

The expression in eq. (3.14) relies on $X_i \perp S_i \mid W_i$, which we will assume throughout.

Since we only observe those for whom $S_i = 1$, the true parameter θ^* cannot be estimated. Thompson and Arah (2014) propose to consider a space of plausible values for θ^* and identify a worst-case lower bound and worst-case upper bound for $\beta_n(\theta^*)$. Inference for $\beta(\theta^*)$ itself was not considered. Formally, we select a parameter space Θ in which we are confident that θ^* resides. We then take the infimum and supremum of $\beta_n(\theta)$ over the space Θ.

3.3.2 Sensitivity parameters

Since we have assumed a logistic form for the selection probabilities (3.13), we can select sensitivity parameters which have a natural interpretation in terms of odds ratios.

Without loss of generality, suppose each W_{ik} has mean zero and standard deviation one within the sample. We can then choose a parameter $\Lambda_1 \geq 1$ such that

$$\Lambda_1^{-1} \leq \exp(\theta_k) \leq \Lambda_1, \ k = 1, \ldots, K. \tag{3.15}$$

We can interpret Λ_1 as the change in the conditional odds of sample selection from a one standard deviation increase in W_{ik}, holding all else fixed. When $\Lambda_1 = 1$, sample selection is completely random. Of course, we could select sensitivity parameters Λ_{1k} on a variable-by-variable basis for $k = 1, \ldots, K$, although choosing a single $\Lambda_1 = \max_k \Lambda_{1k}$ simplifies the interpretation of the sensitivity analysis.

The intercept term θ_0 also needs to be bounded. We can choose two parameters $\Lambda_0^l, \Lambda_0^u \in (0, 1)$ such that

$$\Lambda_0^l \leq \exp(\theta_0) \leq \Lambda_0^u \tag{3.16}$$

which can be interpreted as the odds of sample selection among those for whom $W_{ik} = 0$ for all k.

Rearranging eq. (3.15) and (3.16) shows that the sensitivity parameters $(\Lambda_0^l, \Lambda_0^u, \Lambda_1)$ characterize a compact subset of $\mathbb{R}^{K+1}$:

$$\theta \in \Theta = [\log(\Lambda_0^l), \log(\Lambda_0^u)] \times [\log(1/\Lambda_1), \log(\Lambda_1)]^K \tag{3.17}$$

From here, we can define the estimand and estimator, respectively, for the worst-case lower bound of $\beta(\theta)$ as

$$\nu = \inf\{\beta(\theta) \colon \theta \in \Theta\}, \quad \nu_n = \inf\{\beta_n(\theta) \colon \theta \in \Theta\}.$$

We could of course estimate the worst-case upper bound for $\beta(\theta)$ by taking the supremum of $\beta_n(\theta)$ over Θ (see Remark 3.2). Naturally, we can also consider estimators other than sample means, such as ordinary least squares or two-stage least squares.

3.3.3 Auxiliary information constraints

We now introduce several common examples where there may be discordance between known population quantities and quantities implied by the inverse probability weights. In general, provided we can formulate the constraints as a statistical test with a known null distribution, they can be placed within our framework.

Example 3.2. Suppose we know the response rate for a survey-based sample $r = E\{e(W_i; \theta^*)\}$. It is straightforward to show that $E\{1/e(W_i; \theta^*) \mid S_i = 1\} = 1/r$. This means that the within-sample expectation of the true inverse selection probabilities is equal to the inverse response rate. We therefore only want to consider parameters θ which imply this inverse response rate. The corresponding constraint can be writen as

$$h_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(1/e(W_i;\theta) - 1/r) \leq Z_{\alpha_{1j}/2}\,\sigma_{nj}(\theta)/n^{1/2}, \tag{3.18}$$

where $\sigma_{nj}(\theta)$ is the sample standard deviation of $1/e(W_i;\theta)$.

Example 3.3. Suppose we know the population mean $E(W_{ik})$ of some $W_{ik} \in W_i$. The inverse probability weighted sample mean of W_{ik} should therefore equal this mean in expectation, since

$$\frac{E\{W_{ik}/e(W_i;\theta^*) \mid S_i = 1\}}{E\{1/e(W_i;\theta^*) \mid S_i = 1\}} = E(W_{ik}).$$

This is conceptually similar to the raking procedure in survey sampling (Deming and Stephan, 1940), which adjusts sampling weights to match known marginal totals. The covariate mean constraint can be written as

$$h_{nj}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{W_{ik} - E(W_{ik})\}/e(W_i;\theta) \leq Z_{\alpha_{1j}/2}\,\sigma_{nj}(\theta)/n^{1/2}, \tag{3.19}$$

where $\sigma_{nj}(\theta)$ is the sample standard deviation of $\{W_{ik} - E(W_{ik})\}/e(W_i;\theta)$.

Example 3.4. Suppose we are confident that higher values of W_{ik} are associated with an increased probability of sample selection. For example, W_{ik} could be years of education and we might know from comparisons with representative samples, such as the census, that better educated individuals are more likely to select into the sample, conditional on other selection variables, so that $\theta_j \geq 0$ *a priori*.

Example 3.5. Suppose we know that two variables W_{ik} and $W_{ik'}$ are uncorrelated in the population. The inverse probability weighted correlation between W_{ik} and $W_{ik'}$ should therefore be zero. For example, due to the independent assortment of chromosomes, biological sex and autosomal genetic variants should be independent in the population, however, Pirastu et al. (2021) demonstrate that there is significant correlation within UK Biobank. This constraint can be formulated in several ways, for example fixing the regression coefficient of W_{ik} on $W_{ik'}$ to be zero.

Examples 3.2, 3.3 and 3.5 are two-sided constraints such that we also want these inequalities to hold for $-h_{nj}(\theta)$.

Remark 3.3. In the population means setting, Miratrix, Wager, and Zubizarreta (2018) demonstrate how to place shape constraints on the weighted empirical distribution of the response. Their approach involves constructing the worst-case weighted distribution given the Aronow and Lee (2013) bounding assumptions (see Supplementary Material). This results in a set which contains the oracle weighted distribution with probability approaching one. Provided we have a valid test, we can implement shape constraints within our framework without the need to characterize the worst-case weighted distribution. In the simplest case, we might want a variable to follow a known distribution in the population. For example, the distribution of IQ scores should be normal with mean 100 and standard deviation 15, which is a stronger constraint than Example 3.3. This could be formulated as a Kolmogorov-Smirnov test and the relaxed constraint set could be constructed via the null distribution of that test.

3.4 Extension of Aronow and Lee (2013)

3.4.1 Extension to ratio estimators

Another sensitivity analysis that can be made more informative through the addition of auxiliary constraints is one proposed in Aronow and Lee (2013). This non-parametric sensitivity analysis computes bounds on an inverse probability weighted sample mean under the assumption that each weight is bounded between two known constants. To start, we provide a slight generalization of this sensitivity analysis to ratio estimators. Let the estimator be given by

$$\beta_n = \left\{ \sum_{i=1}^{n} f(T_i)/e(W_i) \right\} \Bigg/ \left\{ \sum_{i=1}^{n} g(T_i)/e(W_i) \right\} \tag{3.20}$$

where $T_i \in \mathcal{T} \subseteq \mathbb{R}^M$ and $f, g \colon \mathbb{R}^M \to \mathbb{R}$. Unlike Section 3.3, we now assume that the functional form of $e(W_i)$ is unknown and some W_i may be unmeasured. As with Aronow and Lee (2013), we assume that $e(W_i)$ lies between two user-specified constants $1 \leq a \leq e(W_i) \leq b < \infty$ with probability one.

To apply our theoretical results in Section 3.2, we need a set Θ of fixed dimension. A simple assumption is that $\mathcal{T}$ is discrete and finite, $\mathcal{T} = \{t_k \colon k = 1, 2, \ldots, K\}$. Under this assumption, we can define $\beta_n(\theta)$ equal to

$$\beta_n(\theta) = \left\{ \sum_{k=1}^{K} \theta_k f(t_k) p_n(t_k) \right\} \Bigg/ \left\{ \sum_{k=1}^{K} \theta_k g(t_k) p_n(t_k) \right\}$$
$$\beta(\theta) = \left\{ \sum_{k=1}^{K} \theta_k f(t_k) p(t_k) \right\} \Bigg/ \left\{ \sum_{k=1}^{K} \theta_k g(t_k) p(t_k) \right\}$$

where $\theta_k = E\{1/e(W_i;\theta) \mid T_i = t_k, S_i = 1\}$ and $p_n(\cdot)$, $p(\cdot)$ are (respectively) the sample and population probability measures. This results in $\Theta = [1/b, 1/a]^K$. From here, the infimum takes the usual form,

$$\nu_n = \inf\{\beta_n(\theta) \colon \theta \in \Theta\}, \quad \nu = \inf\{\beta(\theta) \colon \theta \in \Theta\}.$$

It remains to identify assumptions such that the conditions in Section 3.2 are satisfied.

Assumption 3.8. For all $\theta \in \Theta$, the denominators of $\beta(\theta)$ and $\beta_n(\theta)$ are non-zero with probability one.

Assumption 3.9. For all $t_k \in \mathcal{T}$, $f(t_k)/g(t_k) \neq \nu$.

Assumption 3.8 simply ensures that $\beta_n(\theta)$ and $\beta(\theta)$ are well-defined over Θ. Assumption 3.9 is more subtle but will be needed to ensure a unique solution. If this assumption violated for some k, then the infimum is identical for all values of θ_k, meaning that the solution is not unique. An example of a function violating this condition is the following:

$$\beta(\theta) = \frac{-\theta_1 - 7\theta_2 - 10\theta_3}{\theta_1 + \theta_2 + \theta_3}, \quad \Theta = [1, 2]^3$$

In this example, $\vartheta = (1, 1, 2)$ and $\vartheta = (1, 2, 2)$ are both minimizers over Θ. The minimum value is -7 but $f(t_2)/g(t_2) = -7$ as well, which violates the condition. These assumptions jointly imply that ν has a unique minimizer. We begin with a technical lemma.

Lemma 3.1. *$\nu_n = \beta_n(\vartheta_n)$ is a global minimum over Θ if and only if, for all $k = 1, ..., K$, $q_k f(t_k) \leq \nu\, q_k\, g(t_k)$, where $q_k = \vartheta_k - (1/a + 1/b - \vartheta_k)$.*

This leads to our main proposition.

Proposition 3.4. *Under Assumptions 3.8 and 3.9, the set $\{\theta \in \Theta \colon \beta(\theta) = \nu\}$ is a singleton and, for n sufficiently large, the set $\{\theta \in \Theta \colon \beta_n(\theta) = \nu_n\}$ is a singleton with probability one.*

In fact, Proposition 3.4 provides an explicit form for the sample minimizer,

$$\vartheta_{nk} = \begin{cases} 1/b & \text{if } f(t_k)/g(t_k) \geq \nu_n \\ 1/a & \text{if } f(t_k)/g(t_k) < \nu_n \end{cases}, \tag{3.21}$$

and the population minimizer takes a similar form but with ν_n replaced with ν. Both Proposition 1 of Aronow and Lee (2013) and Section 4.4 of Zhao, Small, and Bhattacharya (2019) propose equivalent algorithms for computing the optimizing weights in the population means setting. Proposition 3.4 shows that we can generalize this algorithm to ratio estimators. In short, we can order $f(t_k)/g(t_k)$ from smallest to largest and evaluate $\beta_n(\theta)$ by enumerating over the weight at which $1/b$ changes to $1/a$, which has computational complexity $O(n)$.

Proposition 3.4 shows that Assumption 3.1 is satisfied for the generalized Aronow and Lee (2013) estimator under relatively weak conditions. Furthermore, Assumptions 3.2, 3.4 and 3.3 are satisfied under Assumption 3.8 and the assumption that $\mathcal{T}$ is finite. We therefore have that Proposition 3.2 is satisfied under these same relatively weak conditions.

3.4.2 Auxiliary information constraints

The constraints described in Examples 3.2 and 3.3 can be applied to this estimator. Let $c_n = Z_{\alpha_{1j}/2}\, n^{-1/2}$, then the response rate constraint can be formulated as

$$h_{nj}(\theta) = \left(1 + c_n^2\right)\left\{\sum_{k=1}^{K}(\theta_k - 1/r)p_n(t_k)\right\}^2 - c_n^2 \sum_{k=1}^{K}(\theta_k - 1/r)^2 p_n(t_k) \leq 0. \tag{3.22}$$

Suppose $w_k \in \mathbb{R}$ is an element of t_k, then the covariate mean constraint can be formulated as

$$h_{nj}(\theta) = \left(1 + c_n^2\right)\left\{\sum_{k=1}^{K}\theta_k(w_k - \bar{w})p_n(t_k)\right\}^2 - c_n^2 \sum_{k=1}^{K}\theta_k^2(w_k - \bar{w})^2 p_n(t_k) \leq 0. \tag{3.23}$$

Both of these constraints are quadratic in θ and can therefore be solved by existing algorithms for quadratically-constrained linear programs. Example 3.4 cannot be extended to this setting because it is tied to a parametric model. Example 3.5 can be extended to this setting in principle, but the resulting optimization problem is intractable.

Uniqueness of ϑ over Θ is needed to invoke Theorem 3.1. The population minimization problem over Θ is a linearly-constrained linear fractional programming problem. For example, the population response rate constraint is

$$h_j(\theta) = \sum_{k=1}^{K}(\theta_k - 1/r)p(t_k),$$

which is linear in θ. Since the level sets of $\beta(\theta)$ are also linear, to establish uniqueness of ϑ it suffices to assume (in addition to Assumptions 3.8 and 3.9) that the coefficient vectors of $h_j(\theta)$ and the level sets of $\beta(\theta)$ over Θ are not parallel.

3.5 Simulations

The aim of these simulations is to provide a brief assessment of the finite sample and limiting properties of the inference procedure described in Section 3.2. For concreteness, we simulate the

sensitivity analysis for selection bias described in Section 3.3. Our parameter $\beta(\theta)$ and estimator $\beta_n(\theta)$ are both the coefficient of a weighted linear regression. In particular, a regression of Y_i on X_i for $(X_i, Y_i) \sim \mathcal{N}(0, I_2)$, where I_2 is the identity matrix. The weights take the form in eq. (3.13) with variables $W_i = (X_i, Y_i)$.

We consider three distinct scenarios for the constraints. In the first scenario, we impose only sensitivity parameters $\Lambda_0^l = 0.11$, $\Lambda_0^u = 0.25$ and $\Lambda_1 = 3$. In the second scenario, we also impose a direction constraint $\theta_1 \geq 0$ as in Example 3.4. In the third scenario, we impose both the previous direction constraint and set the response rate equal to 0.15 as in Example 3.2. In each scenario, we use the discussion in Remark 3.2 to construct a two-sided 95% confidence interval for the identified set $I = [\nu^l, \nu^u]$, where $\nu^l = \inf\{\beta(\theta) \colon \theta \in \Theta\}$, $\nu^u = \sup\{\beta(\theta) \colon \theta \in \Theta\}$ and Θ takes the form in eq. (3.17). The first and second scenarios have no sample constraints and so the confidence interval corresponds to the one in eq. (3.8). The third scenario employs the confidence interval proposed in eq. (3.12) and Theorem 3.1.

Each scenario has distinct properties. In the first scenario, there are two solutions to the population optimization problems, thus violating Assumption 3.1. In the second scenario, the addition of a direction constraint rules out one of the two solutions and satisfies Assumption 3.1. In the third scenario, the introduction of a sample constraint necessitates the use of our relaxed confidence bound. In this scenario, we use our rule-of-thumb from Remark 3.1 to select $\alpha_1 = \alpha_2 = 0.025$ for both the upper and lower bounds of the two-sided confidence interval.

Table 3.1 summarizes the results and broadly aligns with our theoretical predictions. The first scenario violates Assumption 3.1 and the impact of this violation is substantial over-coverage of the confidence interval. Intuitively, this occurs because the sample solution will occur at, or near, the population solution that happens to minimize $\beta_n(\theta)$ in any given sample, which will result in a systematically wider confidence interval. The second scenario satisfies all assumptions for Proposition 3.2 and therefore converges to exact nominal coverage. The third scenario imposes a sample constraint and exhibits some over-coverage. This over-coverage can occur because our confidence bound in Theorem 3.1 sidesteps the covariance between the constraints $h_{nj}(\theta)$ and objective function $Q_n(\theta)$, instead imposing a worst-case intersection bound.

In this simulation exercise, the weight model is comprised of two variables. Section 2 of the Supplementary Material contains an additional simulation exploring the computation time of our R package `selectioninterval` as the number of variables in the weight model increases.

3.6 Applied example: effect of education on income

3.6.1 Description of the design

We implement an instrumental variable design for the effect of education on income in the UK Biobank cohort (Davies, Dickson, et al., 2018). We consider an instrumental variable design

Table 3.1: Coverage frequency for the three scenarios over 5000 Monte Carlo replications

	Sample size						
Scenario	10	25	50	100	200	500	1000
1	0.972	0.992	0.995	0.997	0.998	0.996	0.995
2	0.936	0.974	0.981	0.983	0.979	0.966	0.947
3	0.953	0.985	0.991	0.991	0.987	0.986	0.979

looking at the effect of education on income in the UK Biobank cohort. Our exposure is whether an individual remained in school at least until age 16 and our outcome is whether an individual earned more than £31,000 per year in 2006. Our instrument is based on a September 1972 education reform in England which raised the school leaving age from 15 to 16. Individuals who turned 15 just prior to the implementation of this reform were allowed to leave school, while individuals who turned 15 just after were required to remain in school until they were 16. This created a sharp discontinuity in the policies that the two groups were exposed to. Under the assumption that individuals on either side of the age threshold are otherwise identical, we can use this policy reform as an instrumental variable. We restrict our sample to individuals who turned 15 within 12 months of September 1972 and we control for sex and month-of-birth indicators. The unweighted estimate is 0.18 (95% confidence interval 0.08 - 0.28).

3.6.2 Results from naive sensitivity analysis

We first apply the sensitivity analysis described in Section 3.3.2 without auxiliary constraints, where the probability weights contain sex, years of education, income, age, days of physical activity per week and an interaction term between education and income. We choose sensitivity parameters $\Lambda_0^l = 0.02$, $\Lambda_0^u = 0.25$ and $\Lambda_1 = 2$, so that the average individual in the sample has an odds of sample selection between 0.02 and 0.25 and each variable in the model can induce a marginal odds of sample selection between 0.5 and 2. Our sensitivity analysis suggests that the effect estimate lies in the interval $[-1.34, 0.94]$ (95% confidence interval $-1.84, 1.29]$). This interval is completely uninformative as it spans the full range of possible estimates.

One explanation for this conservativeness is that this simple sensitivity analysis does not utilize all of the information available to us on the target population and the sample selection mechanism. The minimizing (maximizing) weights corresponding to this interval imply that the proportion of males in the population is 38.52% (46.6%) and the proportion of households with a gross income greater than £31000 is 95.84% (95.66%), all of which are inconsistent with known characteristics of the UK population.

3.6.3 Results from constrained sensitivity analysis

To address this incongruence, we consider four constraints that are typical of the information available to applied researchers using datasets such as UK Biobank. The first constraint is the response rate of UK Biobank (5.5%), which is the proportion of individuals who entered the cohort after receiving an invitation. The second constraint is the proportion of males in the UK population within the UK Biobank age range of 40-69 (49.5%). The third is the proportion of UK households earning more than £31000 per year at the date of UK Biobank recruitment in 2006 (21%). The fourth is the average age of individuals within our 2 year age bracket (48.98). All statistics were obtained from publicly available records from the UK's Office of National Statistics.

Figure 3.1 shows the resulting estimated intervals (thicker lines) and their corresponding confidence intervals (thinner lines) where each constraint is added sequentially. The estimated intervals and confidence intervals correspond to those described in Remark 3.2. We can see that each additional constraint reduces the width of the interval, with constraints 3 and 4 (the household income and age constraints respectively) seemingly having the largest marginal impact. The top interval includes all constraints and is quite informative for the desired effect, rejecting the null and suggesting an effect estimate in the range 0.08 - 0.22 (95% confidence interval 0.04 - 0.38). The unweighted estimate still lies within this interval, but our sensitivity analysis suggests some increased uncertainty in the range of effect estimates. These results also suggest that, despite the potential conservativeness of the confidence interval in Theorem 3.1, it can still produce informative bounds in practice.

3.7 Applied example: risk factors for COVID-19

We demonstrate how the sensitivity analysis in Section 3.3 can be applied to estimate risk factors for a positive COVID-19 test in a non-random sub-sample of UK Biobank participants. The data come from the June 2020 wave of the UK Biobank's Coronavirus Infection Study. Volunteers were tested for SARS-CoV-2 antibodies, indicating historical infection, using a lateral flow device. At the time, there was an effort to identify characteristics, such as age and body mass index (BMI), that predispose individuals to infection (Williamson et al., 2020).

A straightforward approach is to measure the correlation between each characteristic, known as a risk factor, and the presence of SARS-CoV-2 antibodies. A potential concern with this approach is that there is self-selection into the study, and individuals with known (or suspected) infection, and those with certain risk factors, are more likely to participate. This could induce collider bias when we condition on participation. This concern is exacerbated when looking at age-related risk factors, since initial UK Biobank recruitment was between 2006 and 2010 and many older participants will have died between then and 2020.

We have access to measures of age, sex and BMI in the full UK Biobank sample that

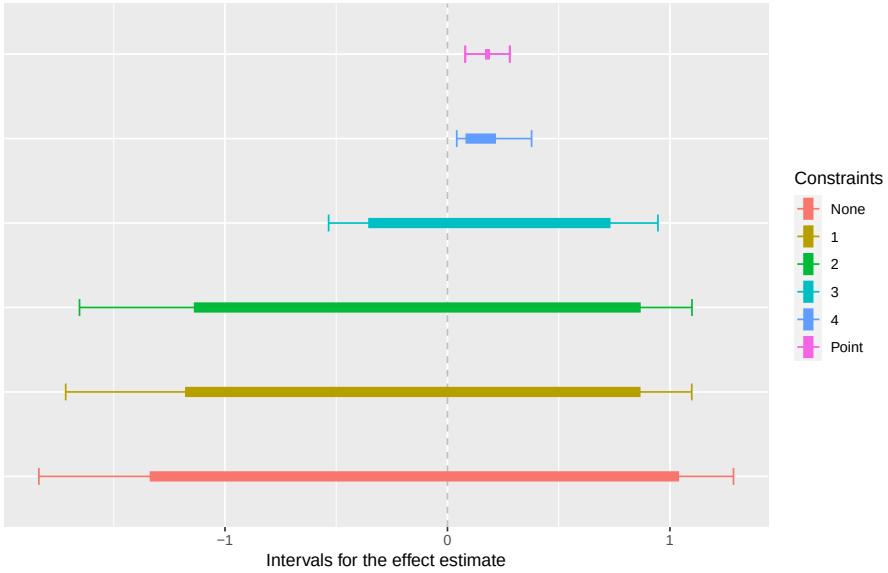


Figure 3.1: Estimated intervals (thick lines) and corresponding confidence intervals (thin lines) for effect estimates in the applied example. 'Point' represents the unweighted point estimate. Each constraint is added sequentially. No constraint means that only the sensitivity parameters $\Lambda_0^l = 0.02$, $\Lambda_0^u = 0.25$ and $\Lambda_1 = 2$ are imposed. Constraint 1 sets the response rate equal to 5.5%. Constraint 2 sets the proportion of males in the population to be 49.5%. Constraint 3 sets the proportion of households earning more than £31000 to be 21%. Constraint 4 sets the average age of individuals to be 48.98 years.

could be used to estimate inverse probability weights. However, one of the most informative characteristics determining selection is COVID-19 infection itself, which we cannot observe in the full sample. We construct intervals containing this missing variable in the weight model and compare them with unweighted estimates and estimates weighted using only variables in the full UK Biobank sample. We also compare the interval without constraints to the interval with several constraints: response rate of the Coronavirus Infection Study and means of age, sex and BMI in the full UK Biobank sample.

Figure 3.2 shows the estimates for two risk factors: age and BMI. The unweighted and naively weighted estimates are similar for both. The interval estimates without constraints are too wide to draw meaningful inferences about the size of the association, however, the estimates with constraints are more informative. For age, the unweighted estimate lies toward the lower end of the interval, although still within the 95% confidence interval. This interval suggests that age could be positively associated with COVID-19 infection, which is plausible given the time frame of the study (Griffith et al., 2020). For BMI, the unweighted estimate also lies within the interval. The corresponding 95% confidence interval does not contain the null, and we have reasonable confidence that BMI is associated with a heightened risk of COVID-19 infection.

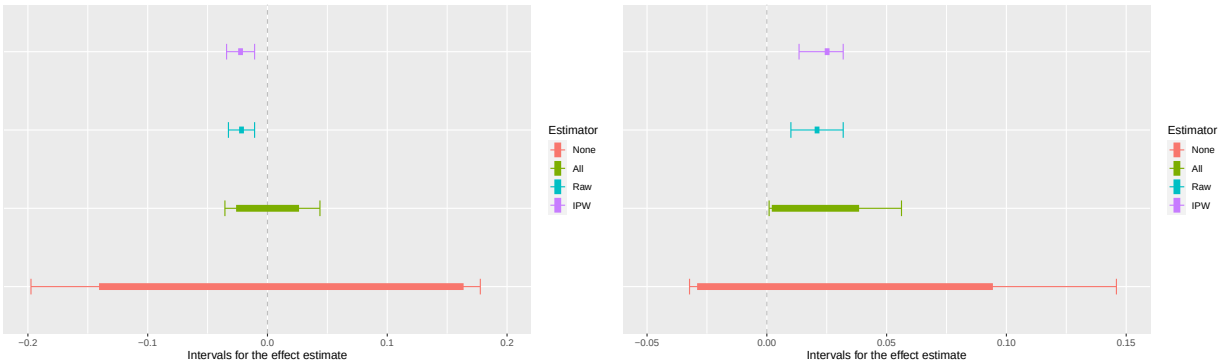


Figure 3.2: Estimated intervals (thick lines) and corresponding confidence intervals (thin lines) for the least squares estimate of age (left) and BMI (right) on Covid-19 test result. 'Raw' represents the unweighted point estimate. 'IPW' represents the estimate using weights estimated with covariates observed in the full UK Biobank cohort. 'None' represents the interval with no constraints. 'All' represents the interval with all constraints. The constraints set the response rate to its estimated value and the mean of age, sex and BMI to their full cohort values.

3.8 Discussion

There has been some existing work on bootstrap inference for Rosenbaum-type sensitivity analyses (Zhao, Small, and Bhattacharya, 2019). This approach considers a fixed parameter space Θ. It is unclear how to select the relaxation parameter $\epsilon_n(\theta)$ in a bootstrap analogue of our method under estimated constraints. Simple approaches, such as constructing Θ_n^r via asymptotic approximations and then bootstrapping the distribution of ν_n^r, are plausible but their statistical properties remain to be explored.

In some instances, including our selection bias application in Section 3.3, the target of inference is $Q(\theta^*)$, where θ^* is some true parameter lying within Θ, rather than ν. Suppose Θ is known and we have a two-sided identified set $[\inf_{\theta\in\Theta} Q(\theta), \sup_{\theta\in\Theta} Q(\theta)]$ as in Remark 3.2, then if $Q(\theta^*)$ lies near the boundary of this set (and the set has positive width), the non-coverage probability of the corresponding confidence interval is effectively one-sided in the limit. A naive two-sided confidence interval constructed around the identified set may be too conservative. Imbens and Manski (2004) discuss approaches for maintaining uniform coverage of $Q(\theta^*)$. The central limit theorem established by Shapiro (1991) for known Θ is amenable to their framework (although, to our knowledge, has not been formally used in this setting); extending this result to sample-constrained problems would be a valuable contribution. Stoye (2009) further extends Imbens and Manski (2004) by developing confidence intervals that exhibit uniform coverage for $Q(\theta^*)$ without relying on assumed superefficiency of the estimated interval width.

A final consideration is the computational burden of our approach. Our general inference procedure in Section 3.2 relies on the optimization problems in eq. (3.11) being solvable, but the computational complexity of these problems will vary depending on the application. Our R package `selectioninterval` relies on out-of-the-box global and local optimization algorithms.

There are no theoretical guarantees of convergence to the global optimum, however, we have not observed a failure of convergence in our simulations.

Chapter 4

Mendelian randomization with coarsened exposures

Publication arising from this chapter

This chapter is published in the journal *Genetic Epidemiology* under the title "Mendelian randomization with coarsened exposures" (Tudball, Bowden, et al., 2021). My contribution was: conceptualization; deriving the technical results; designing the method; designing and analyzing the simulation study and applied example; writing the manuscript and revising in response to co-author and reviewer comments. Co-authors' contribution was: assistance with formalizing the assumptions behind the method; suggestions for the design, analysis and interpretation of the applied example; providing feedback on the manuscript and reviewer response.

SIGNED: Matthew Tudball (First Author) DATE: 24 July, 2022
SIGNED: George Davey Smith (Senior Author) DATE: 24 July, 2022

4.1 Introduction

4.1.1 Motivation

Mendelian randomization proposes to use genetic variants that alter, or mirror the biological effects of, modifiable exposures to study the causal effects of such exposures on downstream outcomes. The principle underlying Mendelian randomization is that genetic variants are randomly passed from parents to offspring at conception, resulting in a plausibly unconfounded source of variation in the exposures with which they are associated. For Mendelian randomization estimates to inform policies or clinical practices, we must additionally assume that genetic and environmental modifiers of the exposure produce similar effects on the outcome (Davey

Smith and Ebrahim, 2003). For example, Mendelian randomization studies of pharmaceutical exposures typically use genetic variants that code for potential drug targets, assuming that similar effects would be observed if those targets were altered therapeutically (Plump and Davey Smith, 2019).

One of the crucial assumptions underlying the Mendelian randomization approach is that the relationship between the genetic instruments and the outcome is fully mediated by the exposure, known as the exclusion restriction assumption. However, it is important to draw a distinction between the true exposure experienced by an individual and our attempt at measuring it. For practical purposes, we are often restricted to coarsened approximations which do not fully encapsulate the mechanism by which the true exposure of interest affects the outcome. Consistent with existing terminology, we define an exposure measurement as coarsened if it is a discrete measure approximating a continuous latent exposure (Marshall, 2016).

In the Mendelian randomization context, coarsened exposures can violate the exclusion restriction assumption. If the genetic instruments are acting on a latent exposure, such as body mass index (BMI), but the measured exposure is a discretization of it, such as obesity status, then there can exist genetically-driven variation in the true exposure within categories of the measured exposure. We could imagine that counterfactually altering some BMI-raising single nucleotide polymorphism (SNP) in an individual could result in a change in their BMI without necessarily changing their obesity status. This can be viewed a form of measurement error which opens up potential pathways from the genetic instruments to the outcome that do not pass through the exposure measure, thus violating the exclusion restriction assumption.

For example, Richardson, Sanderson, et al. (2020) attempt to separate the effects of early and later life adiposity on disease risk. The adiposity variable is a three-category self-report measure ('thinner', 'plumper' and 'about average'). It is reasonable to conceptualize a continuous measure of body mass (e.g., BMI) underlying this coarsened categorical measure, such that genetic variation in this latent continuous measure could occur within categories of the self-report variable. We later re-analyse Richardson, Sanderson, et al. (2020) in Box 4.5.3 using the approach proposed in this paper. Another example is Richmond et al. (2019), who apply Mendelian randomization to investigate the effect of sleep traits (e.g., morning preference, sleep duration) on breast cancer risk, finding large causal effects of several traits. These traits are categorical measures, for example, morning preference is measured in six categories and sleep duration is split into several groups. It is reasonable to conceptualize the true exposures on which the genetic variants are acting as latent continuous sleep traits and preferences, for which the measured exposures are discrete markers.

An important class of latent exposures we consider in this paper is disease liabilities, for which binary disease diagnosis or case status is the typical exposure measurement. There are an increasing number of Mendelian randomization studies investigating the effects of complex diseases such as asthma, schizophrenia and attention deficit hyperactivity disorder on various

outcomes (Lawn et al., 2019; Martins-Silva et al., 2019; Pasman et al., 2018; Sun et al., 2019). Complex diseases which result from the interaction of environment and multiple genetic variants are likely to affect outcomes of interest through pathways other than diagnosis, for example, severity of sub-clinical symptoms. Since genetic instruments are, in turn, likely to influence the manifestation or severity of the underlying symptoms, rather than diagnosis alone, this represents a potential violation of the exclusion restriction.

4.1.2 Previous literature

This specific violation of the exclusion restriction assumption has been raised before in both the economics and political science literatures (Angrist and Imbens, 1995; Marshall, 2016). It has also been raised briefly in the Mendelian randomization context in Burgess and Labrecque (2018), who discuss interpretation of estimates with binary exposures. The authors recommend that findings be framed in terms of this latent exposure but note that the estimates themselves have no meaningful causal interpretation. The authors do not describe how this bias may distort estimates nor clarify how to appropriately frame estimates in terms of the latent exposure, which will depend on the unobservable relationship between the latent exposure and its coarsened measurement.

There is a rich literature in econometrics on instrumental variables for mismeasured or latent exposures (Spady, 2007; Song, Schennach, and White, 2015). See Schennach (2020) for an in-depth review. Hu and Schennach (2008) is the most closely related to the problem in this chapter. The authors consider identification of a class of non-classical, non-linear errors-in-variables models under relatively weak assumptions on the conditional distributions of the instrument, measured exposure and latent exposure. However, Hu and Schennach (2008) cannot be directly applied in this setting because they must assume that these three variables are jointly continuously distributed, whereas we are concerned with discrete exposure measures.

The method described in this chapter is distinguished from these approaches by imposing a stronger set of assumptions grounded in population genetic theory. For example, our assumption that each SNP has an independent, additive effect on the exposure is empirically justified (Heyne et al., 2023) and our linear single index assumption is well-studied (Falconer, 1965; Curnow, 1972). We shall describe and justify these assumptions in further detail later in the chapter.

4.1.3 Our contribution

Our main contributions are to derive an expression for the bias described in Section 4.1.1 and introduce a clear set of identifying assumptions under which one can estimate the causal effect of the latent exposure. We hope to allow researchers to decide whether these assumptions are plausible in the context of their study. In Section 4.2, we outline our technical framework, which assumes a linear single threshold model for the relationship between the latent exposure and

its measurement. That is, we assume that values of the coarsened exposure are determined by whether the latent exposure is above or below some threshold, which could be individual-specific. For example, an individual is classified as obese if their BMI is above 30 and not obese otherwise. This framework also contains the Falconer (1965) liability-threshold model, which assumes that a disease occurs in an individual, or is sufficiently pronounced to be diagnosed, if a build-up of underlying liability crosses some threshold. In this model, liability is assumed to capture all genetic, shared and non-shared environmental risk factors.

In Section 4.3.1, we derive an expression for the bias from the naive approach of using the coarsened measure as the exposure directly. Then, in Section 4.3.2, we show that, if the latent exposure is standardized to have a standard deviation of one, its causal effect can be identified if we have auxiliary information on the genetic variance of the latent exposure. This may be obtained from GWAS or treated as a sensitivity parameter and varied over a plausible range of values. In the context of disease liabilities, we may use the coefficient of determination developed by Lee, Goddard, et al. (2012).

Section 4.4 provides some generalizations to this framework, in particular, allowing two-sample estimation. Section 4.5 provides a real data example by creating artificially dichotomized variables from the continuous BMI measure in UK Biobank. Boxes 4.5.2 and 4.5.3 present reanalyses of two papers which could be interpreted within the framework proposed in this paper (Pasman et al., 2018; Richardson, Sanderson, et al., 2020). In Sections B.1 and B.2 of the Appendix, we examine the bias that can emerge when the assumptions of our framework are violated.

4.2 Framework

We begin by outlining some key notation. Suppose there is a genetic instrument $Z \in \mathbb{R}$, other genetic variants (e.g., pleiotropic, weak) $X \in \mathbb{R}^K$ and an environmental risk factor $V \in \mathbb{R}$, where V is assumed to be continuously distributed with mean zero. We also assume that Z, X and V are mutually independent. We define $G = \mu + \alpha Z + \gamma' X$ as the genetic share of the latent exposure and define the latent exposure itself as

$$\begin{aligned} L &= G - V \\ &= \mu + \alpha Z + \gamma' X - V. \end{aligned} \tag{4.1}$$

It would be equally correct to define $L = G + V$, but the formulation in (4.1) simplifies some later expressions. In the Falconer framework described in Section 4.1, L would represent liability to some disease. We are able to observe a coarsened exposure characterized by a dichotomization of the latent exposure.

$$D = \begin{cases} 1 & \text{if } L \geq 0 \\ 0 & \text{if } L < 0 \end{cases}. \tag{4.2}$$

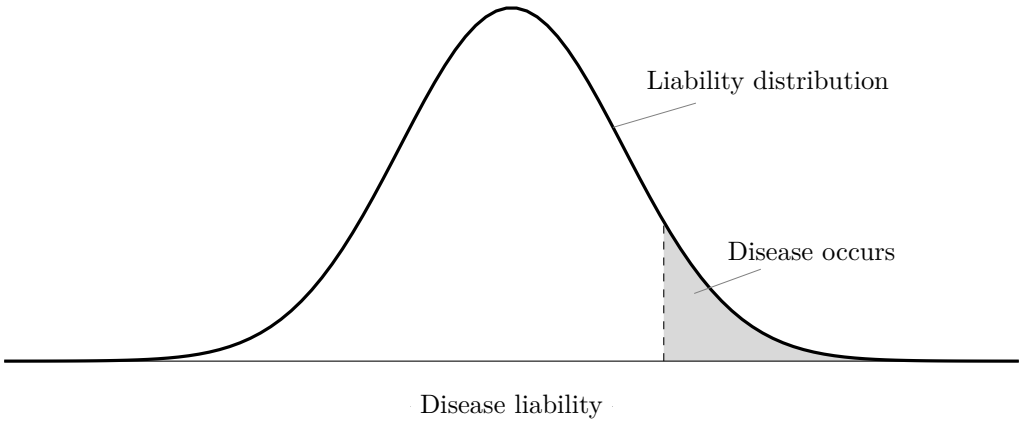


Figure 4.1: In the Falconer framework, liability to a disease is assumed to follow a smooth (often normal) distribution. The disease occurs at the tail of the distribution, with the grey region representing prevalence in the population.

If L is disease liability, then D would represent occurrence of the disease. In practice, we measure diagnosis of the disease, which does not necessarily correspond to occurrence due to under- or over-diagnosis. We will treat the two as equivalent throughout and discuss violations of this equivalence in Section 4.6.

Equation (4.2) is the crucial assumption underlying our approach; namely, that L is a linear index that relates to D according to a single threshold. Section B.1 of the Appendix elaborates on the importance of this structural assumption. Figure 4.1 illustrates our model within the Falconer framework. There is a distribution of disease liabilities and the disease occurs at the right tail of this distribution. The size of the grey region represents the prevalence of the disease in the population.

We also have an observed outcome $Y \in \mathbb{R}$. We restrict ourselves to a linear structural equation model

$$Y = \beta L + \varepsilon \tag{4.3}$$

which is implicitly conditional on covariates, where ε can be correlated with both V and X. We make the standard instrumental variable assumptions, namely, that $\alpha \neq 0$ and Z is independent of ε conditional on covariates. The model (4.3) implicitly captures the assumption described in Section 4.1 that genetic and environmental modifiers of the exposure produce equivalent effects on the outcome. In this setting, the marginal effect (in absolute value) of both G and V is β. Figure 4.2 summarizes this model in a directed acyclic graph. We can see that the exclusion restriction is violated since there exists a path from the latent exposure L to Y which does not pass through the measured exposure D. The structural equation (4.3) assumes no effect of D itself. For a disease such as schizophrenia, liability could have a harmful effect on the outcome but being diagnosed will usually lead to receiving treatment and thus could have a protective effect. We cannot separately identify the two effects in this setting, although possibilities for

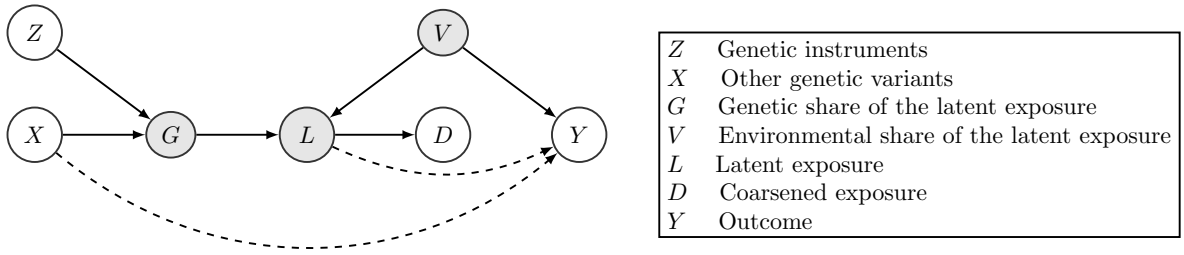


Figure 4.2: The framework proposed in Section 4.2 summarized in a directed acyclic graph. Dotted circles represent latent variables and complete circles represent observed variables.

doing so are discussed in Section 4.4.2. When D is believed to have a distinct effect on the outcome, we may instead identify the total effect of liability on the outcome; i.e., the direct effect β and the indirect effect through D.

The structural assumptions made in this section can be summarized as follows:

Assumption 4.1. (Single threshold) The latent exposure L and its binary measurement D are related by a single threshold model of the form $D = I\{L \geq 0\}$.

Assumption 4.2. (Additivity) $L = G - V$, where G and V are, respectively, the genetic and environmental shares of L.

Assumption 4.3. (Linearity) G is a linear function of the genetic instrument Z and other genetic variants X, such that $G = \mu + \alpha Z + \gamma' X$.

Assumption 4.4. (Environmental share) V has mean zero, standard deviation σ_V and is in some family of continuous distributions, with cumulative distribution function given by $F(v/\sigma_V) = F_V(v)$ and density $f(v/\sigma_V) = f_V(v)$.

Assumption 4.5. (Risk factor independence) Z, X and V are mutually independent.

Assumption 4.6. (Gene-environment equivalence) The outcome model takes the form $Y = \beta L + \varepsilon$, where ε is a random disturbance and X and V may be correlated with ε.

Assumption 4.7. (Instrumental variable assumptions) Z is independent of ε and $\alpha \neq 0$.

4.3 Identification

4.3.1 Bias from the naive approach

The naive approach to Mendelian randomization is to use the coarsened exposure D as the exposure directly. We show in Proposition 4.1 that this results in a 'multiplicative' bias which will scale the true effect β up or down, but not change its direction. When the distribution of

L has a light tail (e.g., normal distribution), we will typically see inflation of effect estimates, with the degree of inflation increasing as the prevalence of D becomes smaller. If D is case status for a disease, for example, then effect estimates will be more inflated for rarer diseases. We see this pattern of inflation occurring in our real data examples in Section 4.5.

Proposition 4.1. *Consider a binary instrument $Z \in \{0, 1\}$ and exposure $G = \mu + \alpha Z$. Under Assumptions 4.1-4.6, the Wald estimand takes the form*

$$\beta_D = \operatorname{cov}(Z, Y)/\operatorname{cov}(Z, D) = \beta/f_V(\mu^*),$$

where $\mu \leq \mu^* \leq \mu + \alpha$.

Proof. We start by noting that

$$\begin{aligned}
\operatorname{cov}(Z, D)/\operatorname{var}(Z) &= \operatorname{pr}(D = 1 \mid Z = 1) - \operatorname{pr}(D = 1 \mid Z = 0) \\
&= \operatorname{pr}(L \geq 0 \mid Z = 1) - \operatorname{pr}(L \geq 0 \mid Z = 0) && \text{(Assumption 4.1)} \\
&= \operatorname{pr}(V \leq \mu + \alpha Z \mid Z = 1) - \operatorname{pr}(V \leq \mu + \alpha Z \mid Z = 0) && \text{(Assumptions 4.2, 4.3 and 4.4)} \\
&= F_V(\mu + \alpha) - F_V(\mu) && \text{(Assumption 4.5)} \\
&= \alpha f_V(\mu^*) && \text{(Assumption 4.4)}
\end{aligned}$$

by the mean value theorem, where $\mu \leq \mu^* \leq \mu + \alpha$.

It follows from Assumption 4.6 that $\operatorname{cov}(Z, Y)/\operatorname{var}(Z) = \alpha\beta$. Thus,

$$\operatorname{cov}(Z, Y)/\operatorname{cov}(Z, D) = \beta/f_V(\mu^*).$$

□

The interpretation of Proposition 4.1 is that β_D is equal to the true latent exposure effect β divided by the density of V at some value μ^*. This quantity $f_V(\mu^*)$ is not identified since the distribution of V is unknown and μ^* is defined on the scale of the latent exposure.

4.3.2 The latent variable approach

The bias formula in Proposition 4.1 indicates that the nuisance term is $f_V(\cdot)$, which is the distribution of the environmental share V. We do not need to know $f_V(\cdot)$ itself to achieve the identification result in this section, but it must lie within a known family of distributions.

Assumption 4.8. The family of distributions F in Assumption 4.4 is known.

Lemma 4.1. *Under Assumptions 4.1-4.5 and 4.8, G/σ_V is identifiable from the observed data (Z, X, D).*

Proof. We can write the distribution of D given Z and X as

$$\begin{aligned}
\text{pr}(D = 1 \mid X, Z) &= \text{pr}(L \geq 0 \mid X, Z) && \text{(Assumption 4.1)} \\
&= \text{pr}(V \leq \mu + \alpha Z + \gamma' X \mid X, Z) && \text{(Assumptions 4.2, 4.3 and 4.4)} \\
&= F\{(\mu + \alpha Z + \gamma' X)/\sigma_V\} && \text{(Assumption 4.5 and 4.8)} \\
&= F\{G/\sigma_V\}
\end{aligned}$$

Thus,

$$G/\sigma_V = F^{-1}\{\text{pr}(D = 1 \mid X, Z)\}.$$

□

Since F is known and (Z, X, D) is observed, G/σ_V can be identified via an appropriate generalized linear regression.

Remark 4.1. In practice, we could specify F directly, for example, as a logistic or normal distribution (corresponding to logistic and probit regressions respectively). Alternatively, to avoid imposing potentially strong distributional assumptions, we could use semi-parametric estimation methods for generalized linear models, which only require some smoothness conditions on F (Ichimura, 1993; Klein and Spady, 1993). Disease liabilities are often assumed be the product of many small, independent traits. Therefore, by the central limit theorem, a normal distribution (i.e., probit model) is a natural choice of link function in this context (Curnow, 1972).

Corollary 4.1. *G/σ_G is identifiable from the observed data (Z, X, D).*

Proof. We begin by noting that

$$\text{var}(G/\sigma_V) = \sigma_G^2/\sigma_V^2$$

is identified and G/σ_V is identified by Lemma 4.1. Thus,

$$(G/\sigma_V)/(\sigma_G/\sigma_V) = G/\sigma_G$$

is identified. □

We are now prepared to introduce our sensitivity parameter.

Definition 4.1. Let $\theta^2 = \sigma_G^2/\sigma_L^2$ be defined as the genetic variance of the latent exposure.

Suppose we have an appropriate choice of θ^2 from external domain knowledge. We can therefore identify the effect of standard deviation increases in the latent exposure L.

Theorem 4.1. *Under Lemma 4.1, Corollary 4.1 and an appropriate choice of the sensitivity parameter θ^2, we can identify*

$$\beta_L = \sigma_L \beta.$$

Proof.

$$\begin{aligned}\operatorname{cov}(Z,Y)/\{\theta \operatorname{cov}(Z, G/\sigma_G)\} &= (\sigma_G/\theta)\,\beta \\ &= (\sigma_L\,\sigma_G/\sigma_G)\,\beta \\ &= \sigma_L\beta.\end{aligned}$$

□

Our latent variable approach can be viewed as proceeding in four steps: 1) estimate the linear predictor of a generalized linear model of D on Z and X; 2) normalize the linear predictor to have mean zero and variance one; 3) use this normalized linear predictor as the exposure in the two stage least squares estimator; and 4) scale the resulting effect estimate up by the genetic variance of the latent exposure θ^2.

The parameter θ^2 can be treated as a sensitivity parameter and varied over a plausible range of values or can, in some instances, be obtained from GWAS which report this measure.

For disease liabilities in particular, Lee, Goddard, et al. (2012) uses the Falconer liability-threshold model to develop a coefficient of determination for GWAS that is interpretable on the liability scale, which corresponds to θ^2. Therefore, θ^2 can be estimated using this approach or selected from GWAS which report this coefficient. For ease of interpretation, liability is often assumed to have mean zero and variance one, in which case $\sigma_L = 1$ and β itself is identified on this scale (Lee, Goddard, et al., 2012).

4.4 Some generalizations

4.4.1 Individual-specific threshold

The formalization of the relationship between disease and liability in equation (4.2) and Figure 4.1 assumes a fixed threshold. That is, all individuals with liability above the threshold will develop or be diagnosed with the disease and all those below the threshold will not. In reality, we might imagine that diagnosis has a random component, driven, for example, by preferences of the diagnosing clinician or imprecision of the testing procedure. It might be more realistic to assume a model such that

$$D = \begin{cases} 1 & \text{if } L \geq R \\ 0 & \text{if } L < R \end{cases} \tag{4.4}$$

where R is a random individual-specific threshold. Provided R is independent of the instrument Z and other variants X, this random threshold will not affect identification of G/σ_V of Lemma 4.1 under correct model specification. However, the link function F of Assumption 4.8 no longer corresponds to the distribution family of V; instead, it corresponds to the distribution

family of $V + R$. This could make correct specification of the link function more difficult and semi-parametric approaches may be warranted.

4.4.2 Identifying effects of the coarsened exposure

The structural model (4.3) assumes no direct effect of the binary exposure measure D on the outcome. As discussed in Section 4.3, when D is diagnosis of a disease, we might expect resulting treatment or therapy to have an effect on the outcome distinct from disease liability, suggesting a structural equation model of the form

$$Y = \beta L + \delta D + \varepsilon. \tag{4.5}$$

The exposure measure is downstream of the latent exposure and there are assumed to be no direct pathways from the genetic instruments to the exposure measure, as illustrated in Figure 4.2. Therefore, we cannot use our genetic instrument Z to estimate the independent effect of the exposure measure on the outcome; the genetic instruments induce no unique variation in the exposure measure independent of the latent exposure. However, consider the individual-specific threshold of Section 4.4.1. The variable R could represent preferences of the clinician for diagnosing the disease or a change in clinical practices affecting some individuals (Brookhart and Schneeweiss, 2007; Davies, Gunnell, et al., 2013). If R is independent of each individual's liability, without directly affecting the outcome, then it is a potential instrument for disease diagnosis. The general rule for separately estimating the effects of the latent exposure and dichotomization is to have instruments which induce distinct variation in both.

4.4.3 Multi-valued discrete exposure

This method generalizes easily to the multi-valued discrete exposure setting. Suppose we observe a discretized variable characterized by

$$D = \begin{cases} 0 & \text{if } L \leq 0 \\ 1 & \text{if } 0 < L \leq d_1 \\ \vdots & \\ K & \text{if } d_{K-1} < L \end{cases} \tag{4.6}$$

where $0 < d_1 < ... < d_{K-1}$ are latent thresholds. D could represent number of years in education and L could represent time in education as a continuous measure. Similar to how the dichotomous exposure can be formulated as a binary response model as in Lemma 4.1, exposures of the form (4.6) can be formulated as an ordered response model and the parameters $\tilde{\mu}$, $\tilde{\alpha}$ and $\tilde{\gamma}$ are still identified, allowing the method to be applied as usual.

4.4.4 Two-sample design with GWAS summary statistics

For rare diseases, it is not always possible to observe the coarse exposure measurement D and the outcome Y in the same sample. It is common practice in Mendelian randomization studies to use summary statistics from separate GWAS of the exposure and outcome to obtain two-sample estimates (Burgess, Scott, et al., 2015). This method also generalizes to the two-sample setting using the popular inverse-variance weighted approach (Burgess, Butterworth, and Thompson, 2013).

Suppose there is a set $\mathcal{Z}_J = \{Z_j : j = 1, ..., J\}$ of SNPs from the exposure GWAS, of which a subset $\mathcal{Z}_{J_0} = \{Z_j : j = 1, ..., J_0\}$, $J_0 \leq J$, is selected as instruments from the outcome GWAS. Suppose we have estimates $\hat{\tilde{\alpha}}_j$ on the log-odds scale of the instrument-exposure relationship $\tilde{\alpha}_j$ for each instrument in $\mathcal{Z}_J$ and estimates of the instrument-outcome relationship $\hat{\Gamma}_j$ for each instrument in $\mathcal{Z}_{J_0}$. Additionally, we need the variance $\sigma^2_{Z_j}$ for each instrument in $\mathcal{Z}_J$, which can be obtained from reported allele frequencies. Lastly, we also need estimates for the inverse-variance weights $w_j = \hat{\tilde{\alpha}}^2_j / \sigma^2_{\hat{\Gamma}_j}$, where $\sigma_{\hat{\Gamma}_j}$ is the standard error of $\hat{\Gamma}_j$. Under the assumption that the instruments in $\mathcal{Z}_J$ are mutually independent, the inverse-variance weighted estimator for $\beta_G = \operatorname{cov}(Z, Y) / \operatorname{cov}(Z, G/\sigma_G)$ can be obtained from the above summary statistics as

$$\left(\sum_{j=1}^{J} \hat{\tilde{\alpha}}^2_j \sigma^2_{Z_j} \right)^{1/2} \frac{\sum_{j=1}^{J_0} w_j \hat{\Gamma}_j / \hat{\tilde{\alpha}}_j}{\sum_{j=1}^{J_0} w_j} \tag{4.7}$$

which is derived in Section B.3 of the Appendix. We can recover the effect in terms of σ_L (i.e., β_L) by rescaling by a suitable choice of θ^2 as described in Section 4.3. Conveniently, the second term in (4.7) is the standard form of the inverse-variance weighted estimator. This means that we can easily readjust existing Mendelian randomization estimates of coarsened exposures using only the exposure GWAS and a choice for θ^2. The large-sample distribution of the estimator (4.7) is derived in Section B.3 of the Appendix.

4.5 Real data examples

4.5.1 Effect of BMI on systolic blood pressure

We can assess the performance of this method in a realistic setting by creating a dichotomized variable from an observed continuous measure, BMI. The idea is to dichotomize BMI at some threshold value and then treat only the dichotomization as observed. We shall compare the true standardized effect of BMI on some outcome with our procedure described in Section 4.3 and with the naive approach of using the dichotomization as the exposure.

Our example is based on the Mendelian randomization analysis performed in Lyall et al. (2017), which estimates the effect of BMI on several cardiometabolic measures in the UK Biobank cohort. In particular, we look at the effect of BMI on systolic blood pressure. This is a

convenient exposure-outcome relationship to estimate because we should not expect there to be threshold effects, i.e., the dichotomizations of BMI should have no distinct effects on systolic blood pressure except through BMI itself.

Consistent with Lyall et al. (2017), we use as potential instruments the 93 genome-wide significant SNPs reported in Locke et al. (2015) available in UK Biobank and we control for age, sex, assessment centre, alcohol intake, smoking status and Townsend deprivation index, along with genetic batch and the first 10 principal components of the genetic relatedness matrix. To avoid weak instrument bias, we prune these SNPs by including those which correlate with BMI with $|t| > 4$ (conditional on the other SNPs) as instruments. We estimate the 'true' standardized effect of BMI on systolic blood pressure via two-stage least squares, finding that a one standard deviation increase in BMI corresponds to an increase in systolic blood pressure of 1.53 mm Hg (95% CI 0.34 - 2.72). At each BMI threshold, we then generate a binary variable equal to 1 if an individual's BMI is above the threshold and 0 otherwise. Treating only this binary measure as observed, we apply the latent variable approach of Section 4.3.2 using a probit link function.

The results of this example are summarized in Figure 4.3, which compares the estimated effects with the 'true' effect of 1.53. The estimates using the dichotomized measure as the exposure are highly sensitive to the choice of threshold. Since we should not expect there to be distinct threshold effects in this setting, this demonstrates that the dichotomized exposure is not capturing the effect of the latent exposure, instead, it is picking up the shape of the distribution of the environmental risk factor for BMI, as discussed in Section 4.3.1. As predicted by the bias formula in Section 4.3.1, the estimates were inflated at the extreme thresholds where the distribution is flatter.

For the latent variable approach, we select a θ^2 of 0.0256 based on the R-squared of our first stage regression of BMI on the genetic variants. The effect estimate from this approach is much less sensitive to the choice of threshold. Furthermore, the estimates appear to accurately recover the 'true' effect of 1.53 regardless of the threshold value, ranging from 1.35 at a BMI cut-off of 30 to 1.92 at a BMI cut-off of 22.5.

We can also investigate this approach in a more realistic setting by re-analysing two existing papers. Box 4.5.2 gives an example of how existing two-sample results which do not have interpretable effect sizes can be reinterpreted using this method. The original paper finds that schizophrenia liability increases one's likelihood of using cannabis, although the effect sizes are not interpretable (Pasman et al., 2018). Using our approach, we find that a one standard deviation increase in liability corresponds to an odds ratio in the range 1.15-1.26 (95% CI 1.10-1.44) for ever using cannabis. This approach allows us to infer the size of this effect which, in this instance, is very modest.

Box 4.5.3 gives an example of how this approach can correct exclusion restriction violations. In the original paper, self-reported adiposity is measured on a three-point scale ('thinner', 'plumper' and 'about average'). Genetic instruments will be acting on the underlying measure

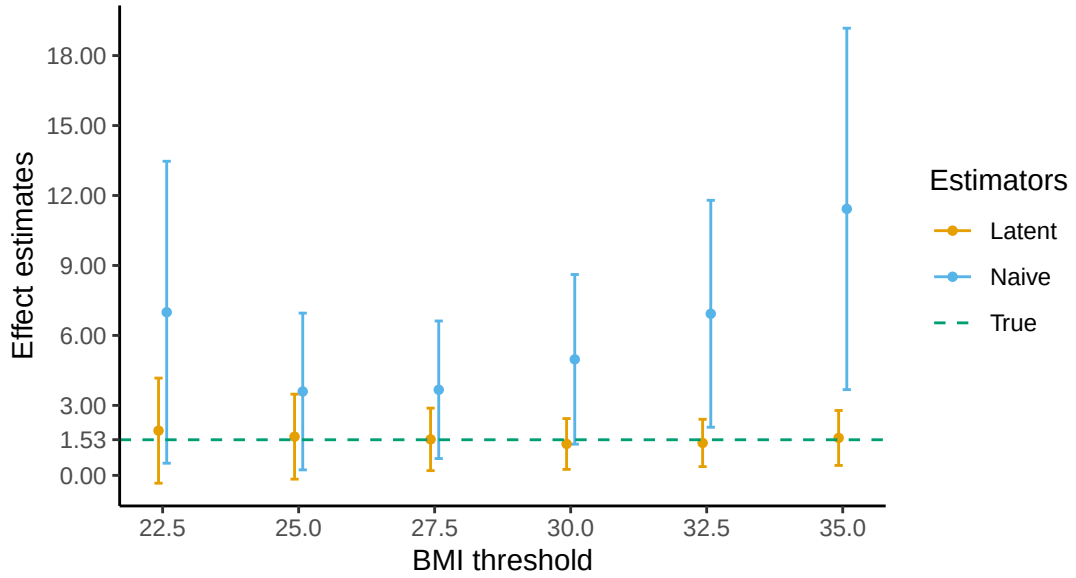


Figure 4.3: Comparison of estimated effect with 'true' effect for various BMI thresholds. $N = 70{,}261$, $\theta^2 = 0.0256$, 95% confidence intervals are generated over 1,000 bootstrap resamples. 'True' corresponds to the sample estimate using BMI as the exposure; 'naive' corresponds to using the dichotomous measurement as the exposure β_D; and 'latent' corresponds to the latent variable estimator β_L of Section 4.3.2.

of child adiposity (e.g., BMI) rather than the three-point scale, so the exclusion restriction is likely to be violated (Richardson, Sanderson, et al., 2020). We use our latent variable approach to ameliorate this bias and to estimate the effect of child BMI directly, which is the exposure of interest.

4.5.2 Re-analysis of Pasman et al. (2018)

Pasman et al. (2018) performs a two-sample bi-directional Mendelian randomization analysis of schizophrenia and cannabis use (Burgess, Scott, et al., 2015). The gene-exposure associations for schizophrenia are pulled from a GWAS of cases and controls and are reported on the log-odds scale (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2015). While this avoids the problem of using the dichotomous diagnosis variable as the exposure (as discussed in Section 4.1), it means that the resulting estimates are interpreted as unit increases in the log-odds, which are scaled by the unobserved parameter σ_V. The authors report an odds ratio of 1.16 (95% CI 1.06 - 1.27) for the effect of genetic liability to schizophrenia. While we can infer the direction of the effect from this estimate, we cannot draw any conclusions about the magnitude.

We apply the two-sample generalization of Section 4.4.4. One of the strengths of this generalization is that we do not need to re-estimate the original inverse-variance weighted Mendelian randomization estimates ourselves. In addition to the estimates reported in the

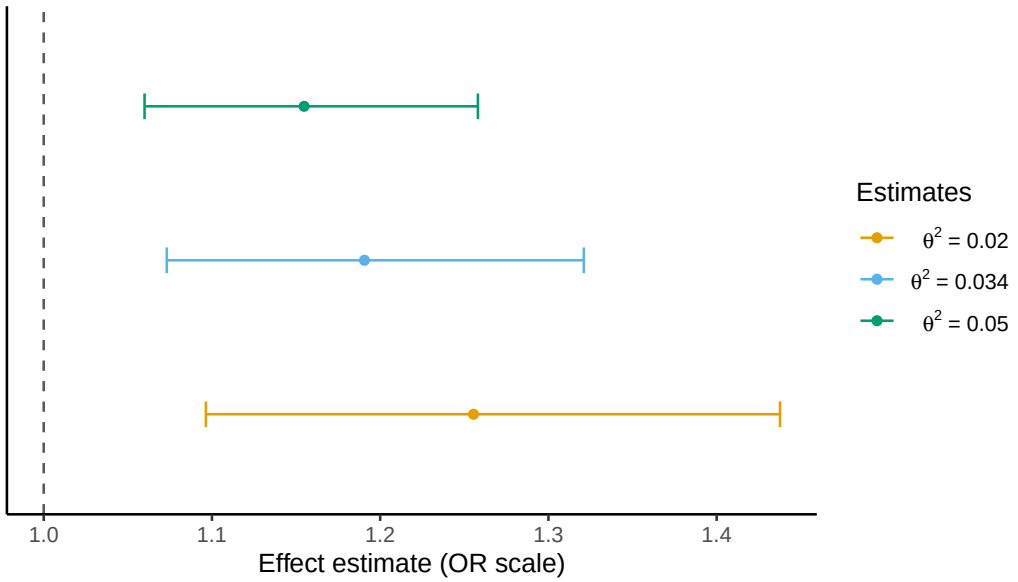


Figure 4.4: Effect of schizophrenia liability on risk of ever using cannabis for several choices of sensitivity parameter θ^2. 95% confidence intervals are estimated as in Section B.3 of the Appendix.

original paper, we need only an estimate of σ_{G_*}, which can be computed from summary data from the schizophrenia GWAS, and some plausible choices for the sensitivity parameter θ^2. The schizophrenia GWAS reports that their genome-wide significant loci explain roughly 3.4% of the variation in schizophrenia liability using the Lee, Goddard, et al. (2012) coefficient of determination. Using this estimate as a baseline, we select three choices for θ^2: 0.02, 0.034 and 0.05.

Our findings are consistent with a modest positive effect of schizophrenia liability on the odds of cannabis use. As shown in Figure 4.4, a one standard deviation increase in schizophrenia liability corresponds to a 1.15-1.26 increase in the odds of cannabis use, with 95% confidence interval range of 1.10-1.44. It is important not to directly compare these estimates with the original estimates: the two are not on the same scale. We must interpret the estimates of Table 4.4 in terms of standard deviation increases in schizophrenia liability.

4.5.3 Reanalysis of Richardson, Sanderson, et al. (2020)

Richardson, Sanderson, et al. (2020) performs two-sample Mendelian randomization analysis of child and adult BMI on risk of several diseases: coronary artery disease, type 2 diabetes, breast cancer and prostate cancer. The instrument-exposure relationship is estimated in the UK Biobank cohort. However, child BMI is not measured directly in UK Biobank, instead, there is a measure of self-reported adiposity in three discrete categories ('thinner', 'plumper' or 'about average'). In this context, the latent exposure is child BMI and the self-report measure

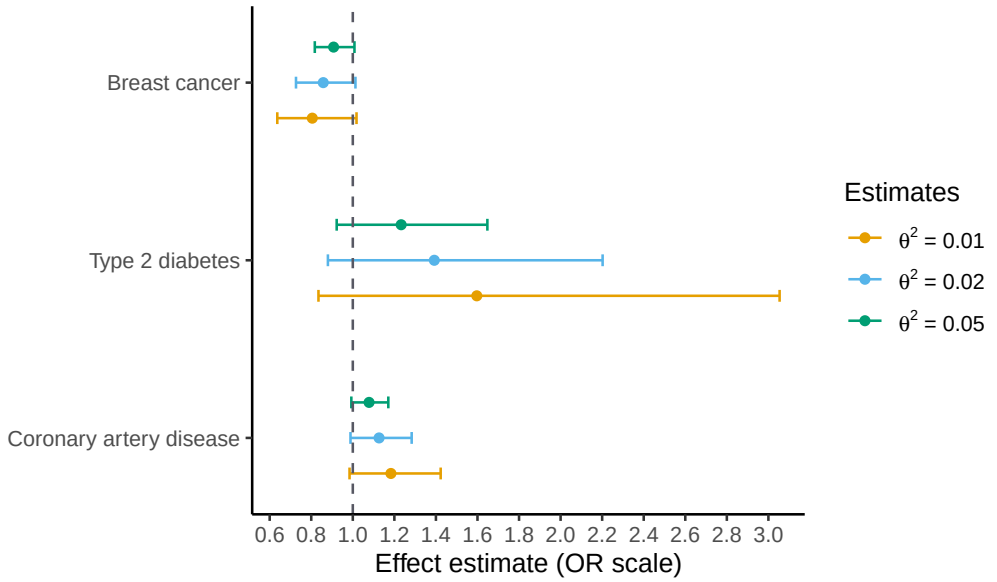


Figure 4.5: Effect of childhood BMI on risk of several diseases for several choices of sensitivity parameter θ^2. 95% confidence intervals are estimated as in Section B.3 of the Appendix.

is a coarsening of child BMI. Since the genetic instruments will act on child BMI directly, the exclusion restriction is likely to be violated.

Therefore, we apply the latent variable method of Section 4.3.2 to this data. We re-analyse the original univariable effect of child BMI on risk of type 2 diabetes (OR 2.32, 95% CI 1.76-3.05), coronary artery disease (1.49, 1.33-1.68) and breast cancer (0.59, 0.50-0.71).

We apply the two-sample generalization of the inverse-variance weighted estimator of Section 4.4.4, estimating the instrument-exposure relationship in UK Biobank using an ordered probit model and the instrument-outcome relationships using the MR-Base platform (Hemani, Zheng, et al., 2018). We choose three values for θ^2 based on a large GWAS of adult BMI: 0.01, 0.02 and 0.05 (Locke et al., 2015). The genetic share of child BMI is estimated using an ordered probit model and standard errors are calculated using the formula in Section B.3 of the Appendix.

Figure 4.5 shows our results for three of the diseases analysed in the paper. Our estimates are in the same direction as the original estimates, which is expected, however, the interpretation of the magnitudes is different. For example, the original paper estimates that a per-category increase in self-reported child adiposity corresponds to an increase in the odds of coronary artery disease of 1.49 (95% CI 1.33-1.68), which could be inflated due to violation of the exclusion restriction. For $\theta^2 = 0.02$, we estimate that a one standard deviation increase in child BMI corresponds to an increase in the odds of coronary artery disease of 1.13 (95% CI 0.99-1.28). It is difficult to directly compare the two sets of estimates since the exposures are different, however, our estimate is suggestive of a modest effect of child BMI on the risk of coronary artery disease.

4.6 Discussion

We propose a simple framework for estimation and interpretation of Mendelian randomization for coarsened measurements of latent continuous exposures. We begin by demonstrating in Section 4.3.1 that using the coarse measurement as the exposure results in a multiplicative bias which will inflate or deflate effect estimates without reversing their sign. However, under the assumptions of our framework, described in Section 4.2, we can recover the effect of the latent exposure in terms of standard deviation increases. Section 4.4.4 shows that it is straight-forward to generalize this approach to the two-sample setting. The key sensitivity parameter in our approach is the genetic share of the variance of the latent exposure, which may be estimated or varied over a plausible range of values (Lee, Goddard, et al., 2012). Section 4.5 evaluates this approach by creating binary exposure measurements from the continuous BMI measure in UK Biobank. We show that we can accurately recover the effect of a standard deviation increase in BMI on systolic blood pressure. We also demonstrate this approach in practice by re-analysing two papers which are likely to suffer from this type of exclusion restriction violation, allowing us to meaningfully interpret their effect sizes.

The approach proposed in this paper relies on a number of strong structural assumptions on the relationship between the latent exposure and its corresponding measurement. The appropriateness of these assumptions must be assessed on a case-by-case basis. Exposure measurements which are defined by strict thresholds of the latent continuous exposure are easiest to conceptualize within this framework. In general, the assumption most difficult to justify is that the thresholds are independent of the genetic share of the latent exposure. One example where this assumption may be violated is self-report measures of mental health status, for example, feelings of depression on a 1-5 scale. Individuals who are genetically predisposed to depression may have different thresholds for reporting their mental wellbeing, either over- or under-reporting.

An additional complication occurs when this method is applied to disease exposures. We have assumed throughout that disease occurrence and disease diagnosis are equivalent; that is, everyone who develops the disease will receive a diagnosis. However, there are often barriers to seeking and accessing the healthcare services needed to receive a diagnosis. These might include stigma surrounding the disease, a lack of trust in healthcare providers or a lack of access to healthcare services due to cost, distance or institutional complexities (Cassim et al., 2019; Stangl et al., 2019). It is therefore possible that individuals with the disease will fail to be diagnosed. This can be viewed as a form of misclassification bias. Misclassification-robust methods for binary exposures could potentially be incorporated into this approach, which we leave for future work (Lewbel, 2000; Rekaya et al., 2016; Smith, Hay, et al., 2013).

In studies where the assumptions in Section 4.2 are believed to be implausible, it is important for researchers to be transparent that the magnitude of their effect estimate will not be well-defined.

Chapter 5

Almost exact Mendelian randomization

Publication arising from this chapter

This chapter has been pre-printed on the arXiv server under the title "Almost exact Mendelian randomization" Tudball2022a. We intend to submit to the journal *Statistical Science*. My contribution was: conceptualization; stating assumptions and proving theorems, lemmas and propositions; designing the causal model; coding and validating the accompanying R package; designing and analyzing the simulation study and applied example; writing the manuscript. Co-authors' contribution was: conceptualization, assistance with technical statements, proofs and design of the causal model; reviewing code for the software package and suggesting improvements; providing feedback on, and making edits to, the manuscript.

SIGNED: Matthew Tudball (First Author) DATE: 24 July, 2022
SIGNED: Qingyuan Zhao (Senior Author) DATE: 24 July, 2022

Software

I prepared an open source R package to accompany this publication. It implements the "almost exact" test described in this chapter. See `https://github.com/matt-tudball/almostexactmr` for installation instructions.

5.1 Introduction

5.1.1 Towards an almost exact inference for MR

As parent-offspring trio data become more widely available, it is increasingly feasible to perform MR within families, as originally proposed by Davey Smith and Ebrahim (2003). There has been some recent methodological development for within-family designs (Davies, Howe, et al., 2019; Brumpton et al., 2020). Thus far this has consisted of extensions of traditional MR techniques in which structural models for the gene-exposure and gene-outcome relationships are proposed and samples are assumed to be drawn according to these models from some large population. In particular, Brumpton et al. (2020) propose a linear structural model with parental genotype fixed effects. Their inference is based on this model and so the role of meiotic randomization is only implicit.

However, one of the unique advantages of MR as a natural experimental design is that it has an explicit inferential basis, namely the exogenous randomness in meiosis and fertilization, which has been thoroughly studied and modelled in genetics since at least Haldane (1919). Haldane developed a simple model for recombination during meiosis that has demonstrated good performance on multiple pedigrees across many species. The connection between this meiosis model and causal inference in parent-offspring trio studies was recently described in the context of locating causal genetic variants (Bates, Sesia, Sabatti, and Candès, 2020) and was implicit in earlier pedigree-based methods, such as the genetic linkage analysis in Morton (1955) and the transmission disequilibrium test in Spielman, McGinnis, and Ewens (1993). Lauritzen and Sheehan (2003) attempted to represent meiosis using graphical models; however, they focused on computatational considerations and did not explore the potential of these models for causal inference.

The idea of the significance test (or hypothesis test, although some authors distinguish the use of these two terms) dates back to Fisher's original proposal for randomized experiments and is well illustrated in his famous 'lady tasting tea' example (Fisher, 1935). Pitman (1937) appears to be the first to fully embrace the idea of randomization tests. This mode of reasoning is usually referred to as randomization inference or design-based inference to contrast with model-based inference. With the aid of the potential outcome framework (Neyman, 1990; Rubin, 1974), we can construct an exact randomization test for the sharp null hypothesis by conditioning on all the potential outcomes (Rubin, 1980; Rosenbaum and Rubin, 1983). Randomization tests are widely used in a variety of settings, including genetics (Spielman, McGinnis, and Ewens, 1993; Bates, Sesia, Sabatti, and Candès, 2020), clinical trials (Rosenberger, Uschner, and Wang, 2019), program evaluation (Heckman and Karapakula, 2019) and instrumental variable analysis (Rosenbaum, 2004; Kang, Peck, and Keele, 2018).

5.1.2 Our contributions

In this article, we propose a statistical framework that enables researchers to use meiosis models as the "reasoned basis" for inference in MR. We propose a randomization test that is *almost exact* in the sense that the test would have exactly the nominal size if the model for meiosis and fertilization were perfect. As detailed below, our methodological developement in Section 5.3 combines several important ideas in the literature.

Our first contribution is a theoretical description of MR and its assumptions in Section 5.3.1 via the language of causal directed acyclic graphs (DAGs) (Spirtes, Glymour, and Scheines, 2000; Pearl, 2009). These graphical tools allow us to visualize and dissect the assumptions imposed on an MR study. In particular, we show how various biological and social processes, including population stratification, gamete formation, fertilization, genetic linkage, assortative mating, dynastic effects, and pleiotropy, can be represented using a DAG and how they can introduce bias in MR analyses. Furthermore, by using single world intervention graphs (SWIGs) (Richardson and Robins, 2013b), we identify sufficient confounder adjustment sets to eliminate these sources of bias in Section 5.3.2. Our results further provide theoretical insights into a fundamental trade-off between statistical power and eliminating pleiotropy-induced bias.

For statistical inference, we propose in Section 5.3.3 a randomization test by connecting two existing literatures. The first literature concerns randomization inference for instrumental variable analyses, which usually assumes that the instrumental variables are randomized according to a simple design (such as random sampling of a binary instrument without replacement) (Rosenbaum, 2004; Kang, Peck, and Keele, 2018). However, in MR, offspring genotypes are very high-dimensional and are randomized based on the parental haplotypes. The second literature attempts to identify the approximate location of ("map") causal genetic variants by modelling the meiotic process (Morton, 1955; Spielman, McGinnis, and Ewens, 1993; Bates, Sesia, Sabatti, and Candès, 2020). In Sections 5.3.4 to 5.3.6, we consider some practical issues with the randomization tests. In particular, we show how the hidden Markov model for meiosis and fertilization implied by Haldane (1919) can greatly simplify the sufficient adjustment sets and the computation of our randomization test.

In addition to the considerable conceptual advantages, our almost exact MR approach has several practical advantages too. First, unlike model-based approaches for within-family MR (Brumpton et al., 2020), our approach does not rely on a correctly specified phenotype model. Nonetheless, the randomization test can take advantage of a more accurate phenotype model to increase its power. Second, Haldane's hidden Markov model implies a propensity score for each genetic instrument given a sufficient adjustment set (Rosenbaum and Rubin, 1983). This can be used as a "clever covariate" (Rose and Laan, 2008) to build powerful test statistics with attractive robustness properties. Third, since the randomization test is exact, it is robust to arbitrarily weak instruments. For an "irrelevant" instrument which induces no variation in the exposure, the test will simply have no power. Finally, by taking advantage of

the DAG representation and using a sufficient confounder adjustment set, our method is also provably robust to biases arising from population structure (including multi-ancestry samples), assortative mating, dynastic effects and pleiotropy by linkage.

In Sections 5.4 and 5.5, we demonstrate the practicality of the almost exact approach to MR with a simulation study and a real data example from the Avon Longitudinal Study of Parents and Children (ALSPAC). The simulation study confirms that the randomization test is exact under the null and explores the power of the test in a number of scenarios. The applied examples consists of a negative control and a positive control. The negative control is the effect of child's body mass index (BMI) at age 7 on mother's BMI pre-pregnancy. Although a causal effect is temporally impossible, the existence of confounders (a.k.a. backdoor paths) may lead to false rejections of the null. The positive control is the effect of child's BMI on itself plus some noise. We compare our results with the results from a "standard" MR analysis that does not condition on parental or offspring haplotypes. We conclude with some further discussion in Section 5.6.

Throughout the paper, we use i to index the parent-offspring trio (or just the offspring) and j to indicate a genomic locus. Bold font is used to represent vectors and script font is used for sets.

5.2 Background

5.2.1 Causal inference preliminaries

We will express our model and assumptions about almost exact MR using causal diagrams, then demonstrate that a randomization test for instrumental variables is a natural vehicle for inference in within-family MR. As such, a good grasp of these concepts is required to understand the remainder of the article. This section lays out some standard notation in causal inference. A lengthier introduction to the causal inference concepts used in this article—including causal graphical models, single world intervention graphs, randomization inference, and instrumental variables—can be found in Section 2.1.

Suppose we have a collection of N individuals indexed by $i = 1, 2, \ldots, N$ and, among these individuals, we are interested in the effect of an exposure D_i on an outcome Y_i. For example, the exposure could be the level of alcohol consumption over some period of time and the outcome could be the incidence of cardiovascular disease. Individual i's *potential* (or *counterfactual*) *outcomes* corresponding to exposure level $D_i = d$ are given by $Y_i(d)$. We make the *consistency* assumption (Hernán and Robins, 2020) which states that the observed outcome corresponds to the potential outcome at the realized exposure level $Y_i = Y_i(D_i)$.

Note that, in denoting the potential outcomes as $Y_i(d)$, we have implicitly made the so-called *stable unit treatment value assignment* (SUTVA) assumption (Rubin, 1980; Imbens and Rubin, 2015). That is, we have assumed that there is *no interference* in the sense that the potential

outcomes of each individual are unaffected by the exposures of other individuals. We have also assumed that there are no hidden versions of the same exposure; this could be violated, for example, if the effect of alcohol consumption on cardiovascular disease has a dose-response relationship but the exposure D_i is only a binary indicator of alcohol consumption.

Potential outcomes may also be defined from a nonparametric structural equation model associated with a causal diagram using recursive substitution (Pearl, 2009). In such diagrams, vertices are used to represent random variables and directed edges are used to represent direct causal influences. The graphical and potential outcomes approaches to causal inference can be nicely unified via the single world intervention graphs (Richardson and Robins, 2013b). A brief review of this can be found in the Appendix.

5.2.2 Genetic preliminaries

Before proceeding to present our model of within-family MR, it is also instructive to provide a basic overview of some relevant concepts in human genetics, with a focus on modelling the processes in genetic inheritance including *meiosis* and *fertilization*. For a thorough exposition on statistical models for pedigree data, see Thompson (2000).

Human somatic cells consist of 23 pairs of chromosomes, with one in each pair inherited from the mother and the other from the father. To simplify the discussion we will only consider autosomal (non-sexual) chromosomes. Each chromosome is a doubled strand of helical DNA composed of complementary nucleotide base pairs. A base pair which exhibits population-level variation in its nucleotides is called a *single nucleotide polymorphism* (SNP). DNA sequences are typically characterized by detectable variant forms induced by different combinations of SNPs. These variant forms are called *alleles*. In this article, we will only consider variants with two alleles. A set of alleles on one chromosome inherited together from the same parent is called a *haplotype* and the unordered pair of haplotypes at the same locus is called a *genotype*.

Meiosis is a type of cell division that results in reproductive cells (a.k.a. gametes) containing one copy of each chromosome. During this process, homologous chromosomes line up and exchange segments of DNA between themselves in a biochemical process called *crossover*. The recombined chromosomes are then further divided and separated into gametes. Since recombinations are infrequent (roughly one to four per chromosome), SNPs located nearby on the same parental chromosome are more likely to be transmitted together, resulting in *genetic linkage*. Fertilization is the process by which gametes in the father (sperm cells) and mother (egg cells) join together to form a zygote, which will then normally develop into an embryo.

We will mainly be concerned with genetic trio studies, in which we observe the haplotypes of the mother, father and their child at p loci. Let $\mathcal{J} = \{1, 2, \ldots, p\}$ be the set of SNP indices. In our discussion below we will assume that the SNPs are on a single chromosome, as different chromosomes are usually modelled independently. We will denote the haplotypes as follows:

offspring's haplotypes: $\boldsymbol{Z}^m = (Z_1^m, \ldots, Z_p^m)$ and $\boldsymbol{Z}^f = (Z_1^f, \ldots, Z_p^f)$,

mother's haplotypes: $\boldsymbol{M}^m = (M_1^m, \ldots, M_p^m)$ and $\boldsymbol{M}^f = (M_1^f, \ldots, M_p^f)$, and

father's haplotypes: $\boldsymbol{F}^m = (F_1^m, \ldots, F_p^m)$ and $\boldsymbol{F}^f = (F_1^f, \ldots, F_p^f)$,

where the superscript m (or f) indicates a maternally (or paternally) inherited haplotype. We only consider SNPs with two alleles, so each of the six haplotype vectors above is in $\{0,1\}^p$. Furthermore, let $\boldsymbol{M}_j^{mf} = (M_j^m, M_j^f)$ denote the mother's haplotypes at locus j and similarly for $\boldsymbol{F}_j^{mf}$ and $\boldsymbol{Z}_j^{mf}$. The offspring's genotype at locus $j \in \mathcal{J}$ is given by $Z_j = Z_j^m + Z_j^f$ and let $\boldsymbol{Z} = \boldsymbol{Z}^m + \boldsymbol{Z}^f \in \{0,1,2\}^p$ denote the vector of offspring genotypes.

Figure 5.1: Illustration of the meiotic process for five sites on a chromosome.

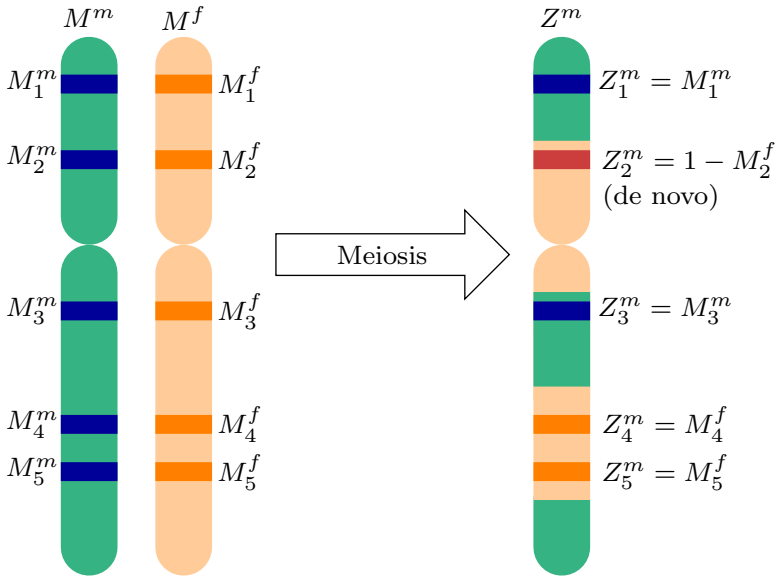


Figure 5.1 illustrates how an offspring's maternally-inherited haplotype $\boldsymbol{Z}^m$ at five loci on a chromosome are related to the mother's haplotypes $\boldsymbol{M}^m$ and $\boldsymbol{M}^f$. In a gamete produced by meiosis, the allele at locus $j \in \mathcal{J}$ is inherited from either the mother's maternal haplotype or father haplotype (excluding mutations). This can be formalized as an ancestry indicator, $U_j^m \in \{m, f\}$. The classical meiosis model of Haldane (1919) assumes that $\boldsymbol{U}^m = (U_1^m, \ldots, U_p^m)$ follows a (discretized) homogeneous Poisson process. Haldane's model is described in Appendix C.1 in detail and can simplify the randomization test considerably (Section 5.3.5). Nonetheless, our "almost exact" approach to MR is modular in the sense that it does not rely on a specific meiosis model and it is theoretically straightforward to incorporate more sophisticated meiosis models that allow for "interference" between the crossovers (Otto and Payseur, 2019). As the meiosis model become more accurate, our test will become closer to exact randomization inference.

The description in the last paragraph does not take genetic mutation into account. Many meiosis models, such as the one in Haldane (1919), assume that there is a small probability of independent mutations:

Assumption 5.1. Given mother's haplotypes $\boldsymbol{M}_j^{mf}$, the ancestry indicator U_j^m, and that fertilization occurs (this is represented as $S = 1$ in Section 5.3),

$$Z_j^m = \begin{cases} M_j^{(U_j^m)}, & \text{with probability } 1-\epsilon, \\ 1 - M_j^{(U_j^m)}, & \text{with probability } \epsilon. \end{cases}$$

The same model holds for the paternally-inherited haplotypes.

The rate of *de novo* mutation ϵ is about 10^{-8} in humans (Acuna-Hidalgo, Veltman, and Hoischen, 2016). When only one generation is considered (as in a genetic trip study), it often suffices to treat $\epsilon = 0$ (i.e. no mutation) for practical purposes, unless it is desirable to obtain the exact randomization distribution under a recombination model.

The above meiosis model assumes no *transmission ratio distortion*. Transmission ratio distortion occurs when one of the two parental alleles is passed on to the (surviving) offspring at more or less than the expected Mendelian rate of 50% (Davies, Howe, et al., 2019). Transmission ratio distortion falls into two categories: segregation distortion, when processes during meiosis or fertilization select certain alleles more frequently than others, and viability selection, when the viability of zygotes themselves depend on the offspring genotype. We sidestep this discussion for now and return to it in Section 5.6.

5.3 Almost exact Mendelian randomization

Returning to the alcohol and cardiovascular disease example in Section 5.2.1, observational studies suggest that moderate alcohol consumption confers reduced risk relative to abstinence or heavy consumption (Millwood et al., 2019). However, this could be a result of systematic differences among people with different drinking patterns (confounding) rather than a causally protective effect of moderate drinking. For this reason, Mendelian randomization has become a popular study design to investigate the long-term health effects of alcohol drinking (Chen et al., 2008).

The *ALDH2* gene regulates acetaldehyde metabolism and is often used as an instrumental variable for alcohol consumption. In East Asian populations, an allele of *ALDH2* produces a protein that is inactive in metabolising acetaldehyde, causing flushing and discomfort while drinking and thereby reducing consumption. Thus, we might like to use the random allocation of variant copies of *ALDH2* as the basis of causal inference. To this end, we need to carefully clarify the conditions under which this inference would be valid.

5.3.1 A causal model for Mendelian inheritance

Next, we construct a general grahical model to describe the process of Mendelian inheritance and genotype-phenotype relationships. This causal model allows us to identify sources of bias

in within-family MR and construct sufficient adjustment sets to control for them. Under this causal model, the central idea behind almost exact MR is to base statistical inference precisely on randomness in genetic inheritance via a model for meiosis and fertilization.

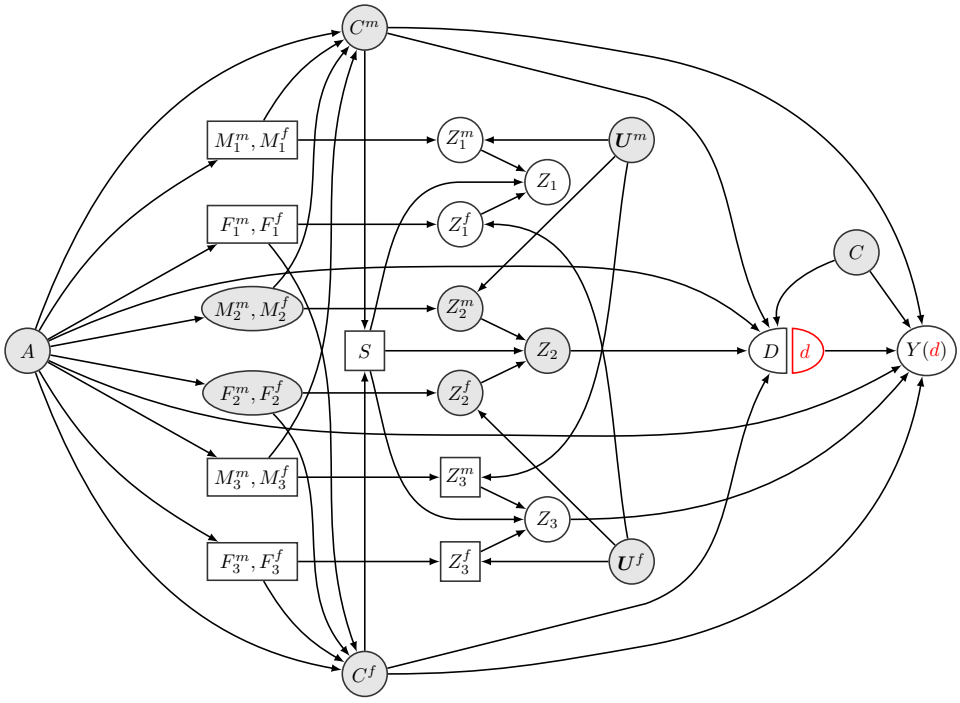


Figure 5.2: The single world intervention graph for a working example of a chromosome with $p = 3$ variants. Transparent nodes are observed and grey nodes are unobserved. Square nodes are the confounders being conditioned on in Proposition 5.2. A is ancestry; $\boldsymbol{M}^f = (M_1^f, M_2^f, M_3^f)$ is the mother's haplotype inherited from her father; $\boldsymbol{M}^m$, $\boldsymbol{F}^m$, and $\boldsymbol{F}^f$ are defined similarly; C^m and C^f are generic phenotypes of the mother and father; S is an indicator of mating; $\boldsymbol{Z}^m = (Z_1^m, Z_2^m, Z_3^m)$ is the offspring's maternal haplotype and $\boldsymbol{U}^m$ is a meiosis indicator; $\boldsymbol{Z}^f$ and $\boldsymbol{U}^f$ are defined simlarly; $\boldsymbol{Z} = (Z_1, Z_2, Z_3)$ is the offspring's genotype; D is the exposure of interest; $Y(d)$ is the potential outcome of Y under the intervention that sets D to d; C is an environmental confounder between D and Y.

Figure 5.2 shows a working example of our causal model on a hypothetical chromosome with $p = 3$ variants. The directed acyclic graph is structured in roughly chronological order from left to right, where A describes the population structure, S is an indicator for mating, and C is any environmental confounder between the exposure D and outcome Y.

At first glance, Figure 5.2 appears to be quite complicated but, by the modularity of graphical models, it can be decomposed into a collection of much simpler subgraphs that describe different biological processes (Figure 5.3). By definition, a joint distribution *factorizes* according to the DAG in Figure 5.2 if its density function can be decomposed as described in Table 5.1.

Next, we describe each term in Table 5.1 and what this factorization implies about our assumptions on the biological processes. To simplify the discussion, we assume all DAGs in

Table 5.1: Factorization of the joint density of all variables in Figure 5.2. Here p is used as a generic symbol for density function.

Terms	Interpretation	More detail
$p(A)p(\boldsymbol{U}^m)p(\boldsymbol{U}^f)p(C)$	Exogenous variables	
$p(\boldsymbol{M}^m, \boldsymbol{M}^f \mid A)p(\boldsymbol{F}^m, \boldsymbol{F}^f \mid A)$	Pop. stratification	Section 5.3.1.1
$p(C^m \mid A, \boldsymbol{M}^m, \boldsymbol{M}^f)p(C^f \mid A, \boldsymbol{F}^m, \boldsymbol{F}^f)$	Parental phenotypes	Section 5.3.1.2
$p(\boldsymbol{Z}^m \mid \boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{U}^m)p(\boldsymbol{Z}^f \mid \boldsymbol{F}^m, \boldsymbol{F}^f, \boldsymbol{U}^f)$	Meiosis	Section 5.3.1.3
$p(S \mid C^m, C^f)$	Assortative mating	Section 5.3.1.3
$p(\boldsymbol{Z} \mid \boldsymbol{Z}^m, \boldsymbol{Z}^f, S)$	Fertilization	Section 5.3.1.3
$p(D \mid A, \boldsymbol{Z}, C^m, C^f, C)p(Y(d) \mid A, \boldsymbol{Z}, C^m, C^f, C)$	Confounding	Section 5.3.1.4

this article are faithful, so conditional independence between random variables is equivalent to d-separation in the DAG.

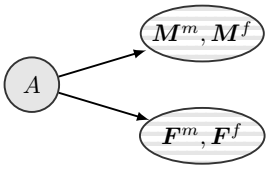
5.3.1.1 Parental genotypes

We assume that parental genotypes originate from some arbitrary, latent population structure. Population stratification is a phenomenon characterized by systematic differences in the distribution of alleles among subgroups of a population. These disparities typically emerge from social and genetic mechanisms including non-random mating, migration patterns and 'founder effects' (Cardon and Palmer, 2003) and can often be detected by principal component analysis (Patterson, Price, and Reich, 2006). Population stratification can introduce spurious associations between genetic variants and traits (Lander and Schork, 1994).

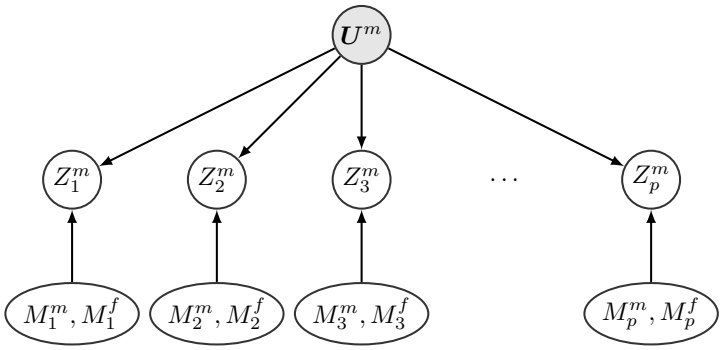
We represent population structure via a latent node A in Figures 5.2 and 5.3a. The arrows from A to $\boldsymbol{M}^m, \boldsymbol{M}^f$ and $\boldsymbol{F}^m, \boldsymbol{F}^f$ indicate that the distribution of parental haplotypes depends on the latent population structure:

$$A \not\perp\!\!\!\perp (\boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f).$$

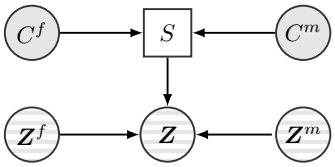
The node A may also capture any linkage disequilibrium in the parental haplotypes. That is, because the parental haplotypes are determined by the same process as the grandparental haplotypes and so on, recombination introduces dependence among nearby genetic variants (see Section 5.3.1.3 below for more discussion). The precise distribution of A and the conditional distribution of the parental haplotypes given A are not important below, because an appropriate subset of the parental haplotypes will be conditioned on and any paths that involve edges from A to $\boldsymbol{M}^m$, $\boldsymbol{M}^f$, $\boldsymbol{F}^m$, and $\boldsymbol{F}^f$ will be blocked.



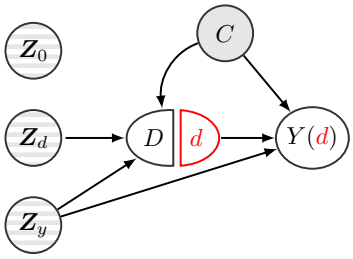
(a) Population structure (Section 5.3.1.1).



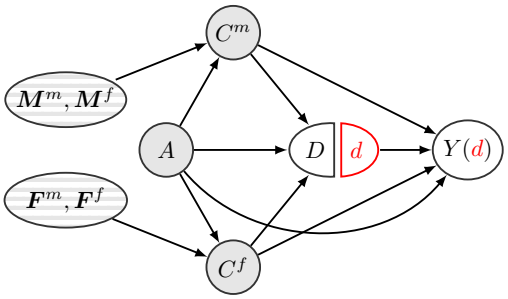
(b) Meiotic recombination of the mother's haplotypes (Section 5.2.2).



(c) Assortative mating (Section 5.3.1.3).



(d) Offspring phenotypes. (Section 5.3.1.4).



(e) Dynastic effects (Section 5.3.1.2).

Figure 5.3: The constituent subgraphs of our within-family Mendelian randomization model. White nodes represent observed variables; grey nodes represent unobserved variables; and striped nodes represent variables for which some elements may be unobserved.

5.3.1.2 Parental phenotypes

We impose no assumptions on the nature and the distribution of the parental phenotypes C^m and C^f. They can depend arbitrarily on the parental haplotypes $\boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f$ and the population structure A:

$$C^m \not\perp\!\!\!\perp (A, \boldsymbol{M}^m, \boldsymbol{M}^f),\ C^f \not\perp\!\!\!\perp (A, \boldsymbol{F}^m, \boldsymbol{F}^f).$$

It is not necessary to model such dependence because the almost exact approach to MR conditions on appropriate parental haplotypes.

5.3.1.3 Offspring genotypes

There are two biological processes involved in the genesis of the offspring's genotypes: meiosis (gamete formation) and fertilization. The meiotic process has been reviewed in Section 5.2.2, and the key Assumption 5.1 can be represented by the causal diagram in Figure 5.3b (for just the mother). A crucial assumption underlying almost exact MR is the exogeneity of the meiosis indicators $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$. This is reflected in Figures 5.2 and 5.3b, as $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$ have no parents and their only children are the offspring's haplotypes. Formally, we assume:

Assumption 5.2. The meiosis indicators are independent of parental haplotypes and phenotypes and any other confounders:

$$(\boldsymbol{U}^m, \boldsymbol{U}^f) \perp\!\!\!\perp (A, C^m, C^f, C, \boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f).$$

Many stochastic processes have been proposed to model the distribution of the ancestry indicator $\boldsymbol{U}^m$; see Otto and Payseur (2019) for a recent review. Due to the dependence in $\boldsymbol{U}^m$, nearby alleles on the same chromosome tend to be inherited together. This phenomenon is known as *genetic linkage*. In Section 5.2.2, we have described the classical model of Haldane (1919) which assumes $\boldsymbol{U}^m$ follows a Poisson process, which was used by Bates, Sesia, Sabatti, and Candès (2020) to map causal variants. We will see in Section 5.3.5 that the Markov properties of a Poisson process greatly simplify randomization inference.

Another mechanism that needs to be modeled is fertilization. In Mendelian inheritance, it is assumed that the potential gametes (sperms and eggs) come together at random. However, mating may not be a random event. *Assortative mating* refers to the phenomenon where individuals are more likely to mate if they have complementary phenotypes. For example, there is evidence in UK Biobank that variant forms of the *ADH1B* gene related to alcohol consumption are more likely to be shared among spouses than non-spouses (Howe, Lawson, et al., 2019). This suggests assortative mating on drinking behaviour and may introduce bias to MR studies on alcohol consumption (Hartwig, Davies, and Davey Smith, 2018).

The subgraph describing assortative mating is shown in Figure 5.3c, where the mating indicator $S \in \{0, 1\}$ is dependent on the parental phenotypes C^m and C^f. Any MR study

necessarily conditions on $S = 1$, otherwise the offspring would not exist and the trio data would not be available. This is formalized in Figure 5.3c by the arrows from S to $\boldsymbol{Z}$. In particular, we may define the offspring's genotype $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \begin{cases} \boldsymbol{Z}^m + \boldsymbol{Z}^f, & \text{if } S = 1, \\ \text{Undefined}, & \text{if } S = 0. \end{cases} \tag{5.1}$$

Notice that the above definition recognizes the fact that the gametes $\boldsymbol{Z}^m$ and $\boldsymbol{Z}^f$ are produced regardless of whether fertilization actually takes place.

Finally, the assumption of no transmission ratio distortion is formalized by the absence of arrows from $\boldsymbol{Z}^m$ and $\boldsymbol{Z}^f$ to S. This is not necessarily a benign assumption, given empirical evidence that gametes may pair up non-randomly (Nadeau, 2017).

Assumption 5.3. There is no transmission ratio distortion in the sense that

$$S \perp\!\!\!\perp (\boldsymbol{Z}^m, \boldsymbol{Z}^f) \mid (\boldsymbol{M}^{mf}, \boldsymbol{F}^{mf}).$$

5.3.1.4 Offspring phenotypes

Finally, we describe assumptions on the offspring phenotypes. We are interested in estimating the causal effect of an offspring phenotype $D \in \mathcal{D} \subseteq \mathbb{R}$ on another offspring phenotype $Y \in \mathcal{Y} \subseteq \mathbb{R}$. We refer to D as the exposure variable and Y as the outcome variable. These phenotypes are determined by the offspring genotypes and environmental factors (including parental traits). For a particular realization of the genotypes $\boldsymbol{z}$, we denote the counterfactual exposure as $D(\boldsymbol{z})$. Furthermore, under an additional intervention that sets D to d, we denote the counterfactual outcome as $Y(\boldsymbol{z}, d)$. These potential outcomes are related to the observed data tuple $(\boldsymbol{Z}, D, Y)$ by

$$D = D(\boldsymbol{Z}),\ Y = Y(\boldsymbol{Z}, D) = Y(\boldsymbol{Z}, D(\boldsymbol{Z})),$$

which is a simple extension of Assumption 2.2 (*consistency*) in Section 2.1.

We are interested in making inference about the counterfactuals $Y(d) = Y(\boldsymbol{Z}, d)$ when the exposure is set to $d \in \mathcal{D}$. As the exposure D typically varies according to the population structure, parental phenotypes and other environmental factors, it is not randomized in the sense that

$$Y(d) \not\perp\!\!\!\perp D \text{ for some or all } d \in \mathcal{D}.$$

For example, if D is alcohol consumption and Y is cardiovascular disease, there may exist offspring confounders such as diet or smoking habits which are common causes of both D and Y. The exact nature of the confunders is not very important for MR as it tries to use unconfounded randomness (in $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$) to make causal inference.

It will be helpful to categorize the genetic variants based on whether they have direct causal effects on D and/or Y.

Assumption 5.4. The set $\mathcal{J} = \{1, \dots, p\}$ of genetic variants can be paritioned as $\mathcal{J} = \mathcal{J}_y \cup \mathcal{J}_d \cup \mathcal{J}_0$, where

- $\mathcal{J}_y$ includes all *pleiotropic* variants with a direct causal effect on Y not mediated by D (some of which may have a causal effect on D as well).
- $\mathcal{J}_d$ includes all causal variants for D with no direct effect on Y.
- $\mathcal{J}_0 = \mathcal{J} \setminus (\mathcal{J}_y \cup \mathcal{J}_d)$ includes all other variants.

In our working example in Figure 5.2, $\mathcal{J}_y = \{3\}$, $\mathcal{J}_d = \{2\}$, and $\mathcal{J}_0 = \{1\}$. If the exposure D indeed has a causal effect on the outcome Y, the loci of the causal variants of Y are given by $\mathcal{J}_y \cup \mathcal{J}_d$.

For subscripts $s \in \{0, d, y\}$, we let $\boldsymbol{Z}_s = \{Z_j : j \in \mathcal{J}_s\}$ denote the corresponding genotypes, which has support $\mathcal{Z}_s = \{0, 1, 2\}^{|\mathcal{J}_s|}$. By Assumption 5.4, our potential outcomes can be written as (with an abuse of notation)

$$D(\boldsymbol{z}) = D(\boldsymbol{z}_d), \; Y(\boldsymbol{z}, d) = Y(\boldsymbol{z}_y, d),$$
$$Y(\boldsymbol{z}) = Y(\boldsymbol{z}_y, D(\boldsymbol{z}_d)) = Y(\boldsymbol{z}_y, \boldsymbol{z}_d),$$

where $\boldsymbol{z} = (\boldsymbol{z}_d, \boldsymbol{z}_y, \boldsymbol{z}_0) \in \mathcal{Z}_d \times \mathcal{Z}_y \times \mathcal{Z}_0 = \mathcal{Z}$ and $d \in \mathcal{D}$.

Figure 5.3d provides the graphical representation of Assumption 5.4. Each element of $\boldsymbol{Z}_0$ has no effect on D or $Y(d)$, each element of $\boldsymbol{Z}_d$ has an effect on D and each element of $\boldsymbol{Z}_y$ has an effect on $Y(d)$ (some are also causes of D). The vector $\boldsymbol{Z}_y$ contains the so-called pleiotropic variants that are causally involved in the expression of multiple phenotypes (Hemani, Bowden, and Davey Smith, 2018). Universal pleiotropy is assumed in the famous infinitesimal model of Fisher (1918). Currently, the widely accepted view is that pleiotropy is at least widespread and some have argued for a omnigenic model (Boyle, Li, and Pritchard, 2017).

Dynastic effects, sometimes called *genetic nurture* (Kong et al., 2018), is a phenomenon characterized by parental genotypes exerting an influence on the offspring's phenotypes via the parental phenotypes. This is depicted in Figure 5.3e, where paths emanate from the parental haplotypes $\boldsymbol{M}^{mf}$ and $\boldsymbol{F}^{mf}$ to the parental phenotypes C^m and C^m and on to the offspring phenotypes D and Y.

5.3.2 Conditions for identification

With the causal model outlined in Section 5.3.1 in mind, we now describe some sufficient conditions under which some $Z_j \in \boldsymbol{Z}$ is a valid instrumental variable for estimating the causal effect of D on Y. Recall that an instrumental variable induces unconfounded variation in the exposure without otherwise affecting the outcome. Due to population stratification (Figure 5.3a), assortative mating (Figure 5.3c), and dynastic effects (Figure 5.3e), the offpsring genotypes $\boldsymbol{Z}$

as a whole are usually not properly randomized without conditioning on the parental haplotypes. That is,

$$\boldsymbol{Z} \not\perp\!\!\!\perp D(\boldsymbol{z}), Y(\boldsymbol{z}, d) \text{ for some or all } \boldsymbol{z} \in \mathcal{Z} \text{ and } d \in \mathcal{D}.$$

To restore validity of genetic instruments, the key idea is to condition on the parental haplotypes as envisioned by Davey Smith and Ebrahim (2003) and used in the context of gene mapping by Spielman, McGinnis, and Ewens (1993) and Bates, Sesia, Sabatti, and Candès (2020). This allows us to use precisely the exogenous randomness in the ancestry indicators $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$ that occurs during meiosis and fertilization. This idea is formalized in the next proposition.

Proposition 5.1. *Under the causal graphical model described in Section 5.3.1, the offspring's maternal haplotype Z_j^m (or genotype Z_j) at some site $j \in \mathcal{J}$ is independent of all ancestral and offspring confounders given the maternal (or parental) haplotypes at site j:*

$$\begin{aligned} Z_j^m &\perp\!\!\!\perp (A, C^m, C^f, C) \mid (\boldsymbol{M}_j^{mf}, S = 1), \\ Z_j &\perp\!\!\!\perp (A, C^m, C^f, C) \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{F}_j^{mf}, S = 1). \end{aligned} \tag{5.2}$$

However, the conditional independence (5.2) alone does not guarantee the validity of Z_j as an instrumental variable. The main issue is that Z_j might be in linkage disequilibrium with other causal variants of Y. Our goal is to find a set of variables $\boldsymbol{V}$ such that Z_j is conditionally independent of the potential outcome $Y(d)$. This is formalized in the definition below.

Definition 5.1. We say a genotype Z_j is a *valid instrumental variable* given $\boldsymbol{V}$ (for estimating the causal effect of D on Y) if the following conditions are satisfied:

1. Relevance: $Z_j \not\perp\!\!\!\perp D \mid \boldsymbol{V}$;
2. Exogeneity: $Z_j \perp\!\!\!\perp Y(d) \mid \boldsymbol{V}$ for all $d \in \mathcal{D}$;
3. Exclusion restriction: $Y(z_j, d) = Y(d)$ for all $z_j \in \{0, 1, 2\}$ and $d \in \mathcal{D}$.

Similarly, we say a haplotype Z_j^m is a valid instrument given $\boldsymbol{V}$ if the same conditions above hold with Z_j replaced by Z_j^m and $z_j \in \{0, 1, 2\}$ replaced by $z_j^m \in \{0, 1\}$.

In our setup (Assumption 5.4), the exclusion restriction is satisfied if and only if $j \notin \mathcal{J}_y$.

To gain intuition on how the set of variables $\boldsymbol{V}$ can be selected, it is helpful to return to the working example in Figure 5.2. We see that Z_3 does not satisfy the exclusion restriction because Z_3 has a direct effect on Y. The causal variant Z_2 for D would be a valid instrument if we condition on the corresponding haplotypes and Z_3, but Z_2 is not observed in this example. This leaves us with one remaining candidate instrument: Z_1 (and potentially its haplotypes

Table 5.2: Some paths between Z_1 and $Y(d)$ in Figure 5.2.

Name of bias	Path	Blocking variable
Dynastic effect	$Z_1^m \leftarrow \boldsymbol{M}_1^{mf} \rightarrow C^m \rightarrow Y(d)$	$\boldsymbol{M}_1^{mf}$
Population stratification	$Z_1^m \leftarrow \boldsymbol{M}_1^{mf} \leftarrow A \rightarrow Y(d)$	$\boldsymbol{M}_1^{mf}$
Pleiotropy	$Z_1^m \leftarrow \boldsymbol{U}^m \rightarrow Z_3^m \rightarrow Z_3 \rightarrow Y(d)$	Z_3^m or Z_3
Assortative mating	$Z_1^m \leftarrow \boldsymbol{M}_1^{mf} \leftarrow C^m \rightarrow \boxed{S} \leftarrow C^f \leftarrow \boldsymbol{F}_3^{mf} \rightarrow Z_3^f \rightarrow Z_3 \rightarrow Y(d)$	$\boldsymbol{M}_1^{mf}$ or Z_3^f or Z_3 or $\boldsymbol{F}_3^{mf}$
Nearly determined ancestry	$Z_1^m \leftarrow \boldsymbol{U}^m \rightarrow \boxed{Z_3^m} \leftarrow \boldsymbol{M}_3^{mf} \leftarrow A \rightarrow Y(d)$	$\boldsymbol{M}_3^{mf}$

Z_1^m and Z_1^f). The relevance assumption is satisfied as long as $\boldsymbol{V}$ does not block both of the following paths

$$Z_1 \leftarrow Z_1^m \leftarrow \boldsymbol{U}^m \rightarrow Z_2^m \rightarrow Z_2 \rightarrow D;$$
$$Z_1 \leftarrow Z_1^f \leftarrow \boldsymbol{U}^f \rightarrow Z_2^f \rightarrow Z_2 \rightarrow D.$$

The exclusion restriction is satisfied because Z_1 is not a causal variant for Y. Finally, exogeneity is satisfied if $\boldsymbol{V}$ blocks the path

$$Z_1 \leftarrow Z_1^m \leftarrow \boldsymbol{U}^m \rightarrow Z_3^m \rightarrow Z_3 \rightarrow Y(d);$$
$$Z_1 \leftarrow Z_1^f \leftarrow \boldsymbol{U}^f \rightarrow Z_3^f \rightarrow Z_3 \rightarrow Y(d).$$

Thus, we have the following result:

Proposition 5.2. *For the example in Figure 5.2, the following conditional independence relationships are true for all $d \in \mathcal{D}$:*

$$Z_1^m \perp\!\!\!\perp Y(d) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{V}_3^m, S = 1), \tag{5.3}$$
$$Z_1 \perp\!\!\!\perp Y(d) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{F}_1^{mf}, \boldsymbol{V}_3, S = 1), \tag{5.4}$$

where $\boldsymbol{V}_3^m = (\boldsymbol{M}_3^{mf}, Z_3^m)$ and $\boldsymbol{V}_3 = (\boldsymbol{M}_3^{mf}, \boldsymbol{F}_3^{mf}, Z_3)$. The adjustment variables above are minimal in the sense that no subsets of them satisfy the same conditional independence.

Proof. The conditional independence follows almost immediately from our discussion above. It is tedious but trivial to show that $\boldsymbol{V} = (\boldsymbol{M}_1^{mf}, \boldsymbol{V}_3^m)$ is minimal for (5.3). Table 5.2 lists the key backdoor paths between Z_1^m and $Y(d)$, describes the corresponding biological mechanism and shows how conditioning on $\boldsymbol{V}$ blocks these paths. The table only includes the maternal paths, but the same blocking also holds for the corresponding paternal paths. □

Remark 5.1. To our knowledge, the bias-inducing path in the last row of Table 5.2, which we term "nearly determined ancestry bias", has not yet been identified in the literature. This is a

form of collider bias introduced because the ancestry indicator can often be almost perfectly determined if we are given the mother's haplotypes and the offspring's maternal haplotype. For example, if the mother is heterozygous $M_3^m = 1, M_3^f = 0$ and the offspring's maternal haplotype is $Z_3^m = 1$, then we know that $U_3^m = m$ is true with an extremely high confidence. Due to genetic linkage, there is also an exceedingly high probability that $U_1^m = m$, inducing dependence between the potential instrument Z_1^m with maternal haplotypes $\boldsymbol{M}_3^{mf}$. This further challenges the widely adopted hypothesis that mapping causal variants is equivalent to testing conditional independence; in our example, Z_3 is the only causal variant of $Y(d)$, but $Z_1 \perp\!\!\!\perp Y(d) \mid Z_3$ may not be true even if there is no population stratification and no causal effect from $\boldsymbol{M}_1^{mf}$ on $Y(d)$.

We conclude this section with a sufficient condition for the validity of Z_j^m and Z_j in our general setting. To simplify the exposition, let $\boldsymbol{V}_{\mathcal{B}}^m = (\boldsymbol{M}_{\mathcal{B}}^{mf}, \boldsymbol{Z}_{\mathcal{B}}^m)$ be a set of maternal adjustment variables, where $\mathcal{B} \subseteq \mathcal{J} \setminus \{j\}$ is a subset of loci. Furthermore, let $\boldsymbol{V}_{\mathcal{B}} = (\boldsymbol{M}_{\mathcal{B}}^{mf}, \boldsymbol{F}_{\mathcal{B}}^{mf}, \boldsymbol{Z}_{\mathcal{B}})$.

Theorem 5.1. *Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ is a full chromosome. Consider the general causal model for Mendelian randomization in Section 5.3.1 and let $j \in \mathcal{J}$ be the index of a candidate instrument. Then Z_j^m is a valid instrument conditional on $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m)$ if the following conditions are satisfied:*

1. $Z_j^m \not\perp\!\!\!\perp \boldsymbol{Z}_d^m \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m, S = 1)$;
2. $Z_j^m \perp\!\!\!\perp \boldsymbol{Z}_y^m \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m, S = 1)$.

Proof. The relevance of Z_j^m follows from the first condition, because Z_j^m is dependent on some causal variants (or is itself a causal variant) of D. The exclusion restriction ($j \notin \mathcal{J}_y$) follows directly from the second condition. For exogeneity, paths from Z_j^m to $Y(d)$ either go through the confounders A, C^f, C^m, or C, which are blocked by $\boldsymbol{M}_j^{mf}$ by Proposition 5.1, or through some causal variants of the outcome as in $Z_j^m \leftarrow \boldsymbol{U}^m \rightarrow \boldsymbol{Z}_y^m \rightarrow \boldsymbol{Z}_y \rightarrow Y(d)$, which are blocked by the second condition. □

It is straightforward to extend Theorem 5.1 to establish validity of the genotype Z_j at locus j as an instrumental variable. Details are omitted.

Since Proposition 5.1 ensures that, after conditioning on $\boldsymbol{M}_j^{mf}$, the instrument Z_j^m is independent of all ancestral and offspring confounders (A, C^m, C^f, C), the only remaining threats to the validity of Z_j^m are irrelevance and pleiotropy. The set $\mathcal{B}$ is chosen to ensure that Z_j^m is independent of all pleiotropic variants conditional on $\boldsymbol{V}_{\mathcal{B}}^m$ (condition 2 of Theorem 5.1) but not independent of the set of causal variants (condition 1 of Theorem 5.1).

This highlights an intrinsic trade-off in choosing the adjustment set $\boldsymbol{V}_{\mathcal{B}}$: by choosing a larger subset $\mathcal{B}$, the second condition is more likely but the first condition is less likely to be satisfied. The reason is that, when conditioning on more genetic variants, we are more likely to block the pleiotropic paths to Y but we are also more likely to block the path between the instrument and the causal variant.

For now, we will continue our discussion under the assumption that an appropriate set $\mathcal{B}$ can be chosen. In Section 5.3.5, we will return to this and describe a simple construction of $\mathcal{B}$ using Markov properties in Haldane's model for meiosis.

5.3.3 Hypothesis testing

We are now ready to describe the randomization-based inference for within-family Mendelian randomization. We will focus on the simplest case where a single genetic variant from the offspring's maternally-inherited haplotype is used as an instrumental variable and defer the discussion on multiple instruments to Section 5.3.6.

Suppose we observe N trios of parent and offspring and within each trio the observed variables can be described by the causal diagram described in Section 5.3.1. Let $i \in \{1, \dots, N\}$ be the index of the trio. Following Rosenbaum and Rubin (1983), we define the *propensity score* of the instrument Z_{ij}^m at locus j of individual i as

$$\pi_{ij}^m = \mathbb{P}(Z_{ij}^m = 1 \mid \boldsymbol{M}_{ij}^{mf}, \boldsymbol{V}_{i\mathcal{B}}^m) \tag{5.5}$$

where $\mathcal{B} \subseteq \mathcal{J}$ is an appropriately chosen set of loci that satisfies the conditions in Theorem 5.1. In words, π_{ij}^m describes the randomization distribution of the haplotype Z_{ij}^m conditional on a set of parental and offspring haplotypes or genotypes. For now, we will treat π_{ij}^m as known.

Let us consider the following model for the potential outcomes that assumes a constant treatment effect β:

$$Y_i(d) = Y_i(0) + \beta d \text{ for all } d \in \mathcal{D} \text{ and } i = 1, \dots, N. \tag{5.6}$$

Let $\mathcal{F} = \{Y_i(0) \colon i = 1, \dots, N\}$ denote the collection of potential outcomes for all individuals i under no exposure $d = 0$. Our goal is to test null hypotheses of the form

$$H_0 \colon \beta = \beta_0, \;\; H_1 \colon \beta \neq \beta_0 \tag{5.7}$$

where β_0 is some hypothetical value of the causal effect. If the null hypothesis is true, equation (5.6) implies that the potential outcome under no exposure ($d = 0$) can be identified from the observed data by

$$Y_i(0) = Y_i(D_i) - \beta_0 D_i = Y_i - \beta_0 D_i.$$

For ease of notation, let $Q_i(\beta_0) = Y_i - \beta_0 D_i$ be the adjusted outcome.

Theorem 5.2 and the model (5.6) imply that we are essentially testing the following conditional independence:

$$\begin{aligned} H_0 &\colon Z_{ij}^m \perp\!\!\!\perp Q_i(\beta_0) \mid (\boldsymbol{M}_{ij}^{mf}, \boldsymbol{V}_{i\mathcal{B}}^m), \\ H_1 &\colon Z_{ij}^m \not\perp\!\!\!\perp Q_i(\beta_0) \mid (\boldsymbol{M}_{ij}^{mf}, \boldsymbol{V}_{i\mathcal{B}}^m). \end{aligned} \tag{5.8}$$

Let $\boldsymbol{Z}_j^m = (Z_{1j}^m, \dots, Z_{Nj}^m)$ and similarly define other vector-valued genotypes. Suppose we have selected a test statistic $T(\boldsymbol{Z}_j^m \mid \mathcal{F})$ whose dependence on $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m)$ is implicit. For example,

this could be the coefficient from a regression of the adjusted outcome on the instrument. The randomization-based p-value for H_0 can then be written as

$$\begin{aligned} P(\boldsymbol{Z}_j^m \mid \mathcal{F}) &= \tilde{\mathbb{P}}\big(T(\tilde{\boldsymbol{Z}}_j^m \mid \mathcal{F}) \leq T(\boldsymbol{Z}_j^m \mid \mathcal{F})\big) \\ &= \sum_{\tilde{\boldsymbol{z}}^m \in \{0,1\}^N} I\big\{T(\tilde{\boldsymbol{z}}_j^m \mid \mathcal{F}) \leq T(\boldsymbol{Z}_j^m \mid \mathcal{F})\big\} \times \\ &\quad \times \prod_{i=1}^{N} (\pi_{ij}^m)^{\tilde{z}_i} (1 - \pi_{ij}^m)^{1-\tilde{z}_i}, \end{aligned} \tag{5.9}$$

where $I\{\cdot\}$ is the indicator function, $\tilde{\boldsymbol{Z}}_j^m$ denotes an independent, random draw from (5.5) and $\tilde{\mathbb{P}}$ denotes its probability distribution. Given the propensity score and the null hypothesis, this p-value can be computed exactly by enumerating over all 2^N possible values of $\tilde{\boldsymbol{Z}}^m$, or using a Monte Carlo approximation by drawing a large sample of $\tilde{\boldsymbol{Z}}^m$ using the propensity scores $\boldsymbol{\pi}_j^m$; see Algorithm 1 for the pseudo-code. It is straightforward to replace the haplotype Z_{ij}^m with the genotype Z_{ij}; the randomization distribution of $Z_{ij} \in \{0, 1, 2\}$ is a simple function of π_{ij}^m and π_{ij}^f since meioses in the mother and father are independent.

Equation (5.9) highlights the importance of the vector $\boldsymbol{\pi}_j^m$ of propensity scores in randomization inference. However, $\boldsymbol{\pi}_j^m$ describes a biochemical process occurring in the human body which is not precisely known to, or controlled by, us. Therefore, the best we can do is perform *almost exact* inference by replacing $\boldsymbol{\pi}_j^m$ with an accurate approximation. The model we use in this paper is Haldane's hidden Markov model described in Appendix C.1. As discussed in Section 5.2.2 our method is modular in the sense that more sophisticated meiosis models can easily be substituted as the randomization distribution; see Broman and Weber (2000) and Otto and Payseur (2019) for discussion and comparison of alternative models.

Remark 5.2. One case we did not consider is multi-levelled exposures where

$$Y_i = Y_i(0) + \sum_{t=1}^{T} \beta(t) D_i(t)$$

and $D_i(t) \in \{D_i(1), \ldots, D_i(T)\}$ is a collection of mutually exclusive binary exposures (Newey and Stouli, 2022). Multi-levelled exposures arise naturally in many Mendelian randomization applications, for example, alcohol consumption is often measured in several categories of standardized drinks per day. We could test the null $H_0 : \beta(1) = \beta_0(1), \ldots, \beta(T) = \beta_0(T)$, but it is unclear when the instrument Z_{ij}^m induces sufficient variation in $D_i(1), \ldots, D_i(T)$ to reject this null. We leave this extension for future work.

5.3.4 Choice of test statistic

Our randomization test retains the nominal size under the null hypothesis, regardless of the choice of the test statistic. Nonetheless, a well chosen statistic may substantially increase

Algorithm 1: Almost exact test

Compute the test statistic on the observed data $t = T(\boldsymbol{Z}_j^m \mid \mathcal{F})$;
for $k = 1, \ldots, K$ **do**
  Sample a counterfactual instrument $\tilde{\boldsymbol{Z}}_j^m$ from the randomization distribution (e.g. using Theorem C.1 in Appendix C.1 based on Haldane's model);
  Compute the test statistic using the counterfactual instrument $\tilde{t}_k = T(\tilde{\boldsymbol{Z}}_j^m \mid \mathcal{F})$;
end
Compute an approximation to the p-value in Equation (5.9) via the proportion of $\tilde{t}_1, \ldots, \tilde{t}_K$ which are larger than t:

$$\hat{P}(\boldsymbol{Z}_j^m \mid \mathcal{F}) = \frac{|\{k \colon t \leq \tilde{t}_k\}|}{K}.$$

the power of the test. A practical challenge is that the adjustment set $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}})$ may be high-dimensional and highly correlated, and their role in designing the test statistic is unclear. We propose to use the following "clever covariate" in the test statistic:

$$X_{ij}^m = \frac{Z_{ij}^m}{\pi_{ij}^m} - \frac{1 - Z_{ij}^m}{1 - \pi_{ij}^m}.$$

We may then use the weighted difference-in-means statistic

$$T(\boldsymbol{Z}_j^m \mid \mathcal{F}) = \sum_{i=1}^{N} Q_i(\beta_0) X_{ij}^m$$

or the F-statistic in a linear regression of $Q_i(\beta_0)$ on Z_{ij}^m and X_{ij}^m. A simulation example in Section 5.4.2 shows that using this clever covariate can increase the power of the test dramatically.

The idea of using a "clever covariate" is proposed in Scharfstein, Rotnitzky, and Robins (1999) and Rose and Laan (2008) and is commonly used in semiparametric estimators of the average treatment effect. Heuristically, the clever covariate exploits the so-called "central role" of the propensity score (Rosenbaum and Rubin, 1983)

$$Y_i(d) \perp\!\!\!\perp Z_{ij}^m \mid \pi_{ij}^m,$$

provided that $0 < \pi_{ij}^m < 1$. Thus, the propensity score π_{ij}^m may be viewed as a one-dimensional summary of the sufficient adjustment set $(\boldsymbol{M}_{ij}^{mf}, \boldsymbol{V}_{i\mathcal{B}})$ and is particularly convenient here because it can be directly computed from a meiosis model.

5.3.5 Simplification via Markovian structure

Thus far, we have sidestepped the issue of choosing the adjustment set $\boldsymbol{V}_{\mathcal{B}}$ and computing the propensity score. Conditional independencies implied by Haldane's meiosis model allows us to greatly simplify the sufficient confounder adjustment set. We explain this below.

The conditions in Theorem 5.1 are trivially satisfied with $\mathcal{B} = \emptyset$ if $\mathcal{J}_y = \emptyset$ and $\mathcal{J}_d \neq \emptyset$, i.e., all causal variants of Y on this chromosome only affect Y through D. However, this is a rather unlikely situation. More often, we need to condition on some variants to block the pleiotropic paths (such as Z_3 in the working example in Figure 5.2). To this end, we can utilize the Markovian structure on the meiosis indicators $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$ implied by Haldane's model. Roughly speaking, such structure implies that

$$Z_j \perp\!\!\!\perp Z_l \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{F}_j^{mf}, \boldsymbol{V}_k \text{ for all } j < k < l,$$

if there are no mutations and mother's genotype at locus k is heterozygous (i.e. $M_k^f \neq M_k^m$).

More generally, let b_1 and b_2 ($b_1 < j < b_2$) be two heterozygous loci in the mother's genome, i.e., $M_{b_1}^f \neq M_{b_1}^m$ and $M_{b_2}^f \neq M_{b_2}^m$. Let $\mathcal{A} = \{b_1 + 1, \ldots, b_2 - 1\}$ be the set of loci between b_1 and b_2, which of course contains the locus j of interest.

Theorem 5.2. *Consider the setting in Theorem 5.1 and suppose*

1. *The meiosis indicator process is a Markov chain so that $U_j^m \perp\!\!\!\perp U_l^m \mid U_k^m$ for all $j < k < l$;*
2. *There are no mutations: $\epsilon = 0$.*

Then Z_j^m is a valid instrumental variable conditional on $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\{b_1,b_2\}}^m)$ if the following conditions are true:

3. *$\mathcal{A} \cap \mathcal{J}_d \neq \emptyset$;*
4. *$\mathcal{A} \cap \mathcal{J}_y = \emptyset$.*

Proof. Because there are no mutations and M_{b_1} and M_{b_2} are heterozygous, we can uniquely determine $U_{b_1}^m$ and $U_{b_2}^m$ from $\boldsymbol{V}_{\{b_1,b_2\}}^m$. By the assumed Markovian structure, this means that

$$Z_j^m \perp\!\!\!\perp Z_l^m \mid \boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\{b_1,b_2\}}^m \text{ for all } j < b_1 \text{ or } j > b_2.$$

Thus, the last two conditions in Theorem 5.2 imply the two conditions in Theorem 5.1. □

One can easily mirror the above result for using the paternal haplotype Z_j^f as an instrument variable. Furthermore, let b_1' and b_2' ($b_1' < j < b_2'$) be two heterozygous loci in the father's genome. Then it is easy to see that $Z_j = Z_j^m + Z_j^f$ is a valid instrument conditional on $(\boldsymbol{M}_j^{mf}, \boldsymbol{F}_j^{mf}, \boldsymbol{V}_{\{b_1,b_2\}}^m, \boldsymbol{V}_{\{b_1',b_2'\}}^f)$ if the last two conditions hold for the union $\mathcal{A} = \{\min(b_1, b_1') + 1, \ldots, \max(b_2, b_2') - 1\}$.

Under the setting in Theorem 5.2, we can partition the offspring genome into mutually independent subsets by conditioning on heterozygous parental genotypes. This partition is useful for constructing independent p-values when we have multiple instruments. Suppose we have a collection of genomic position $\mathcal{B} = \{b_1, \ldots, b_k\}$ that will be conditioned on and let

$\mathcal{A}_k = \{b_{k-1}+1, \ldots, b_k - 1\}$ be the loci in between (suppose $b_0 = 0$ and $b_{k+1} = p+1$). This indues the following partition of the chromosome:

$$\mathcal{J} = \mathcal{A}_1 \cup \{b_1\} \cup \mathcal{A}_2 \cup \{b_2\} \cup \ldots \cup \mathcal{A}_k \cup \{b_k\} \cup \mathcal{A}_{k+1}. \tag{5.10}$$

Proposition 5.3. *Suppose $M_j^m \neq M_j^f$ for all $j \in \mathcal{B}$. Then, under the first two assumptions in Theorem 5.2, we have*

$$Z_j^m \perp\!\!\!\perp Z_{j'}^m \mid (\boldsymbol{M}_j^{mf}, \boldsymbol{M}_{j'}^{mf}, \boldsymbol{V}_{\mathcal{B}}^m).$$

for any $j \in \mathcal{A}_l$ and $j' \in \mathcal{A}_{l'}$ such that $l \neq l'$.

Proof. The proof follows from an almost identical argument to Theorem 5.1. The assumption that $\epsilon = 0$ means that U_j^m is uniquely determined for all $j \in \mathcal{B}$ from $\boldsymbol{M}_j^{mf}$ and Z_j^m. Therefore the assumed Markovian structure implies that conditioning on $\boldsymbol{V}_{\mathcal{B}}^m$, along with the parental haplotypes $\boldsymbol{M}_j^{mf}$ and $\boldsymbol{M}_{j'}^{mf}$, then induces the conditional independence. □

5.3.6 Multiple instruments

Proposition 5.3 allows us to formalize the intuition that genetic instruments across the genome may provide independent pieces of evidence.

Corollary 5.1. *Suppose $j \in \mathcal{A}_l$, $j' \in \mathcal{A}_{l'}$, and $l \neq l'$. Then Z_j^m and $Z_{j'}^m$ are independent, valid instruments given $(\boldsymbol{M}_j^{mf}, \boldsymbol{M}_{j'}^{mf}, \boldsymbol{V}_{\mathcal{B}}^m)$ if*

1. *The first two assumptions of Theorem 5.2 hold;*
2. *$\mathcal{A}_l \cap \mathcal{J}_d \neq 0$ and $\mathcal{A}_{l'} \cap \mathcal{J}_d \neq 0$;*
3. *$\mathcal{A}_l \cap \mathcal{J}_y = 0$ and $\mathcal{A}_{l'} \cap \mathcal{J}_y = 0$.*

Corollary 5.1 says that any two instruments are valid and independent if they lie in separate regions in the partition (5.10) and each region contains at least one causal variant of the exposure and no pleiotropic variants (i.e. variants with a direct effect on Y not mediated by D).

As a direct application of this corollary, we can use standard procedures to combine randomization p-values obtained from different genetic instruments and test the null hypothesis (see e.g. Bretz, Hothorn, and Westfall, 2016). One such procedure is Fisher's method (Fisher, 1925): let $\{p_1, p_2, \ldots, p_k\}$ be a collection of independent p-values, then $-2\sum_{j=1}^k \log(p_j) \sim \chi^2_{2k}$ when all of the corresponding null hypotheses are true. We will use this method to aggregate p-values in the applied example in Appendix A.1.

As some instruments may violate the exclusion restriction, a more robust approach is to test the partial conjunction of the null hypotheses (Benjamini and Heller, 2008; Wang and Owen, 2019); loosely speaking, this means that we reject the causal null hypothesis only if quite a few genetic instruments appear to provide evidence against it.

It may be impossible in practice to separate closely linked instruments into partitions separated by a heterozygous variant, in which case the hypothesis (5.8) can be tested using $(Z_j^m, Z_{j'}^m)$ jointly. Corollary C.2 in Appendix C.2 derives the joint randomization distribution of a collection of instruments.

Remark 5.3. The multiple instruments case has parallels with the problem of overidentification in econometrics (Sargan, 1958; Sargan, 1988). Each instrument admits a valid confidence interval for β in Section 5.3.3 under the randomization of Z_j^m. If all instruments are valid, then the intersection of all confidence intervals should be non-empty with known probability. An empty intersection can be viewed as evidence that at least one instrument is invalid. See Diemer et al. (2020) for further exposition of this idea.

5.4 Simulation

5.4.1 Setup and illustration

In this section we explore the properties of our almost exact test via a numerical simulation. The set up of the simulation is described in detail in Appendix C.3. To summarize, we generate a dataset of 15,000 parent-offspring trios with a null effect of an exposure on an outcome (i.e. $\beta = 0$), both of which have variance one, and consider 5 genetic instruments on a chromosome with $p = 150$ loci. The instruments are non-causal markers for nearby causal variants and there are also pleiotropic variants in linkage disequilibrium with the instruments.

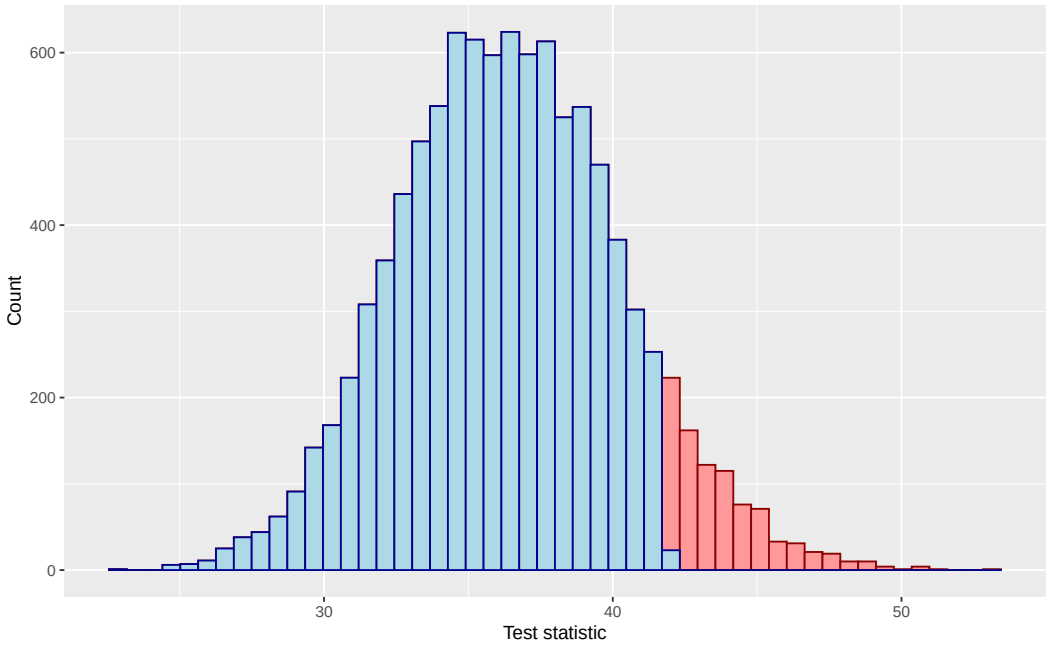
To make our setup more tangible, Table 5.3 shows the first 6 lines of observed and counterfactual data (in red) from the simulation for one of the instruments and corresponding parental haplotypes. We can see that individual 4 will provide almost no information for a test of the null hypothesis; both of her parents are homozygous so there is no randomization in her genotype outside of de novo mutations. Conversely, both of individual 1's parents are heterozygous so she could receive both major alleles, both minor alleles or one of each.

Suppose we wish to test the null hypothesis $H_0 : \beta = -0.3$. Column $\tilde{Z}_i$ in Table 5.3 shows a counterfactual draw of each individual's instrument conditional on the adjustment set given in Equation (C.9) in Appendix C.3, along with the adjusted outcome $Q_i(-0.3)$. Note that $\tilde{Z}_i$ is independent of $Q_i(-0.3)$ by construction, so the null hypothesis is necessarily satisfied for this counterfactual. As expected individual 4 has the same genotype in this counterfactual, however, individual 1 inherits both minor alleles in this case. Figure 5.4 plots a distribution of 10,000 counterfactual test statistics drawn under the null hypothesis. The test statistic is the F-statistic from a regression of the adjusted outcome on the instruments. The bars highlighted in red are larger than the observed test statistic, such that the almost exact p-value is around 0.13.

Table 5.3: First 6 rows of observed data from the simulation

i	Z_i	$\tilde{Z}_i$	M_i^m	M_i^f	F_i^m	F_i^f	D_i	Y_i	$Q_i(-0.3)$
1	1	**2**	1	0	1	0	1.11	0.73	1.06
2	0	**1**	1	0	0	0	0.83	-0.52	0.77
3	1	**1**	1	0	0	0	0.94	0.31	0.59
4	0	**0**	0	0	0	0	1.43	3.30	3.73
5	0	**0**	0	0	0	0	0.15	1.34	1.38
6	0	**0**	0	0	0	0	-0.14	1.60	1.56

Figure 5.4: Histogram of 10,000 test statistics under the exact null hypothesis $H_0 : \beta = -0.3$



5.4.2 Power

We now use the simulation to study the power of the randomization test, assuming a correct adjustment set is used (see Equation (C.9) in Appendix C.3). As the haplotypes are simulated according to Haldane's meiosis model, the randomization test should be exact. This is verified by the near-uniform distributions of the p-values for testing the correct null hypothesis $H_0 : \beta = 0$ with three different test statistics in the top panels of Figure 5.5.

The histograms in the bottom panels of Figure 5.5 depict the distribution of p-values for a test of a false null hypothesis $H_0 : \beta = 0.5$. The power of the test varies significantly according to the choices of test statistic. The simple F-statistic based on a linear regression of the adjusted outcome on the instruments (test statistic 1) has almost no power, while the test statistic obtained from the same model but with the propensity score included as a clever covariate (test statistic 2) has a reasonable power of about 0.52.

Figure 5.6 expands upon the previous figure by plotting a power curve for each test statistic. We can see that test statistic 1 has almost no power between $\beta_0 = 0$ and $\beta_0 = 1$. Since test statistic 1 is unconditional on the adjustment set, resampled offspring haplotypes retain their correlation with the confounders via the parental haplotypes. This can cause under-rejection of false null hypotheses around the unconditional instrumental variable estimator. In this simulation, the Anderson-Rubin 95% confidence interval is 0.64–0.89, which aligns with the region of under-rejection.

Test statistic 2, on the other hand, conditions on the confounders via a clever covariate. It has a power curve that is centred on the true null $\beta_0 = 0$ and has significantly improved power in the region between $\beta_0 = 0$ and $\beta_0 = 1$. However, it should be noted that test statistic 2 is not uniformly more powerful than test statistic 1.

5.5 Applied example

5.5.1 Preliminaries

We illustrate our approach with a pair of negative and positive controls using the Avon Longitudinal Study of Parents and Children (ALSPAC). Our dataset consists of 6,222 mother-child duos from ALSPAC, a longitudinal cohort initially comprising pregnant women resident in Avon, UK with expected dates of delivery from 1 April 1991 to 31 December 1992. The initial sample consisted of 14,676 fetuses, resulting in 14,062 live births and 13,988 children who were alive at 1 year of age. In subsequent years, mothers, children and occasionally partners attended several waves of questionnaires and clinic visits, including genotyping. For a more thorough cohort description, see Boyd et al., 2013 and Fraser et al., 2013.[1]

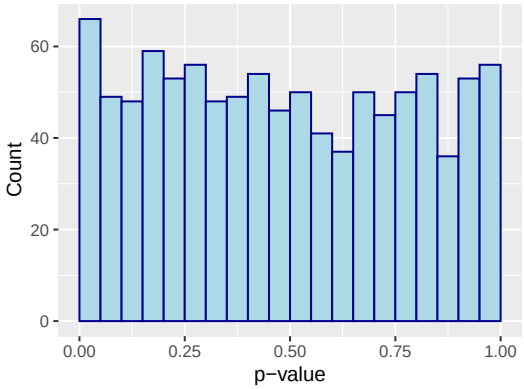
The negative control is the effect of child's BMI at age 7 on mother's BMI pre-pregnancy. Dynastic effects, as depicted in Figure 5.3e, could induce a spurious correlation between child's BMI-associated variants and their mother's BMI pre-pregnancy. Blocking this backdoor path is crucial for reliable causal inference. The positive control is the effect of child's BMI at age 7 on a simulated, noisy version of itself. We vary the proportion of the outcome that is attributable to noise to assess the power of our test.

5.5.2 Data processing

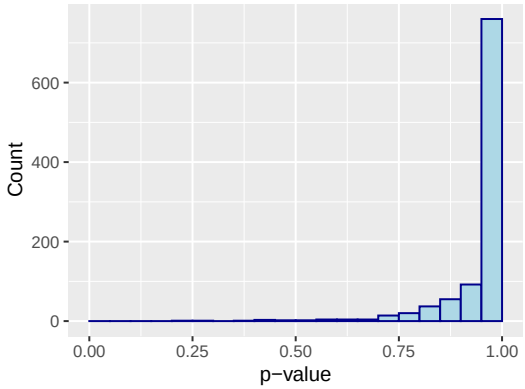
We use ALSPAC genotype data generated using the Illumina HumanHap550 chip (for children) and Illumina human660W chip (for mothers) and imputed to the 1000 Genomes reference panel. We remove SNPs with missingness of more than 5% and minor allele frequency of less than 1%. Haplotypes are phased using the `SHAPEIT2` software with the `duoHMM` flag, which ensures that phased haplotypes are consistent with known pedigrees in the sample. We obtain recombination

[1]Please note that the study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool (`https://www.bristol.ac.uk/alspac/researchers/our-data/`).

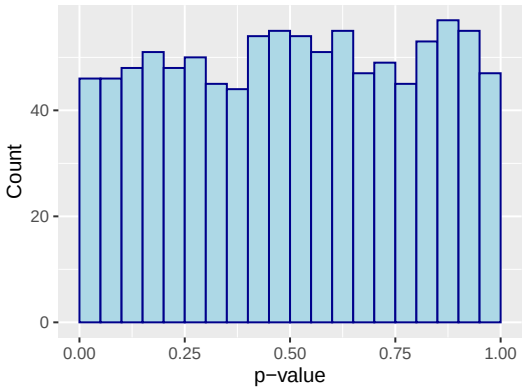
Figure 5.5: Histograms of 1,000 p-values for several null hypotheses and test statistics. Test statistic 1 is the F-statistic from a linear regression of the adjusted outcome on the instruments. Test statistic 2 is similar but includes the propensity scores for each instrument as covariates. Test statistic 3 includes only the parental genotypes for each instrument as covariates.



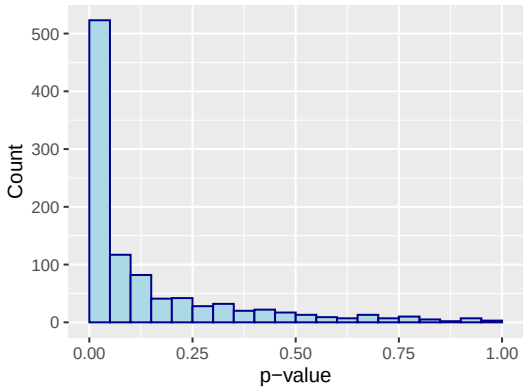
(a) $H_0 : \beta = 0$ and test statistic 1



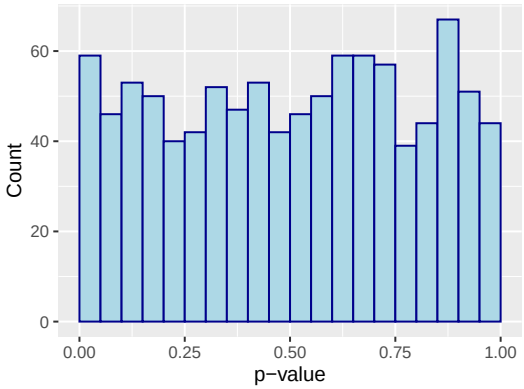
(b) $H_0 : \beta = 0.5$ and test statistic 1



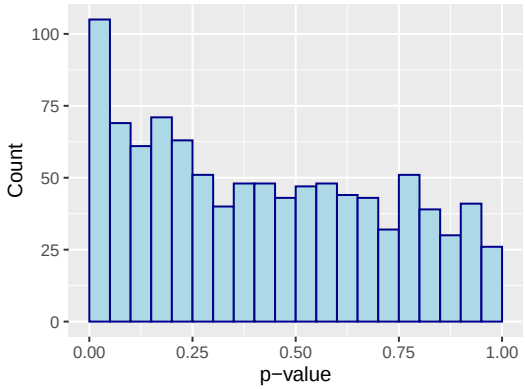
(c) $H_0 : \beta = 0$ and test statistic 2



(d) $H_0 : \beta = 0.5$ and test statistic 2

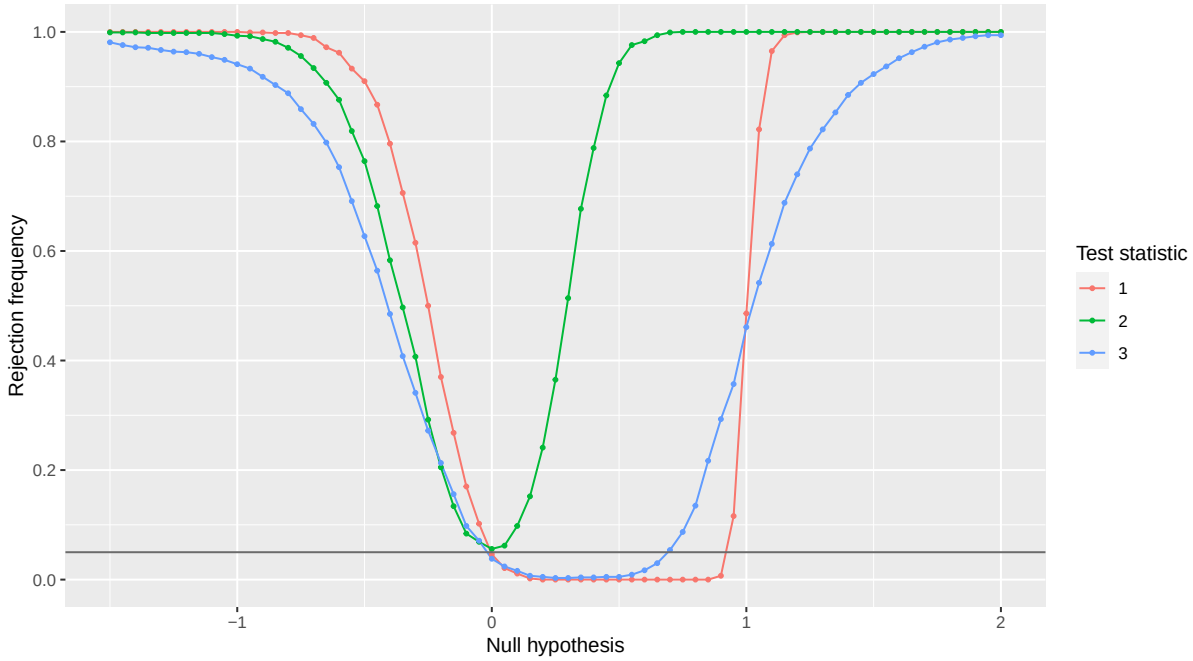


(e) $H_0 : \beta = 0$ and test statistic 3



(f) $H_0 : \beta = 0.5$ and test statistic 3

Figure 5.6: Power curves for the three choices of test statistic. Test statistic 1 is the F-statistic from a linear regression of the adjusted outcome on the instruments. Test statistic 2 includes the propensity scores for each instrument as covariates. Test statistic 3 includes the parental haplotypes as covariates. Each point on the figure is the rejection frequency over 1,000 replications.



probabilities from the 1000 Genomes genetic map file on Genome Reference Consortium Human Build 37.

Our instruments are selected from the genome-wide association study (GWAS) of Vogelezang et al., 2020, which identifies 25 genetic variants for childhood BMI, including 2 novel loci located close to *NEDD4L* and *SLC45A3*. Of the genome-wide significant variants in the discovery sample, we select 11 with a p-value of less than 0.001 in the replication sample. ALSPAC is included in the discovery sample, so independent replication is important for avoiding spurious associations with the exposure. Two of our instruments, rs571312 and rs76227980, are located close together near *MC4R* and need to be tested jointly. We exclude rs62107261 because it is not contained in the 1000 Genomes genetic map file. We condition on all variants outside of a 500 kilobase window around each instrument.

5.5.3 Results

Tables 5.4 and 5.5 show results for the negative and positive controls, respectively. The last row of each table shows the p-value from Fisher's method aggregated across all independent p-values. The aggregated p-value for the negative control is 0.21, indicating little evidence

Table 5.4: Results from the ALSPAC negative control example.

Instrument (rsID)	Chromosome	Proximal gene	P-value
rs11676272	2	*ADCY3*	0.45
rs7138803	12	*BCDIN3D*	0.55
rs939584	2	*TMEM18*	0.39
rs17817449	16	*FTO*	0.06
rs12042908	1	*TNNI3K*	0.35
rs543874	1	*SEC16B*	0.07
rs56133711	11	*BDNF*	0.59
rs571312, rs76227980	18	*MC4R*	0.48
rs12641981	4	*GNPDA2*	0.62
rs1094647	1	*SLC45A3*	0.19
Fisher's method			0.21

Table 5.5: Results from the ALSPAC positive control example. (Chr. = chromosome)

Instrument (rsID)	Chr.	Gene	P-value for noise of		
			10%	20%	50%
rs11676272	2	*ADCY3*	0.01	0.01	0.01
rs7138803	12	*BCDIN3D*	0.01	0.01	0.01
rs939584	2	*TMEM18*	0.98	0.95	0.88
rs17817449	16	*FTO*	0.33	0.35	0.44
rs12042908	1	*TNNI3K*	0.77	0.79	0.85
rs543874	1	*SEC16B*	0.48	0.64	0.92
rs56133711	11	*BDNF*	0.12	0.14	0.25
rs571312, rs76227980	18	*MC4R*	0.31	0.39	0.63
rs12641981	4	*GNPDA2*	0.49	0.56	0.76
rs1094647	1	*SLC45A3*	0.23	0.25	0.35
Fisher's method			0.03	0.05	0.16

against the null. The aggregated p-values for the positive control range from 0.03 (when 10% of the variance of the simulated outcome is noise) to 0.16 (when 50% of the variance of the simulated outcome is noise). This indicates weak evidence against the null even when the effect is quite strong.

We can also compare the results in Tables 5.4 and 5.5 with per-instrument p-values obtained from two-stage least squares (2SLS) using the same offspring haplotypes as instruments, unconditional on parental or other offspring haplotypes. For the negative control, the p-value from Fisher's method is 0.02, indicating some evidence against the null. This is expected, given that the backdoor paths remain unblocked. For the positive control, the p-values from Fisher's method range from less than 10^{-20} (when 10% of the variance of the simulated outcome is noise) to 4.5×10^{-11} (when 50% of the variance of the simulated outcome is noise). This indicates that

the unconditional analysis has significantly more power to detect non-zero effects compared to our "almost exact" test. We discuss potential reasons for, and implications of, this low power in Section 5.6

5.6 Discussion

We have presented an almost exact approach to within-family MR, which has a number of conceptual and practical advantages over model-based approaches to MR using population GWAS data. However, the applied example in Appendix A.1 demonstrates that power may be limited relative to conventional MR analyses in unrelated individuals. Since our test leverages the precise amount of information available in a single meiosis, this suggests that MR in unrelated individuals is drawing power from elsewhere.

Besides the obvious distinction that conventional MR analyses are model-based (and thus are not robust to model misspecification), another likely reason for the large difference in the empirical results is that MR in unrelated individuals use randomness in meioses across many generations. For example, an offspring with parents who are homozygous for the non-effect allele offers no power in our test, since their genotype will not vary across meioses. However, if we assume that genotypes are randomly distributed at the population level (as in MR studies with unrelated individuals), that same offspring can act as a comparator for individuals with the effect allele. Brumpton et al., 2020 corroborate this loss of power for their within-family method, but do not elaborate on the broader implications for how Mendelian randomization is typically justified. It would be extremely valuable for the MR literature to discuss the extent to which Mendelian inheritance across multiple generations is driving the power behind existing results, as such uncontrolled randomness may introduce bias when there are strong dynastic effects and natural selection.

Continuing the discussion on multiple instruments in Section 5.3.6, our approach closely resembles the usage of evidence factors in observational studies as advocated by Rosenbaum (2010) and Rosenbaum (2021) and Zhao, Lee, et al. (2022). Using (conditionally) independent instruments in different genomic regions may also be viewed as a form of triangulation to improve causal inference (Lawlor, Tilling, and Davey Smith, 2017). Although all the p-values are obtained using the same study design, different genetic variants may influence the exposure through different biological mechanisms and the fact that they provide corroborating evidence strengthens the causal conclusion.

We must also return to the problem of transmission ratio distortion (TRD) discussed in Section 5.2.2. TRD violates the assumptions of our meiosis model that alleles are (unconditionally) passed from parents to offspring at the Mendelian rate of 50%. We could represent TRD in our causal model in Figure 5.2 via an arrow from the gametes $(\boldsymbol{Z}^m, \boldsymbol{Z}^f)$ to the mating indicator S. This indicates that the gametes themselves influence survival of their corresponding zygote

to term. If our putative instrument Z_1^m is in linkage with any variant exhibiting TRD, then this invalidates it as an instrument. Suppose Z_3^m exhibits TRD, then this opens collider paths via the parental phenotypes C^m and C^f, for example, $Y(d) \leftarrow C^m \rightarrow \boxed{S} \leftarrow Z_3^m \leftarrow \boldsymbol{U}^m \rightarrow Z_1^m$. The intuition is that parental phenotypes related to the likelihood of mating become associated with offspring variants related to the likelihood of offspring survival. Within our causal model, this pathway can be closed by conditioning on Z_3^m, with unconditioned variants obeying the meiosis model. If any unconditioned variants exhibit TRD, then this bias will remain and our meiosis model will incorrectly describe the inheritance patterns of any linked variants, resulting in an erroneous randomization distribution. Expanding resources of parent-offspring data may allow us to test the prevalence of transmission ratio distortion, which will help to inform the reasonableness of maintaining Mendel's First Law in our meiosis and fertilization model.

Chapter 6

Discussion

6.1 Recap

The motivation underlying all three chapters in my thesis is that causal inference always rests on untestable assumptions and interrogating, or assessing sensitivity to, those assumptions can lead to more reliable and interpretable scientific results. Although the chapters tackle seemingly disconnected problems, this theme unites them. In what follows, I briefly review the main takeaways of each chapter. I then discuss the implications of my thesis work for two stakeholders – methodologists and practitioners – and potential areas for future work.

In Chapter 3, I consider the problem of selection bias in large population cohort studies, demonstrating that causal inference in this setting can be highly sensitive to sample selection patterns. I also show that relatively few population-level auxiliary constraints are needed to significantly improve the precision of partially-identified intervals for inverse probability weighted estimators. In the process of developing this sensitivity analysis for selection bias, I developed a flexible procedure for conducting statistical inference in stochastic optimization problems with estimated constraint sets.

In Chapter 4, I explore the exclusion restriction violation that can occur in an instrumental variable analysis when the exposure measure is a coarsened approximation to some true latent exposure. I derived a simple expression for the resulting bias and proposed a sensitivity analysis for the effect estimate on the scale of the latent exposure, with the sensitivity parameter being the genetic variance of the latent exposure.

Finally, in Chapter 5, I describe how Mendelian randomization (MR) studies in unrelated individuals make a number of strong assumptions about the population-level distribution of genotypes and the absence of biasing pathways from population structure, parental phenotypes and assortative mating, among others. The "almost exact" test proposed in that chapter is based on an explicit model for Mendelian inheritance, embedded within a broader causal model for MR itself. I use the causal model to identify sufficient confounder adjustment sets and I use the model for meiosis and fertilization to derive a randomization distribution for

offspring genotypes conditional on this adjustment set. This approach to sensitivity analysis relaxes the need for many of the assumptions made in MR studies among unrelated individuals. Disagreement between my test and these existing studies could therefore indicate a violation of one or more of these assumptions.

6.2 Extended discussion for methodologists

6.2.1 Chapter 3

One of the central contributions of Chapter 3 is developing a procedure for statistical inference in a difficult class of stochastic optimization problems where the constraint set must also be estimated. This has implications beyond the selection bias example considered in the chapter. For example, Duarte et al. (2021) propose an "automated" approach to partial identification with discrete data, specifically, a procedure for generating asymptotically sharp bounds given a desired causal estimand and causal graph. The authors note that statistical inference over the resulting bounds remains to be developed. The procedure in Chapter 3 could be applied in this setting, allowing valid statistical inference in a broad range of causal inference problems.

Chapter 3 also contributes to the literature on inverse probability weighting for sample selection. Chattopadhyay, Hase, and Zubizarreta (2020) discuss two approaches for weighting: balancing and modelling. Modelling weights attempt to explicitly model the selection process using observed covariates, for example, estimating a logistic regression with sample selection ($S \in \{0, 1\}$) as the response. By contrast, balancing weights are chosen to minimize the difference in moments of the covariate distribution (typically means or variances) between the selected and non-selected group. The authors demonstrate conditions under which these two approaches are equivalent.

A feature of Chapter 3 is that it utilizes both approaches: a model is specified for the weights and the model parameters are selected subject to some balancing conditions. Other authors have considered this formulation (Nevo, 2003; Signorovitch et al., 2012) but I frame it as a partial identification problem. An advantage of the partial identification formulation is flexibility. Since I am not concerned with exactly identifying the weights, I can use a more flexible model for the probability weights (e.g., include interactions or higher order terms that are unlikely to be available as population moments) and a more diverse set of auxiliary information (e.g., shape constraints on the covariate distribution). A drawback of this flexibility is that I do not formally characterize the information content of each constraint – it is unclear which constraints are informative, in the sense of narrowing the identified interval, and which are not.

6.2.2 Chapter 4

As discussed in Section 2.1.8, Chapter 4 is not the only contribution to the problem of coarsening bias in instrumental variable analyses. Angrist and Imbens (1995) introduce the

common interpretation of the Wald estimand with a multi-valued exposure as a weighted average of the complier average causal effect at each exposure level. In doing so, they discuss the bias that arises when a multi-valued exposure is miscoded as a binary exposure. The authors demonstrate that this results in a multiplicative bias that will tend to inflate effect estimates (see Section 2.1.8 for a more formal discussion). This is consistent with my bias formula in Section 4.3.1 under a linear index relationship between the latent continuous exposure and coarsened exposure. In my set-up with a continuously-distributed latent exposure (as opposed to a multi-valued discrete latent exposure), coarsening can also induce a downward bias.

I did not explore in much detail the connection between my sensitivity analysis and the literature on coefficients of determination on the liability scale. The sensitivity parameter θ in my sensitivity analysis is the genetic variance of the latent exposure. Lee, Goddard, et al. (2012) propose a method of estimating this quantity under the same Falconer (1965) liability-threshold model that motivated my work. It is likely that an estimator for β could be conceived in which θ is replaced with an appropriate estimator. This would sidestep the need to view this as a sensitivity analysis, provided the estimator for θ is reliable.

6.2.3 Chapter 5

Chapter 5 reveals some novel – often subtle – characteristics of the MR design that could serve as an impetus for methodological work in MR and population genetics more generally. For example, nearly determined ancestry bias (Table 5.2) can arise in my causal model when we condition on offspring genotypes without conditioning on the corresponding parental genotypes. This type of conditioning is common in genetic fine-mapping methods, which attempt to locate causal variants for diseases and other phenotypes (Wang, Sarkar, et al., 2020). In the presence of nearly determined ancestry bias, existing approaches to fine-mapping could lead to erroneous inferences about causal variants, although this has not yet been recognized, to my knowledge.

Another characteristic that was not explicitly explored in this chapter is the unique challenge implied by our causal model in diverse and admixed samples. *Admixture* is said to occur when two previously divergent or isolated populations produce offspring (Gopalan et al., 2022). The majority of studies in human population genetics have been conducted in samples of white Europeans, ostensibly for ease of data collection and to minimize bias from population structure. However, there is growing recognition of the need for studies in non-European populations (Sirugo, Williams, and Tishkoff, 2019; Caliebe et al., 2022). This recognition is occurring in MR studies as well (Zollner et al., 2022).

Admixture can induce long-acting linkage disequilibrium along a chromosome. This can be illustrated with a simple derivation using the HMM in Chapter 5. Suppose an offspring's mother is admixed, with her mother from population A and her father from population B. Suppose we also have two variants indexed by $j \in \{1, 2\}$ on the same chromosome, with the probability of a crossover occurring in a single meiosis given by p. We can write $E(M_j^m) = f_j^A$

and $E(M_j^f) = f_j^B$, where f denotes population allele frequencies. It follows from the HMM in Appendix C.1 that

$$\text{pr}(Z_2^m = 1 \mid Z_1^m = 1) = \left[f_2^A(1-p) + f_2^B p\right] \frac{f_1^A}{f_1^A + f_1^B} + \left[f_2^A p + f_2^B(1-p)\right] \frac{f_1^B}{f_1^A + f_1^B}.$$

This expression is a generalization of the one in Price et al. (2008), who consider two perfectly-linked variants such that $p = 0$. If $f_1^A = f_1^B$ then $\text{pr}(Z_2^m = 1 \mid Z_1^m = 1) = \frac{1}{2} f_2^A + \frac{1}{2} f_2^B$, indicating that Z_1^m and Z_2^m are independent. If $f_2^A = f_2^B = f_2$ then $\text{pr}(Z_2^m = 1 \mid Z_1^m = 1) = f_2$, also indicating independence. If the variants are far apart, such that $p \approx \frac{1}{2}$ then $\text{pr}(Z_2^m = 1 \mid Z_1^m = 1) \approx \frac{1}{2} f_2^A + \frac{1}{2} f_2^B$, once again indicating independence. This suggests that admixture resulting in differential allele frequencies, along with some correlation between the meiosis indicators (captured by p), is necessary to induce this form of linkage disequilibrium. Since p only converges to one half over long distances (for example, two variants that are 10 Mb apart will have $p \approx 0.1$), this suggests that linkage disequilibrium from admixture can extend far along a chromosome. This is problematic for MR studies because it can induce correlation between the instrument and potentially many pleiotropic variants located along the chromosome. Conditioning on parental genotypes is sufficient to address this bias, but methodologists developing approaches for MR in unrelated admixed samples should be cognizant of the need to consider pleiotropy induced by long-acting linkage disequilibrium.

6.3 Extended discussion for practitioners

Each chapter in my thesis proposes a sensitivity analysis for a problem of relevance to applied researchers, and it is my hope that my proposals will receive some uptake. I aim to provide a compelling case for addressing each of the three methodological problems and to remove barriers to implementing my proposed solution. To this end, I have written open source R packages for two of my chapters: `selectioninterval` (Chapter 3) and `almostexactmr` (Chapter 5). These can be freely installed and modified by users. I also released a GitHub repository containing all scripts used to generate the results in Chapter 4 (`https://github.com/matt-tudball/mrlat_replication`). Alongside software packages and scripts, each chapter contains several applications of the proposed methods to real data.

6.3.1 Chapter 3

Chapter 3 can be viewed in the broader push for awareness of the problem of selection bias in applied research. This was highlighted by Griffith et al. (2020), who note that many epidemiological studies of COVID-19 rely on highly selected samples, such as the influential COVID Symptom Tracker Study (Menni et al., 2020). Recent studies have also been raising concerns about the representativeness of large population cohorts such as UK Biobank (Fry et al., 2017; Munafò et al., 2018; Huang, 2021). Given the growing importance of rigorously addressing

selection bias in applied studies, Chapter 3 contributes by proposing a way of quantifying selection bias even in the absence of data on non-selected observations or insufficiently many population moments to identify balancing weights (Chattopadhyay, Hase, and Zubizarreta, 2020). This is most likely to be useful in self-selected surveys or cohorts where there is little data on individuals who deny to participate.

A practical challenge of implementing my sensitivity analysis is selection of the sensitivity parameters. This is a challenge common to many sensitivity analyses (Thompson and Arah, 2014; Smith and VanderWeele, 2019; Cinelli and Hazlett, 2020) and there is no optimal choice. Smith (2020) advocates selecting sensitivity parameters which shift a non-zero estimate toward the null. An advantage of this conceptualization is that it yields sensitivity parameters which result in a qualitative change in a study's conclusions. From there, the practitioner can decide whether sensitivity parameters of that magnitude are plausible or not. Cinelli and Hazlett (2020) instead advocate choosing sensitivity parameters based on domain knowledge, that is, referring to previous studies or consulting domain experts. An advantage of this conceptualization is that the sensitivity parameters are more justifiable and take advantage of existing knowledge. The two approaches are, of course, not mutually exclusive.

A feature of the sensitivity analysis in Chapter 3 is that there are two avenues for reducing the width of the identified interval: decreasing the sensitivity parameters and adding additional constraints. I argue that the most compelling sensitivity analysis is one in which the sensitivity parameters are set sufficiently large to capture most plausible selection patterns, but the interval is tightened using auxiliary data – in the UK, this can often be obtained from the Office of National Statistics (`https://www.ons.gov.uk/`) which publishes the census and other population-level metrics. This approach allows us to utilize existing information about the population in a formal way.

A practical problem with my sensitivity analysis is that the optimization problem (3.11) can theoretically induce very extreme weights. In practice, this is often handled via trimming, where observations with extreme weights are removed from the analysis (Lee, Lessler, and Stuart, 2011), such that the sample is given by $\{i = 1, \dots, N : e(W_i; \theta) \in [\alpha, 1 - \alpha]\}$ for some specified $\alpha > 0$. Crump et al. (2009) propose a more principled approach of only using observations whose covariates lie within a subspace of their support such that the variance of the estimator is minimized. Under some conditions, this is equivalent to trimming. It is unclear how to utilize weight trimming within my sensitivity analysis. It is tempting to follow a naive approach of including constraints such that the weights cannot be too extreme, however, this does not trim outlying observations; indeed, it could lead to artificially tight intervals by forcing the weights of outlying observations to lie within a certain range. I view the potential dependence of the identified interval on a few extreme weights as a drawback of my sensitivity analysis.

6.3.2 Chapter 4

Chapter 4 serves two purposes for practitioners. Firstly, it raises awareness about the problem of coarsening bias in MR studies. Much of the discussion about exclusion restriction violations in the MR literature has centred on pleiotropy (Hemani, Bowden, and Davey Smith, 2018), which occurs when a variant (or group of closely linked variants) exert biological effects on multiple phenotypes. Exclusion restriction violations related to coarsening, and measurement error more generally, have received considerably less attention. Secondly, Chapter 4 proposes a method for recovering an interpretable effect estimate on the scale of the latent exposure, under some strong structural assumptions.

Since publication of this chapter, a couple of studies have implemented my method to address coarsening bias. Lai et al. (2022) estimate the bidirectional effects of hypertension and gout. Given that both phenotypes are measured as a binary diagnosis variable, the authors apply my coarsening bias method to estimate effects on the liability scale. The authors select ranges for the sensitivity parameter θ^2 based on GWAS estimates of closely-related continuous phenotypes: blood pressure for hypertension and uric acid for gout. I view this as a clever approach that could be utilized in other applications of my method. Wang, Richardson, et al. (2022) estimate the effect of atrial fibrillation on a large number of phenotypes (referred to as a *phenome-wide* study). Similar to the previous paper, atrial fibrillation is measured as a binary diagnosis variable and the authors instead estimated effects on the liability scale, using θ^2 derived from Lee, Goddard, et al. (2012)'s coefficient of determination introduced in Section 4.3.2.

6.3.3 Chapter 5

Chapter 5 also serves a dual purpose for practitioners. Firstly, one of the primary motivations for this paper is the lack of a rigorous biologically-grounded justification for MR as a study design, and a clear description of the underlying assumptions. I view this chapter as being a useful pedagogical tool, particularly for econometricians, applied statisticians, and others with a technical background who would like to see a formal exposition of the assumptions and testing procedure. Secondly, it provides a statistical procedure for parent-offspring MR, with accompanying software package. It is increasingly being recognized that certain traits are more heavily influenced by patterns of ancestry and demography than others. For example, Howe, Nivard, et al. (2022) demonstrate that behavioural traits, such as educational attainment, exhibit greater attenuation from a within-sibship analysis than molecular traits, such as lipids. The authors argue that this attenuation is driven by blocking some of the backdoor paths described in Table 5.2. Researchers studying such traits could strengthen their analyses by demonstrating consistency between my robust, but less powerful, "almost exact" test and more conventional approaches in samples of unrelated individuals.

6.4 Future work

There are several promising avenues for expanding and generalizing the work in this thesis. Chapter 5 is restricted to parent-offspring designs, however, we could build upon parent-offspring principles to develop a general framework for causal inference in family-based studies containing arbitrary pedigrees (e.g., parent-offspring, siblings, cousins, grandparents). The observation that would guide my initial approach to this problem is that the genotypes of any two related individuals can be characterized by their identity-by-descent (IBD) state (Thompson, 2000). Furthermore, there are algorithms that can infer IBD states quickly and accurately (Young et al., 2022). For a sibling pair, the IBD state of a particular variant simply reflects whether the siblings inherited it from the same parental haplotypes or different parental haplotypes. This could serve as an instrumental variable for the difference in the siblings' alleles, independent of biases introduced by assortative mating or dynastic effects. Hypothetically, this principle could be generalized to more distant kin. The potential impact of this work is twofold. Firstly, we could improve the power of within-family designs by allowing valid inference for a broader range of pedigrees, potentially overcoming the power limitations of relying on a single meiosis (see Section 5.6). Secondly, we could assess the extent of, and alleviate, potential biases that exist in traditional methods for within-family designs, e.g., first differencing, linear fixed effects (Brumpton et al., 2020).

Within the parent-offspring design, there are opportunities to address applied questions for which existing methods are ill-suited. One such problem is the effect of skin tone discrimination on mental health or educational outcomes. This is an important question from a public health perspective, however, there are likely to be large ancestral biases, restricting us to a within-family design. Furthermore, the admixture that drives much of the genetic variation in skin tone is likely to introduce pleiotropy via long-acting linkage disequilibrium. My method can adjust for this, whereas other within-family methods cannot. My first step would be to apply for data from Born in Bradford, which is a parent-offspring birth cohort in the English city of Bradford, in which roughly 50% of newborns are of South Asian ancestry (Wright et al., 2013).

The sensitivity analysis in Chapter 3 allows users to specify population characteristics in the form of statistical tests, such that inverse probability weights that are inconsistent with these characteristics will be unlikely under the null distribution. I have identified three areas of future work. Firstly, the general statistical inference procedure for sample-constrained stochastic optimization problems is overly conservative because it does not consider the covariance between the objective function and constraint set. Bootstrapping the plug-in estimator ν_n^p in eq. (3.4) could yield a more efficient, if computationally expensive, inference, but the statistical properties of such an approach remain to be explored. Secondly, genetic data has unique properties that could be characterized within my framework and there is growing recognition that genetic association studies are not immune to sample selection bias (Mitchell et al., 2022). For example, autosomal variants should be independent of biological sex in the population (Pirastu et al.,

2021) and genetic segments shared IBD between two related individuals should not be enriched for certain alleles compared to non-IBD segments (Benonisdottir and Kong, 2022). I could formulate statistical tests for these characteristics and use them as constraints in the sensitivity analysis. Thirdly, while the sensitivity analysis in Chapter 3 is flexible, it is unclear that it is the most efficient way of utilizing these population characteristics. Dorn and Guo (2022) derive an asymptotically sharp interval for a class of inverse probability weighted estimators that utilizes quantile balancing of covariates. It would be valuable to assess whether this quantile balancing approach could be applied to the broader class of estimators I consider.

I view three important areas of future work for the coarsening bias method in Chapter 4. Firstly, my method assumes that we have a single coarsened exposure for a single latent exposure. In practice, we often have multiple coarsened measures of the same latent trait. Jin et al. (2021) describe an approach for integrating multiple measures of a latent exposure in MR studies using second-order summary statistics, for example, CRP, IL-6 and MCP-1 as biomarkers for underlying inflammation. It would be valuable to allow multiple coarsened exposures to provide evidence for the causal effect of a latent exposure. Secondly, my method relies on a strong structural assumption that the coarsened and latent exposures are related via a linear index model. It may be possible to relax some of these assumptions, for example, by assuming that the coarsened exposure is simply rank-preserving with respect to the latent exposure. Alternatively, we could consider a sensitivity analysis which partially identifies the effect of the latent exposure under a weaker set of assumptions. Thirdly, my method does not provide a way of assessing whether the exclusion restriction is violated due to coarsening bias. This last extension has been undertaken since publication of Chapter 4 by Tian et al. (2022), who propose a Gelman–Rubin uniformity statistic to test for exclusion restriction violations.

Appendix A

Supplementary material for Chapter 3

A.1 Further details for the applied example

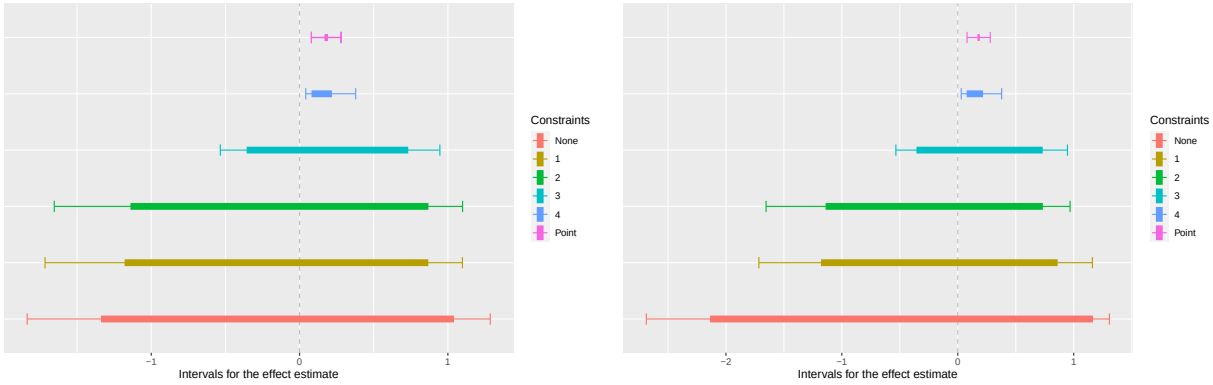
A.1.1 Varying the sensitivity parameters

It is important to report a few choices for the sensitivity parameters to understand which parameters are driving the width of the interval. Selecting $\Lambda_1 < 1.75$ results in an empty constraint set, indicating that there are no parameters which satisfy all of the auxiliary information constraints provided. The response rate constraint of Example 3.2 appears to be more informative when the interval $(\Lambda_0^l, \Lambda_0^u)$ is wider, which is expected. For all choices of parameters we consider, the constraints are informative and recover an interval that rejects the null.

A.1.2 Visualizing the feasible region

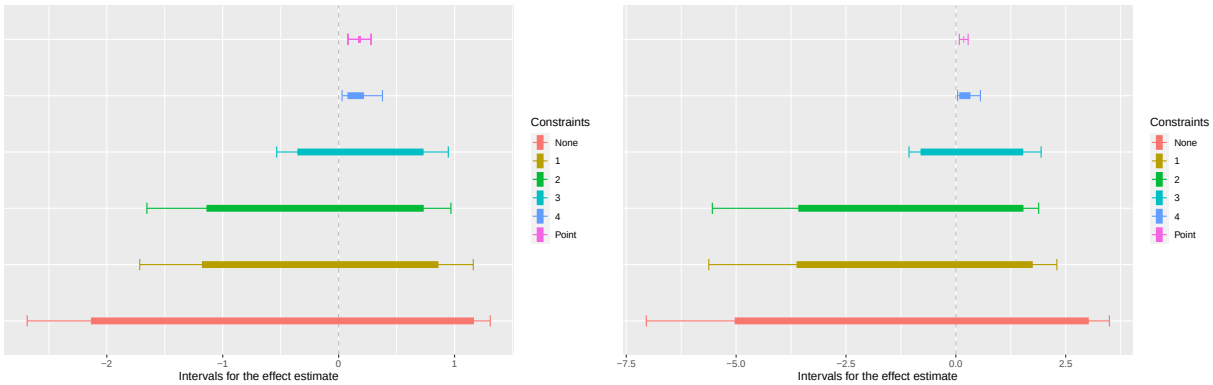
It is also illustrative to plot the feasible region for a couple of simple examples. Suppose that W_i consists only of the sex variable. We select sensitivity parameters $(\Lambda_0^l, \Lambda_0^u, \Lambda_1) = (0.02, 0.2, 2)$ as usual and consider two constraints: setting the response rate to be 0.055 and setting the population mean of male sex to be 0.495.

Figure A.2 plots the two feasible regions. The feasible regions are both small in comparison to the space implied by the sensitivity parameters. Imposing both constraints simultaneously will result in a non-empty feasible region. In fact, Nevo (2003) shows that the parameters of the selection model are exactly identified in this case provided each constraint provides a unique restriction on θ, in the sense that the outer product of the corresponding equality constraints is of full rank.



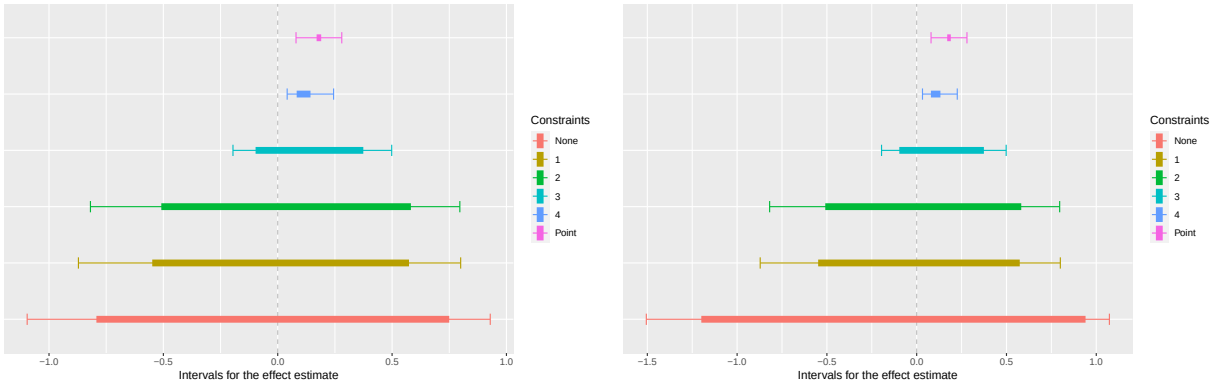
(a) $\Lambda_0^l = 0.02, \Lambda_0^u = 0.2, \Lambda_1 = 2$

(b) $\Lambda_0^l = 0.01, \Lambda_0^u = 0.5, \Lambda_1 = 2$



(c) $\Lambda_0^l = 0.02, \Lambda_0^u = 0.2, \Lambda_1 = 2.5$

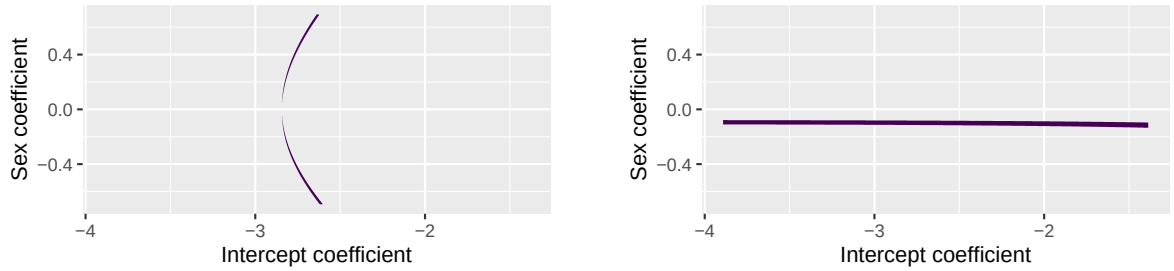
(d) $\Lambda_0^l = 0.01, \Lambda_0^u = 0.5, \Lambda_1 = 2.5$



(e) $\Lambda_0^l = 0.02, \Lambda_0^u = 0.2, \Lambda_1 = 1.75$

(f) $\Lambda_0^l = 0.01, \Lambda_0^u = 0.5, \Lambda_1 = 1.75$

Figure A.1: This figure presents several choices of sensitivity parameters for the applied example described in Section 3.6.



(a) Response rate constraint

(b) Covariate mean constraint

Figure A.2: This figure plots the feasible region (purple region) for a simple selection model with one variable and an intercept. Panel (a) sets the response rate to be 0.055 and panel (b) sets the population mean of male sex to be 0.495.

A.1.3 Implied selection probabilities within covariate strata

We can check the implied probabilities of sample selection within different covariate strata. Strata for which the implied probabilities are extreme, or inconsistent with known patterns of sample selection, could be addressed by introducing additional constraints. To illustrate this idea, Tables A.1 and A.2 show the implied probabilities within strata of sex and educational attainment of the sensitivity analysis in Section 3.6. Table A.1 shows the probabilities with only the response rate constraint and Table A.2 shows the probabilities with all constraints.

The probabilities in Table A.1 exhibit sample selection patterns that are inconsistent with known characteristics of UK Biobank (Fry et al., 2017). In particular, at both the lower and upper bounds, better educated individuals are less likely to select into the sample. At the lower bound, men are more likely to select into the sample than women.

The probabilities in Table A.2 are more consistent with UK Biobank selection patterns. At the lower bound, better educated individuals are now more likely to select into the sample. Women are more likely to select into the sample than men across most strata of educational attainment. Despite this, the selection pattern for education is still contrary to our expectations at the upper bound. This could indicate that a constraint on average educational attainment would tighten the upper bound. Alternatively, we could constrain the coefficient for educational attainment to be non-negative in the weight model.

A.2 Computation time of selection bias method

This simulation illustrates the computation time of our R package `selectioninterval` as the number of weight model increases. We replicate the simulation set-up in Section 3.5. To reiterate, our parameter is the regression coefficient of Y_i on X_i for $(X_i, Y_i) \sim \mathcal{N}(0, I_2)$, $i = 1, \ldots, 200$. The variables in the weight model are $W_i = (X_i, Y_i, Z_i^d)$, where $Z_i^d \sim \mathcal{N}(0, I_d)$. Figure A.3 shows the single-core computation time in seconds of the lower and upper bounds as the dimension d

Table A.1: Implied probabilities across sex and education strata, response rate constraint

Age finished school	Probabilities at lower bound		Probabilities at upper bound	
	Female	Male	Female	Male
14	0.59	0.72	0.33	0.17
15	0.43	0.60	0.27	0.12
16	0.40	0.56	0.35	0.17
17	0.28	0.43	0.31	0.13
18	0.22	0.33	0.28	0.11
19	0.21	0.30	0.22	0.08
20	0.14	0.30	0.19	0.07
21	0.09	0.17	0.17	0.06
22	0.10	0.16	0.11	0.04

Table A.2: Implied probabilities across sex and education strata, all constraints

Age finished school	Probabilities at lower bound		Probabilities at upper bound	
	Female	Male	Female	Male
14	0.07	0.06	0.20	0.17
15	0.10	0.09	0.20	0.17
16	0.13	0.12	0.21	0.18
17	0.16	0.17	0.21	0.17
18	0.19	0.20	0.19	0.16
19	0.20	0.22	0.15	0.13
20	0.24	0.22	0.14	0.10
21	0.29	0.28	0.13	0.10
22	0.27	0.31	0.09	0.08

is varied from 1 to 20. For this simulation, computational complexity appears to be sub-linear.

A.3 Sufficient conditions for Assumption 3.6

Assumption A.1. For sufficiently large n, $\Theta_n^r \subseteq B$ with probability one.

Assumption A.2. For all $j = 1, \ldots, J$, $h_{nj}(\theta)$ converges to $h_j(\theta)$ and $\epsilon_{nj}(\theta)$ converges to 0 with probability one as $n \to \infty$ uniformly on B.

Assumption A.3. For any $\theta \in \Theta$, let $\mathcal{A}(\theta) = \{j \colon h_j(\theta) = 0\}$ be the indices of active constraints, which could be empty. We assume that the gradient vectors $\nabla h_j(\vartheta)$, $j \in \mathcal{A}(\vartheta)$, are linearly independent.

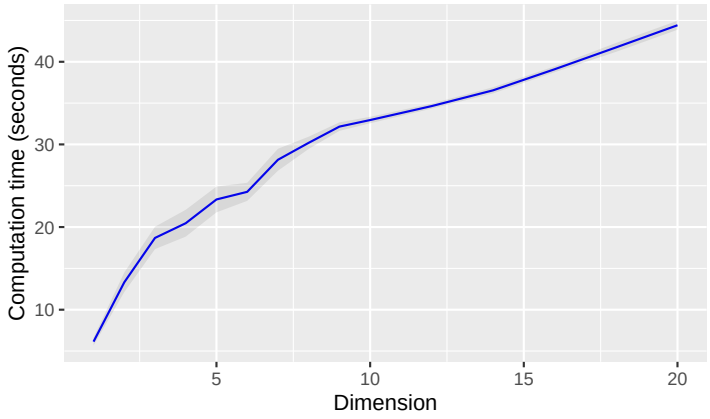


Figure A.3: This figure plots the computation time in seconds of our R package for a sample size of 200. The dimension d is varied between 1 and 20.

A.4 Technical details

Proof of Lemma 3.1

Proof. Consider the population problem. We will prove the 'if' statement since the 'only if' statement follows from some simple algebra. We will begin by noting that, since ν is the solution to a linear fractional programming problem and since Θ is a compact, convex polyhedron, the maximizing weight vector ϑ will lie at a vertex. In other words, $\vartheta \in \{1/b, 1/a\}^K$. Take an arbitrary weight vector $\theta \in \{1/b, 1/a\}^K$. Suppose there are $1 < m \leq K$ elements of θ which differ from ϑ. Without loss of generality, suppose these are the first m elements. Then we can write

$$
\begin{aligned}
\beta(w) &= \frac{\sum_{k=1}^{K} \theta_k f(t_k) p(t_k)}{\sum_{k=1}^{K} \theta_k g(t_k) p(t_k)} \\
&= \frac{\sum_{k=1}^{K} \vartheta_k f(t_k) p(t_k) - \sum_{k=1}^{m} q_k f(t_k) p(t_k)}{\sum_{k=1}^{K} \vartheta_j g(t_k) p(t_k) - \sum_{k=1}^{m} q_k g(t_k) p(t_k)} \\
&\geq \frac{\nu \sum_{k=1}^{K} \vartheta_k g(t_k) p(t_k) - \nu \sum_{k=1}^{m} q_k g(t_k) p(t_k)}{\sum_{k=1}^{K} \vartheta_k g(t_k) p(t_k) - \sum_{k=1}^{m} q_k g(t_k) p(t_k)} \\
&= \nu \frac{\sum_{k=1}^{K} \vartheta_k g(t_k) p(t_k) - \sum_{k=1}^{m} q_k g(t_k) p(t_k)}{\sum_{k=1}^{K} \vartheta_k g(t_k) p(t_k) - \sum_{k=1}^{m} q_k g(t_k) p(t_k)} \\
&= \nu
\end{aligned}
$$

The same holds with probability one for the sample problem by replacing $p(t_k)$ with $p_n(t_k)$. □

Proof of Proposition 3.4

Proof. Consider the population problem. Suppose that there are two global minima, $\nu_1 = \beta(\vartheta_1)$ and $\nu_2 = \beta(\vartheta_2)$, such that $\nu_1 = \nu_2 = \nu$ and $\vartheta_1 \neq \vartheta_2$. Since ν_1 and ν_2 are both global minima then, by Lemma 3.1, for all $k = 1, \ldots, K$,

$$\begin{aligned} q_k\, f(t_k) &\leq \nu_1\, q_k\, g(t_k) \\ q_k\, f(t_k) &\leq \nu_2\, q_k\, g(t_k) \end{aligned} \tag{A.1}$$

Without loss of generality, we assume that ϑ_1 and ϑ_2 differ by the first m elements. Then,

$$\begin{aligned} \nu = \nu_1 &= \frac{\sum_{k=1}^{K} \vartheta_{1j} f(t_k) p(t_k)}{\sum_{k=1}^{K} \vartheta_{1k} g(t_k) p(t_k)} \\ &= \frac{\sum_{k=1}^{K} \vartheta_{2k} f(t_k) p(t_k)}{\sum_{k=1}^{K} \vartheta_{2k} g(t_k) p(t_k)} \\ &= \frac{\sum_{k=1}^{K} \vartheta_{1k} f(t_k) p(t_k) - \sum_{k=1}^{K} q_k f(t_k) p(t_k)}{\sum_{k=1}^{K} \vartheta_{1k} g(t_k) p(t_k) - \sum_{k=1}^{m} q_k g(t_k) p(t_k)} \end{aligned}$$

where $q_k = \vartheta_{1k} - (1/a + 1/b - \vartheta_{1k})$. Rearranging, we obtain,

$$\sum_{k=1}^{m} q_k\, f(t_k)\, p(t_k) \;=\; \nu \sum_{k=1}^{m} q_k\, g(t_k)\, p(t_k)$$

However, (A.1) implies that this equality will only hold if, for all $k = 1, ..., m$, $f(t_k)/g(t_k) = \nu$, which cannot be true by Assumption 3.1. Therefore, by contradiction, the set $\{\theta \in \Theta \colon \beta(\theta) = \nu\}$ must be a singleton. The same holds with probability one for the sample problem by replacing $p(t_k)$ with $p_n(t_k)$. □

Proof of Proposition 3.3

Proof.

$$\begin{aligned} |\sigma_n^2(\vartheta_n) - \sigma^2(\vartheta)| &= |\sigma_n^2(\vartheta_n) - \sigma^2(\vartheta_n) + \sigma^2(\vartheta_n) - \sigma^2(\vartheta)| \\ &\leq |\sigma_n^2(\vartheta_n) - \sigma^2(\vartheta_n)| \;+\; |\sigma^2(\vartheta_n) - \sigma^2(\vartheta)| \\ &\leq \sup_{\theta \in \Theta} |\sigma_n^2(\theta) - \sigma^2(\theta)| \;+\; |\sigma^2(\vartheta_n) - \sigma^2(\vartheta)| \\ &\to 0 \text{ with probability one,} \end{aligned}$$

where the last inequality holds by the uniform strong consistency and the second term on the last line goes to zero with probability one since $\vartheta_n \to \vartheta$ with probability one by Proposition 3.1 and $\sigma^2(\theta) \in C(S)$. □

Before proving Theorem 3.1, we begin with some notation and preliminary lemmas. Denote the confidence bound for a particular θ and α as

$$C_n(\theta, \alpha) = Q_n(\theta) - Z_\alpha \, \sigma_n(\theta) \, n^{-1/2}$$

and the sample minimum over Θ_n^r at α_2 as

$$\zeta_n^r \in \arg\min\{C_n(\theta, \alpha_2) \colon \theta \in \Theta_n^r\}.$$

Recall that $\vartheta_n^r \in \arg\min\{Q_n(\theta) \colon \theta \in \Theta^r\}$ and $\vartheta = \arg\min\{Q(\theta) \colon \theta \in \Theta\}$, which is assumed to be unique. These quantities can be ordered deterministically as

$$C_n(\zeta_n^r, \alpha) \, \leq \, C_n(\vartheta_n^r, \alpha) \, \leq \, Q_n(\vartheta_n^r) \, \leq \, Q_n(\zeta_n^r).$$

The first lemma provides a lower bound for the coverage probability of $C_n(\vartheta_n^r, \alpha_2)$.

Lemma A.1. *Let $\delta_n = Z_{\alpha_2} \, \epsilon \, n^{-1/2}$ be a deterministic sequence where $\epsilon > 0$ is any positive constant, then*

$$\mathrm{pr}\Big\{ C_n(\vartheta_n^r, \alpha_2) \, \leq \, \nu \Big\} \geq \mathrm{pr}\Big\{ C_n(\zeta_n^r, \alpha_2) \, \leq \, \nu - \delta_n \Big\} + \mathrm{pr}\Big\{ \big|\sigma_n(\zeta_n^r) \, - \, \sigma_n(\vartheta_n^r)\big| \, \leq \, \epsilon \Big\} \, - \, 1.$$

Proof.

$$\mathrm{pr}\Big\{ C_n(\vartheta_n^r, \alpha_2) \, \leq \, \nu \Big\}$$

$$= \mathrm{pr}\Big\{ Q_n(\vartheta_n^r) \, - \, Z_{\alpha_2} \, \sigma_n(\vartheta_n^r) \, n^{-1/2} \, \leq \, \nu \Big\}$$

$$= \mathrm{pr}\Big[Q_n(\zeta_n^r) \, - \, Z_{\alpha_2} \sigma_n(\zeta_n^r) n^{-1/2} \, + \, \big\{ (Q_n(\vartheta_n^r) - Q_n(\zeta_n^r) \big\} \, + \, Z_{\alpha_2} \big\{ \sigma_n(\zeta_n^r) \, - \, \sigma_n(\vartheta_n^r) \big\} n^{-1/2} \, \leq \, \nu \Big]$$

$$\geq \mathrm{pr}\Big[Q_n(\zeta_n^r) \, - \, Z_{\alpha_2} \sigma_n(\zeta_n^r) n^{-1/2} \, + \, Z_{\alpha_2} \big\{ \sigma_n(\zeta_n^r) \, - \, \sigma_n(\vartheta_n^r) \big\} n^{-1/2} \, \leq \, \nu \Big]$$

$$\geq \mathrm{pr}\Big\{ Q_n(\zeta_n^r) \, - \, Z_{\alpha_2} \sigma_n(\zeta_n^r) n^{-1/2} \, + \, \delta_n \, \leq \, \nu, \, Z_{\alpha_2} \big| \sigma_n(\zeta_n^r) \, - \, \sigma_n(\vartheta_n^r) \big| n^{-1/2} \, \leq \, \delta_n \Big\}$$

$$\geq \mathrm{pr}\Big\{ C_n(\zeta_n^r, \alpha_2) \, \leq \, \nu - \delta_n \Big\} \, + \, \mathrm{pr}\Big\{ Z_{\alpha_2} \big| \sigma_n(\zeta_n^r) \, - \, \sigma_n(\vartheta_n^r) \big| n^{-1/2} \, \leq \, \delta_n \Big\} \, - \, 1$$

$$= \mathrm{pr}\Big\{ C_n(\zeta_n^r, \alpha_2) \, \leq \, \nu - \delta_n \Big\} \, + \, \mathrm{pr}\Big\{ \big|\sigma_n(\zeta_n^r) \, - \, \sigma_n(\vartheta_n^r)\big| \, \leq \, \epsilon \Big\} \, - \, 1.$$

□

Lemma A.2. *Suppose a sequence of functions $\tilde{Q}_n\colon \mathbb{R}^p \to \mathbb{R}$ is in $C(B)$ and converges to $Q(\theta)$ with probability one as $n \to \infty$ uniformly on B. Furthermore, let $\tilde{\nu}_n = \inf\{\tilde{Q}_n(\theta)\colon \Theta_n^r\}$ and $\tilde{\vartheta}_n \in \{\theta\colon \tilde{Q}_n(\theta) = \tilde{\nu}_n\}$. Then, under Assumptions 3.1, 3.2 and A.1 - A.3, $\tilde{\nu}_n \to \nu$ and $\tilde{\vartheta}_n \to \vartheta$ with probability one as $n \to \infty$.*

Proof. The proof of this lemma combines Theorem 5.3 and Theorem 5.5 in Shapiro, Dentcheva, and Ruszczyński (2009). Theorem 5.3 establishes consistency of optimal values and solutions when Θ is known. Theorem 5.5 generalizes this result to an estimated constraint set, in our case Θ_n^r.

In particular, Theorem 5.5 requires that the following two conditions are satisfied:

a) If $\theta_n \in \Theta_n^r$ and θ_n converges with probability one to a point θ^*, then $\theta^* \in \Theta$.

b) There exists a sequence $\theta_n \in \Theta_n^r$ such that θ_n converges to ϑ with probability one.

We begin with the proof of condition (a). Suppose $\theta^* \notin \Theta$. Then there exists some $j = 1, \ldots, J$ such that $h_j(\theta^*) \geq \delta$, where $\delta > 0$ is some constant. By the triangle inequality,

$$|h_{jn}(\theta_n) - h_j(\theta^*)| \leq |h_{jn}(\theta_n) - h_{jn}(\theta^*)| + |h_{jn}(\theta^*) - h_j(\theta^*)|.$$

The first term converges to zero with probability one because θ_n converges to θ^* and $h_{jn}(\theta) \in C(B)$, both with probability one. The second term also converges to zero with probability one by Assumption A.2 because $h_{jn}(\theta) \to h_j(\theta)$ uniformly on B with probability one. This means that for all $\epsilon > 0$ there exists an $n \geq n_\epsilon$ such that

$$|h_{jn}(\theta_n) - h_j(\theta^*)| < \epsilon.$$

However, since $h_j(\theta^*) \geq \delta$ and $\theta_n \in \Theta_n^r$, such that $h_{jn}(\theta_n) \leq \epsilon_{jn}(\theta_n)$, this is equivalent to

$$h_j(\theta^*) - h_{jn}(\theta_n) < \epsilon$$

whenever $\delta > \epsilon_{jn}(\theta_n)$. Without loss of generality, we can set $\epsilon = \delta - \epsilon_{jn}(\theta) > 0$. From here, we can rearrange,

$$h_{jn}(\theta_n) > h_j(\theta^*) - \epsilon \geq \delta - \epsilon \geq \epsilon_{jn}(\theta_n),$$

which means that $\theta_n \notin \Theta_n^r$, which is a contradiction. Therefore, it must be that $\theta^* \in \Theta$.

Condition (b) follows from the constraint qualification imposed in Assumption A.3 and the discussion in Shapiro, Dentcheva, and Ruszczyński (2009, p. 161-162). □

Proof of Theorem 3.1

Proof. We begin by invoking Lemma A.1, which states that

$$\text{pr}\Big\{C_n(\vartheta_n^r, \alpha_2) \leq \nu\Big\} \geq \text{pr}\Big\{C_n(\zeta_n^r, \alpha_2) \leq \nu - \delta_n\Big\} + \text{pr}\Big\{\big|\sigma_n(\zeta_n^r) - \sigma_n(\vartheta_n^r)\big| \leq \epsilon\Big\} - 1,$$

where $\delta_n = Z_{\alpha_2}\,\epsilon\,n^{-1/2}$ and $\epsilon > 0$ is any positive constant. We claim that

$$\lim_{n\to\infty} \text{pr}\Big\{\big|\sigma_n(\zeta_n^r) \,-\, \sigma_n(\vartheta_n^r)\big| \,\leq\, \epsilon\Big\} = 1.$$

This follows from Proposition 3.3 and Assumption 3.6. A sufficient condition for satisfying this assumption is that $\vartheta_n^r \to \vartheta$ and $\zeta_n^r \to \vartheta$ with probability one. This follows from Lemma A.2 since $Q_n(\theta)$ and $C_n(\theta, \alpha)$ both converge to $Q(\theta)$ with probability one uniformly on B by Assumption 3.3. Therefore, we have that

$$\begin{aligned}
&\lim_{n\to\infty} \text{pr}\Big\{C_n(\vartheta_n^r, \alpha_2) \,\leq\, \nu\Big\} \\
&\geq \lim_{n\to\infty} \text{pr}\Big\{C_n(\zeta_n^r, \alpha_2) \,\leq\, \nu - \delta_n\Big\} \\
&\geq \lim_{n\to\infty} \text{pr}\Big\{C_n(\zeta_n^r, \alpha_2) \,\leq\, \nu - \delta_n,\ \Theta \,\subseteq\, \Theta_n^r\Big\} \\
&\geq \lim_{n\to\infty} \text{pr}\Big\{C_n(\vartheta_n, \alpha_2) \,\leq\, \nu - \delta_n,\ \Theta \,\subseteq\, \Theta_n^r\Big\} \\
&= \lim_{n\to\infty} \left[\text{pr}\Big\{Q_n(\vartheta_n) - Z_{\alpha_2}\sigma_n(\vartheta_n)n^{-1/2} \,\leq\, \nu - \delta_n\Big\} - \text{pr}\Big\{C_n(\vartheta_n, \alpha_2) \,\leq\, \nu - \delta_n,\ \Theta \,\not\subseteq\, \Theta_n^r\Big\}\right] \\
&\geq \lim_{n\to\infty} \left(\text{pr}\left[n^{1/2}\,\frac{\big\{Q_n(\vartheta_n) \,-\, \nu\big\}}{\sigma(\vartheta)}\,\frac{\sigma(\vartheta)}{\sigma_n(\vartheta_n)} \,\leq\, Z_{\alpha_2}(1-\epsilon)\right] \,-\, \text{pr}\Big\{\Theta \,\not\subseteq\, \Theta_n^r\Big\}\right) \\
&\geq \Phi\{Z_{\alpha_2}(1-\epsilon)\} \,-\, \alpha_1
\end{aligned}$$

where the last inequality follows by Slutsky's theorem, Proposition 3.3 and Proposition 3.2. Since $\epsilon > 0$ is an arbitrarily small constant, this lower bound can be set arbitrarily close to $1 - \alpha_2 - \alpha_1$. □

Appendix B

Supplementary material for Chapter 4

B.1 Importance of the identifying assumptions

Assumptions 4.1 and 4.3 require that the latent exposure and measurement are related by a linear single index model. This assumption imposes considerable structure on the relationship between the two. To see why this assumption is necessary for identification, consider the more general model $L = G - V = \nu(Z, X) - V$, where $\nu(\cdot)$ is some continuous function. D is invariant to any monotone transformation $t(\cdot)$ in the sense that

$$D = I\{\nu(Z, X) \geq V\} = I\{t(\nu(Z, X)) \geq t(V)\} \tag{B.1}$$

One such monotone transformation we can take is $t(\cdot) = F_V(\cdot)$, where $F_V(\cdot)$ is the cumulative distribution of V, such that

$$D = I\{F_V(\nu(Z, X) \geq F_V(V)\} = I\{\text{pr}(D = 1 \mid Z, X) \geq U\} \tag{B.2}$$

where $U \sim Unif(0, 1)$. This means that the observable data distribution $f(Z, X, D, Y)$ is consistent with any monotone transformation of $\nu(Z, X)$, including $\text{pr}(D = 1 \mid Z, X)$ itself. By imposing the structural assumption that $G = \nu(Z, X) = \mu + \alpha Z + \gamma' X$, we reduce the class of models that the observed data distribution is consistent with to $\nu(Z, X)$ that are proportional to G. This allows us to separate G, which is linear in parameters, from the non-linear link function. In the absence of this linear index assumption, this separation does not occur. This approach to identification is within the class of 'identification by functional form' methods described in Lewbel (2019), which provides an overview of this class of methods and discusses their limitations. Section B.2 provides some simulation results when other assumptions fail, namely, correct specification of the link function and independence between the threshold and Z and X, both of which can introduce considerable bias. Bias from misspecification of the link

function can be ameliorated by using more flexible semi-parametric binary outcome estimators (Ichimura, 1993; Klein and Spady, 1993). Independence of the threshold from Z and X is a reasonable assumption when these are genetic factors and D is disease diagnosis or when D is a deterministic categorisation of the latent exposure (i.e., splitting BMI into obesity status).

B.2 Simulating violations of the identifying assumptions

In this section, we present some simulation results which violate the identifying assumptions stated in Section 4.2. Our data generating process is as follows:

$$Z \overset{iid}{\sim} \mathcal{N}(0,1),\ X \overset{iid}{\sim} \text{Exp}(1),\ V \overset{iid}{\sim} \text{SN}(0,1,a)$$
$$G = \alpha_Z Z + \alpha_X X$$
$$L = G - V,$$
$$D = I(L \geq bX)$$
$$Y = \beta_L L + \beta_X X + \beta_V V + \varepsilon,\ \varepsilon \overset{iid}{\sim} \mathcal{N}(0,1)$$

We set parameters $(\alpha_Z, \alpha_X, \beta_L,)$ equal to 1, (β_X, β_V) equal to 0.2, normalize Z, X and V to have mean 0 and normalize the variances as follows: $\text{var}(Z) = 2\theta^2/5$, $\text{var}(X) = 3\theta^2/5$ and $\text{var}(V) = 1 - \theta^2$, where $\theta^2 = 0.1$, meaning that $\sigma_L = 1$. $\text{SN}(0,1,a)$ denotes the skew normal distribution with skewness parameter a. We vary the skewness parameter over a range of values in Table B.1. When $a = 0$, this is equivalent to the standard normal distribution, meaning that the probit link will be correctly specified. As V becomes more skewed, the bias increases. This bias can be ameliorated with semi-parametric methods for binary outcomes (Ichimura, 1993; Klein and Spady, 1993).

Table B.1: Ratio of estimated to true β_L with link function misspecification

Link function	Skewness parameter a					
	0	1	2	3	4	5
Logistic	1.01	1.02	1.03	1.05	1.06	1.07
	(1.01-1.02)	(1.01-1.03)	(1.03-1.04)	(1.04-1.06)	(1.05-1.07)	(1.06-1.07)
Probit	1.01	1.02	1.03	1.05	1.06	1.07
	(1.00-1.02)	(1.01-1.03)	(1.02-1.04)	(1.04-1.06)	(1.05-1.07)	(1.06-1.08)
Semi-parametric†	1.00	1.00	1.00	1.00	1.01	1.01
	(0.99-1.00)	(0.99-1.01)	(0.99-1.01)	(1.00-1.01)	(1.00-1.01)	(1.00-1.02)

†Klein and Spady (1993) estimator. Estimates are means over 1,000 draws each of sample size 2,500. The parameter $b = 0$ throughout. 95% Monte Carlo confidence intervals are in parentheses.

Another assumption that can be violated is independence between the threshold of D and the observed variables Z and X. This dependence is captured by the parameter b. Since $\alpha_X = 1$, b can be interpreted as the relative contribution of X to the threshold compared to the latent

exposure L (e.g., $b = 0.5$ means that X contributes half as much to the threshold as to the latent exposure). In Table B.2, we vary the parameter b over a range of values and report the resulting bias. Despite the link function being correctly specified, there is significant bias from dependence in the threshold, which is roughly equal to b (e.g., when $b = 0.5$, relative bias in β_L is roughly 50%). Unlike misspecification of the link function, semi-parametric techniques cannot correct this bias. When X determines the threshold value, we cannot separately identify G in this framework. This simulation also suggests that threshold dependence may be bigger concern in this approach than misspecification of the link function.

Table B.2: Ratio of estimated to true β_L with threshold dependence

Link function	Skewness parameter b				
	0	0.1	0.25	0.5	1
Logistic	1.01	1.08	1.17	1.34	1.71
	(1.01-1.02)	(1.07-1.08)	(1.16-1.18)	(1.33-1.36)	(1.69-1.72)
Probit	1.01	1.07	1.17	1.33	1.69
	(1-1.02)	(1.06-1.08)	(1.16-1.18)	(1.32-1.34)	(1.68-1.70)
Semi-parametric†	1.00	1.05	1.15	1.30	1.67
	(0.99-1)	(1.05-1.06)	(1.14-1.16)	(1.29-1.31)	(1.66-1.69)

†Klein and Spady (1993) estimator. Estimates are means over 1,000 draws each of sample size 2,500. The parameter $a = 0$ throughout. 95% Monte Carlo confidence intervals are in parentheses.

B.3 Two-sample estimator and variance derivation

We begin by deriving equation (4.7). For some instrument Z_{ki} in $\mathcal{Z}_{J_0}$, the estimand β_G can be written as

$$\begin{aligned}
\sigma_G\beta &= \operatorname{cov}(Z_{ki}, Y)/\operatorname{cov}(Z_{ki}, G/\sigma_G) \\
&= \sigma_G\sigma_{Z_kY}/\operatorname{cov}(Z_{ki}, G) \\
&= \sigma_G\sigma_{Z_kY}/\operatorname{cov}(Z_{ki}, \textstyle\sum_{j=1}^{J}\alpha_j Z_{ji}) \\
&= \sigma_G\sigma_{Z_kY}/\alpha_k\sigma^2_{Z_k} \\
&= \sigma_{\tilde{G}}\Gamma_k/\tilde{\alpha}_k
\end{aligned} \tag{B.3}$$

which we can estimate from GWAS summary data. We can use inverse-variance weighting to 'meta-analyse' over these estimates for each Z_{ki} in $\mathcal{Z}_{J_0}$, which recovers the estimator (4.7). Denote $\hat{\beta}$ as the inverse-variance weighted estimator for $\sigma_V\beta$, then our two-sample estimator can be written as

$$\hat{\sigma}_{\tilde{G}}\hat{\beta} = \left(\sum_{j=1}^{J}\hat{\tilde{\alpha}}_j^2\sigma^2_{Z_j}\right)^{1/2}\hat{\beta}. \tag{B.4}$$

If we make the common assumption that $\hat{\tilde{\alpha}}_j$ has negligible uncertainty (i.e., $\hat{\tilde{\alpha}}_j \approx \tilde{\alpha}_j$), then we can write an estimator for the variance of (B.4) as

$$\left(\sum_{j=1}^{J} \hat{\tilde{\alpha}}_j^2 \sigma_{Z_j}^2\right) \sigma_{\hat{\beta}}^2. \tag{B.5}$$

Appendix C

Supplementary material for Chapter 5

C.1 Randomization distribution of offspring alleles

The distribution of offspring haplotypes is often approximated by a first order hidden Markov model (HMM) (Haldane, 1919; Thompson, 2000; Bates, Sesia, Sabatti, and Candès, 2020). This model assumes "no interference", such that the location of crossover events are independent and the likelihood of an offspring inheriting a SNP from a given maternal or paternal haplotype depends only on the inheritance at adjacent loci. This induces a Poisson renewal process for the distribution of distances between crossovers, however, it should be noted that there is evidence of positive crossover interference in human meioses which results in a more even spread of crossovers than would be expected with random placement. Recent literature has therefore suggested that a Gamma renewal process may be a more appropriate model, although we do not provide this extension here (Otto and Payseur, 2019).

The randomness in our randomization distribution arises from both the location of crossover events (i.e., the transition distribution) and the small probability of independent de novo mutations (i.e., the emission distribution). Without loss of generality, we describe the distribution of offspring alleles inherited from the mother $\boldsymbol{Z}^m$ given maternal haplotypes $\boldsymbol{M}^m$ and $\boldsymbol{M}^f$. Inheritance from the father is an independent instance of the same model. The transition distribution for the meiosis indicator at site j is assumed to be Poisson with mean equal to the genetic distance in centimorgans r_j between site $j-1$ and j:

$$\begin{aligned}\mathbb{P}(U_j^m = u_{j-1}^m \mid U_{j-1}^m = u_{j-1}^m) &= \mathbb{P}(\text{even number of recombinations between } j-1 \text{ and } j) \\ &= \tfrac{1}{2}(1 + e^{-2r_j}); \\ \mathbb{P}(U_j^m = U_{j'}^m) &= \tfrac{1}{2}(1 + e^{-2(d_j + \ldots + d_{j'})})\end{aligned}$$

where $u_{j-1}^m \in \{m, f\}$ and $j < j'$. Genetic distance is not proportional to physical distance

on the chromosome due to the presence of recombination hotspots where crossover events are more likely to occur (Belmont et al., 2005; Bherer, Campbell, and Auton, 2017). As r_j becomes large, the likelihood of an even number of recombinations approaches one half since genetically distant sites are transmitted almost independently.

The emission distribution is characterized by the probability of independent de novo single nucleotide mutations. A de novo mutation is said to occur when the base pair at some offspring SNP differs from the base pair they inherited from the parental haplotype. Within the context of the model, conditional on $U_j^m = u_j^m \in \{m, f\}$, each Z_j^m is sampled according to

$$\mathbb{P}(Z_j^m = M_j^{(u_j^m)} \mid U_j^m = u_j^m) = 1 - \epsilon \tag{C.1}$$

The probability of a de novo mutation ϵ is approximately $1 \cdot 10^{-8}$ in humans (Acuna-Hidalgo, Veltman, and Hoischen, 2016).

The graphical structure of the hidden Markov model is shown in Figure C.1. This graph differs from the more general structure shown in Figure 5.3b in that each meiosis indicator U_j^m depends only on the previous indicator U_{j-1}^m. Figure C.2 embeds the hidden Markov model within the complete causal model used throughout Section 4.3.

Our primary use of the Markovian structure described above is to derive propensity scores for offspring haplotypes $Z_j^m \in \{0, 1\}$. In particular, our goal is to express the propensity score of some SNP Z_j^m given the adjustment set $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}}^m)$ of Theorem 5.1, where $\boldsymbol{V}_{\mathcal{B}}^m = (\boldsymbol{M}_{\mathcal{B}}^{mf}, \boldsymbol{Z}_{\mathcal{B}}^m)$ and $\mathcal{B} \subseteq \mathcal{J} \setminus \{j\}$. Throughout this section we will assume that $\mathcal{B} = \{1, \dots, l\} \cup \{h, \dots p\}$ for $l < j < h$. Suppressing conditioning on $\boldsymbol{M}_j^{mf}$ and $\boldsymbol{M}_{\mathcal{B}}^{mf}$ for ease of notation, the propensity score for Z_j^m can be written as

$$\mathbb{P}(Z_j^m = 1 \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) = \sum_{u \in \{m,f\}} \mathbb{P}(Z_j^m = 1 \mid U_j^m = u) \mathbb{P}(U_j^m = u \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m). \tag{C.2}$$

It is therefore more convenient to consider the conditional probability of U_j^m. We state the following theorem:

Theorem C.1. *Using the conditional independence properties implied by Figure 5.3b, the conditional probability of $U_j^m = m$ can be factorized as*

$$\begin{aligned} &\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \\ \propto \; &\Bigg[\sum_{u \in \{m,f\}} \beta_{h-1}^m(u) \mathbb{P}(U_{h-1}^m = u \mid U_j^m = m) \Bigg] \Bigg[\sum_{u \in \{m,f\}} \mathbb{P}(U_j^m = m \mid U_l^m = u) \, \alpha_l^m(u) \Bigg]. \end{aligned}$$

The forward weights are defined recursively as

$$\begin{aligned} \alpha_1^m(u_1^m) &= \begin{cases} \frac{1}{2}(1 - \epsilon) & \text{if } M_1^{u_1^m} = z_1^m \\ \frac{1}{2}\epsilon & \text{if } M_1^{u_1^m} \neq z_1^m \end{cases} \\ \alpha_k^m(u_k^m) &= \sum_{u \in \{m,f\}} \mathbb{P}(Z_k^m = z_k^m \mid U_k^m = u_k^m) \mathbb{P}(U_k^m = u_k^m \mid U_{k-1}^m = u) \alpha_{k-1}^m(u), \; k = 2, \dots, p \end{aligned}$$

and the backward weights are defined recursively as

$$\begin{aligned}\beta_p^m(u_p^m) &= 1\\ \beta_k^m(u_k^m) &= \sum_{u\in\{m,f\}} \beta_{k+1}^m(u)\mathbb{P}(U_{k+1}^m = u \mid U_k^m = u_k^m)\mathbb{P}(Z_{k+1}^m = z_{k+1}^m \mid U_{k+1}^m = u),\ k = 1,\ldots,p-1,\end{aligned}$$

for $u_k^m \in \{m, f\}$ *and* $j, k \in \mathcal{J}$.

If we impose the simplifying assumption that $\epsilon = 0$, so that there is zero probability of de novo mutations, then the distribution of U_j^m derived in Theorem C.1 can be simplified further.

Corollary C.1. *Suppose the probability of a single nucleotide de novo mutation is* $\epsilon = 0$ *and suppose that the maternal haplotypes at* $b_1, b_2 \in \mathcal{J}$ *are heterozygous, where* $b_1 < l < j < h < b_2$. *That is,* $M_{b_1}^m \neq M_{b_1}^f$ *and* $M_{b_2}^m \neq M_{b_2}^f$. *Then the propensity score in Theorem C.1 can equivalently be written as*

$$\begin{aligned}&\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m)\\ \propto\ &\Big[\sum_{u\in\{m,f\}} \tilde{\beta}_{h-1}^m(u)\mathbb{P}(U_{h-1}^m = u \mid U_j^m = m)\Big]\Big[\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m \mid U_l^m = u)\,\tilde{\alpha}_l^m(u)\Big].\end{aligned}$$

where

$$\begin{aligned}\tilde{\alpha}_{b_1+1}^m(u_{b_1+1}^m) &= \mathbb{P}(Z_{b_1+1}^m = z_{b_1+1}^m \mid U_{b_1+1}^m = u_{b_1+1}^m)\mathbb{P}(U_{b_1+1}^m = u_{b_1+1}^m \mid U_{b_1}^m = u_{b_1}^m)\\ \tilde{\alpha}_k^m(u_k^m) &= \sum_{u\in\{m,f\}} \mathbb{P}(Z_k^m = z_k^m \mid U_k^m = u_k^m)\mathbb{P}(U_k^m = u_k^m \mid U_{k-1}^m = u)\tilde{\alpha}_{k-1}^m(u),\ k = b_1+2,\ldots,p;\end{aligned}$$

and

$$\begin{aligned}\tilde{\beta}_{b_2-1}^m(u_{b_2-1}^m) &= \mathbb{P}(U_{b_2}^m = u_{b_2}^m \mid U_{b_2-1}^m = u_{b_2-1}^m)\mathbb{P}(Z_{b_2}^m = z_{b_2}^m \mid U_{b_2}^m = u_{b_2}^m)\\ \beta_k^m(u_k^m) &= \sum_{u\in\{m,f\}} \tilde{\beta}_{k+1}^m(u)\mathbb{P}(U_{k+1}^m = u \mid U_k^m = u_k^m)\mathbb{P}(Z_{k+1}^m = z_{k+1}^m \mid U_{k+1}^m = u),\ k = 1,\ldots,b_2-2.\end{aligned}$$

We will occasionally have multiple instruments lying in the same window. We will then need to compute a multivariate propensity score. We state the following corollary without proof because it follows almost immediately from Theorem C.1.

Corollary C.2. *Suppose we have a collection of instruments* $\mathcal{J} = \{j_1, j_2, \ldots, j_r\}$ *such that* $l < j_1 < j_2 < \ldots < j_r < h$. *Then the propensity score can be written as*

$$\mathbb{P}(U_{j_1}^m = u_{j-1}^m, U_{j_2}^m = u_{j_2}^m, \ldots, U_{j_r}^m = u_{j_r}^m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \tag{C.3}$$

$$= \mathbb{P}(U_{j_1}^m = u_{j_1}^m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m)\prod_{k=2}^{r}\mathbb{P}(U_{j_k}^m = u_{j_k}^m \mid U_{j_{k-1}}^m = u_{j_{k-1}}^m, \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \tag{C.4}$$

The first propensity score $\mathbb{P}(U_{j_1}^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m)$ *takes the form in Theorem C.1 and*

$$\mathbb{P}(U_{j_k}^m = m \mid U_{j_{k-1}}^m = u_{j_{k-1}}^m, \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \tag{C.5}$$

$$\propto \mathbb{P}(U_{j_k}^m = m \mid U_{j_{k-1}}^m = u_{j_{k-1}}^m)\Big[\sum_{u\in\{m,f\}} \beta_{h-1}^m(u)\mathbb{P}(U_{h-1}^m = u \mid U_{j_k}^m = m)\Big] \tag{C.6}$$

where $\beta_{h-1}^m(u)$ *is the backward weight defined in Theorem C.1.*

Figure C.1: Graphical representation of Haldane's hidden Markov model

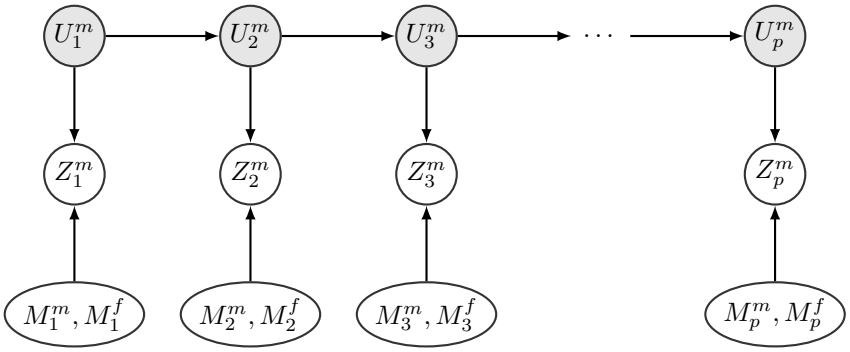
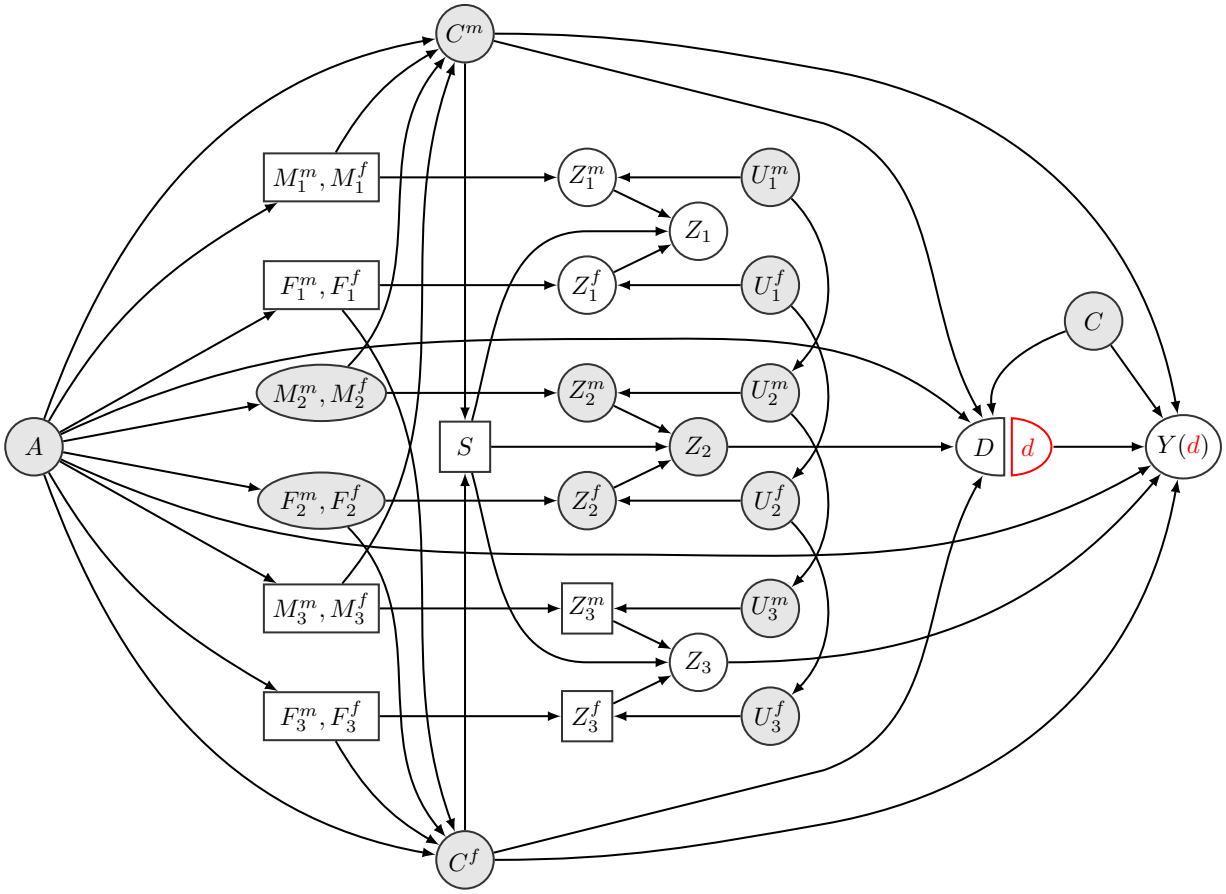


Figure C.2: Haldane's hidden Markov model embedded in our full causal model



C.2 Technical proofs

Proposition 5.1

Proof. From Assumption 5.1 we know that, conditional on $(M_j^m, M_j^f, F_j^m, F_j^f)$, Z_j^m and Z_j^f only depend on $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$, respectively, and exogenous mutation events. By (5.1), $Z_j = Z_j^m + Z_j^f$ given that $S = 1$ (fertilization occurs). Finally, by Assumption 5.2, the meiosis indicators $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$ are independent of all confounders (A, C^m, C^f, C). Therefore, the conditional independence statement immediately follows. □

Theorem C.1

Proof. The conditional probability of U_j^m can be factorized as

$$
\begin{aligned}
&\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \\
&\propto \mathbb{P}(U_j^m = m, \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \\
&= \mathbb{P}(\boldsymbol{Z}_{h:p}^m = \boldsymbol{z}_{h:p}^m \mid U_j^m = m)\mathbb{P}(U_j^m = m, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m) \\
&= \Bigg[\sum_{u\in\{m,f\}} \mathbb{P}(\boldsymbol{Z}_{(j+1):p}^m = \boldsymbol{z}_{(j+1):p}^m, U_{h-1}^m = u \mid U_j^m = m)\Bigg]\Bigg[\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m, U_l^m = u, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m)\Bigg] \\
&= \Bigg[\sum_{u\in\{m,f\}} \mathbb{P}(\boldsymbol{Z}_{(j+1):p}^m = \boldsymbol{z}_{(j+1):p}^m \mid U_{h-1}^m = u)\mathbb{P}(U_{h-1}^m = u \mid U_j^m = m)\Bigg] \\
&\quad \Bigg[\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m \mid U_l^m = u)\mathbb{P}(U_l^m = u, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m)\Bigg] \\
&= \Bigg[\sum_{u\in\{m,f\}} \beta_{h-1}^m(u)\mathbb{P}(U_{h-1}^m = u \mid U_j^m = m)\Bigg]\Bigg[\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m \mid U_l^m = u)\alpha_l^m(u)\Bigg].
\end{aligned}
$$

The forward weight $\alpha_1^m(u_1^m)$ for some $u_1^m \in \{m, f\}$ can be derived as

$$
\begin{aligned}
\alpha_1^m(u_1^m) &= \mathbb{P}(U_1^m = u_1^m, Z_1^m = z_1^m) \\
&= \mathbb{P}(Z_1^m = z_1^m \mid U_1^m = u_1^m)\mathbb{P}(U_1^m = u_1^m) \\
&= \tfrac{1}{2}\mathbb{P}(Z_1^m = z_1^m \mid U_1^m = u_1^m)
\end{aligned}
$$

where the emission probability is known. A recursive expression for the forward weight $\alpha_j^m(u_j^m)$ for $j = 2, \ldots, p$ can be derived as

$$
\begin{aligned}
\alpha_j^m(u_j^m) &= \mathbb{P}(U_j^m = u_j^m, \boldsymbol{Z}_{1:j}^m = \boldsymbol{z}_{1:j}^m) \\
&= \sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = u_j^m, U_{j-1}^m = u_{j-1}^m, \boldsymbol{Z}_{1:j}^m = \boldsymbol{z}_{1:j}^m) \\
&= \sum_{u\in\{m,f\}} \mathbb{P}(Z_j^m = z_j^m \mid U_j^m = u_j^m)\mathbb{P}(U_j^m = u_j^m \mid U_{j-1}^m = u)\mathbb{P}(U_{j-1}^m = u, \boldsymbol{Z}_{1:(j-1)}^m = \boldsymbol{z}_{1:(j-1)}^m) \\
&= \sum_{u\in\{m,f\}} \mathbb{P}(Z_j^m = z_j^m \mid U_j^m = u_j^m)\mathbb{P}(U_j^m = u_j^m \mid U_{j-1}^m = u)\,\alpha_{j-1}^m(u).
\end{aligned}
$$

The backward weight $\beta_j^m(u_j^m)$ for some $u_j^m \in \{m, f\}$ and $j = 1, \ldots, p-1$ can be derived as

$$\begin{aligned}
&\beta_j^m(u_j^m) = \mathbb{P}(\boldsymbol{Z}_{(j+1):p}^m = \boldsymbol{z}_{(j+1):p}^m \mid U_j^m = u_j^m) \\
&= \sum_{u \in \{m,f\}} \mathbb{P}(\boldsymbol{Z}_{(j+1):p}^m = \boldsymbol{z}_{(j+1):p}^m, U_{j+1}^m = u \mid U_j^m = u_j^m) \\
&= \sum_{u \in \{m,f\}} \mathbb{P}(\boldsymbol{Z}_{(j+2):p}^m = \boldsymbol{z}_{(j+2):p}^m \mid U_{j+1}^m = u)\mathbb{P}(Z_{j+1}^m = z_{j+1}^m \mid U_{j+1}^m = u)\mathbb{P}(U_{j+1}^m = u \mid U_j^m = u_j^m).
\end{aligned}$$

Writing the probability of U_p^m shows that $\beta_p^m(u) = 1$ for all $u \in \{m, f\}$. □

Corollary C.1

Proof. The proof involves some manipulation of conditional independencies. We simplify the probability with respect to b_1 and omit simplification with respect to b_2 for brevity. As with the proof of Theorem C.1 we begin by factorising the conditional probability of U_j^m.

$$\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) = \frac{\mathbb{P}(\boldsymbol{Z}_{h:p}^m = \boldsymbol{z}_{h:p}^m \mid U_j^m = m)\mathbb{P}(U_j^m = m, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m)}{\mathbb{P}(\boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m)}. \tag{C.7}$$

Since $b_1 < j$ we are concerned with simplifying the second probability in the numerator of equation (C.7).

$$\begin{aligned}
&\mathbb{P}(U_j^m = m, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m) \\
&= \sum_{u \in \{m,f\}} \mathbb{P}(U_j^m = m, U_{b_1}^m = u, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m) \\
&= \sum_{u \in \{m,f\}} \mathbb{P}(U_j^m = m, \boldsymbol{Z}_{(b_1+1):l}^m = \boldsymbol{z}_{(b_1+1):l}^m \mid U_{b_1}^m = u)\mathbb{P}(U_{b_1}^m = u, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \\
&= \mathbb{P}(U_j^m = m, \boldsymbol{Z}_{(b_1+1):l}^m = \boldsymbol{z}_{(b_1+1):l}^m \mid U_{b_1}^m = m)\mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \\
&= \mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \sum_{u \in \{m,f\}} \mathbb{P}(U_j^m = m \mid U_{j-1}^m = u)\mathbb{P}(U_{j-1}^m = u, \boldsymbol{Z}_{(b_1+1):(j-1)}^m = \\
&\qquad \boldsymbol{z}_{(b_1+1):(j-1)}^m \mid U_{b_1}^m = m) \\
&= \mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \sum_{u \in \{m,f\}} \mathbb{P}(U_j^m = m \mid U_{j-1}^m = u)\tilde{\alpha}_{j-1}^m(u).
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\alpha}_{b_1+1}^m(u_{b_1+1}^m) &= \mathbb{P}(Z_{b_1+1}^m = z_{b_1+1}^m \mid U_{b_1+1}^m = u_{b_1+1}^m)\mathbb{P}(U_{b_1+1}^m = u_{b_1+1}^m \mid U_{b_1}^m = m) \\
\tilde{\alpha}_k^m(u_k^m) &= \sum_{u \in \{m,f\}} \mathbb{P}(Z_k^m = z_k^m \mid U_k^m = u_k^m)\mathbb{P}(U_k^m = u_k^m \mid U_{k-1}^m = u)\tilde{\alpha}_{k-1}^m(u), \\
&\quad \text{for } k = b_1 + 2, \ldots, j-1.
\end{aligned}$$

We now factorize the denominator of equation (C.7).

$$\mathbb{P}(\boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) = \mathbb{P}(\boldsymbol{Z}_{(b_1+1):l} = \boldsymbol{z}_{(b_1+1):l}, \boldsymbol{Z}_{h:p} = \boldsymbol{z}_{h:p} \mid U_{b_1}^m = m)\mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1} = \boldsymbol{z}_{1:b_1}).$$

Substituting these simplified expressions back in equation (C.7) we obtain

(C.8)

$$\begin{aligned}
&\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \\
= &\frac{\mathbb{P}(\boldsymbol{Z}_{h:p}^m = \boldsymbol{z}_{h:p}^m \mid U_j^m = m)\mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m)\sum_{u\in\{m,f\}}\mathbb{P}(U_j^m = m \mid U_{j-1}^m = u)\tilde{\alpha}_{j-1}^m(u)}{\mathbb{P}(\boldsymbol{Z}_{(b_1+1):l} = \boldsymbol{z}_{(b_1+1):l}, \boldsymbol{Z}_{h:p} = \boldsymbol{z}_{h:p} \mid U_k^m = m)\mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1} = \boldsymbol{z}_{1:b_1})} \\
= &\frac{\mathbb{P}(\boldsymbol{Z}_{h:p}^m = \boldsymbol{z}_{h:p}^m \mid U_j^m = m)\sum_{u\in\{m,f\}}\mathbb{P}(U_j^m = m \mid U_{j-1}^m = u)\tilde{\alpha}_{j-1}^m(u)}{\mathbb{P}(\boldsymbol{Z}_{(b_1+1):l} = \boldsymbol{z}_{(b_1+1):l}, \boldsymbol{Z}_{h:p} = \boldsymbol{z}_{h:p} \mid U_{b_1}^m = m)}.
\end{aligned}$$

which does not depend on $\boldsymbol{Z}_{1:k}^m$. □

C.3 Simulation description

Table C.1: Description of the simulation variables and parameters

Variable	Description of how the variable is constructed	Parameters
$\boldsymbol{M}_i^m, \boldsymbol{M}_i^f, \boldsymbol{F}_i^m, \boldsymbol{F}_i^f$	The parental haplotypes are constructed to allow linkage disequilibrium in nearby SNPs. For each parental haplotype we first sample from a p-variate normal such that $X_{ij} \sim \mathcal{N}(0,1)$ and $Cov(X_{ij}, X_{ik}) = \rho^{\lvert j-k\rvert}$, $0 < \rho < 1$, $j,k \in \mathcal{J}$. Thresholds $V_{ij} \sim Unif(a,b)$ are sampled and the haplotypes are defined as $M_{ij}^m = I\{X_{ij} > V_{ij}\}$ where $I\{\cdot\}$ is the indicator function (and similarly for the other haplotypes).	$\rho = 0.75$ $a = \Phi^{-1}(0.6)$ $b = \Phi^{-1}(0.95)$ where $\Phi^{-1}(\cdot)$ is the inverse normal CDF.

C_i^m, C_i^f	We first define a variable $$\hat{\mu}_i^m = \frac{1}{p}\sum_{j=1}^{p}(M_{ij}^m + M_{ij}^f).$$ It follows from our construction of the parental haplotypes that $$\mu^m = E[\hat{\mu}^m] = 2\left(1 - \frac{1}{b-a}\int_a^b \Phi(x)dx\right).$$ where $\Phi(\cdot)$ is the normal CDF. For each individual i we sample the parental confounder such that $$C_i^m \sim \mathcal{N}(\hat{\mu}_i^m - \mu^m, 1).$$ We follow an identical procedure for C_i^f.	N/A
C_i	We construct the offspring confounder as $$C_i \sim \mathcal{N}(0, 1).$$	N/A
$\boldsymbol{Z}_i^m, \boldsymbol{Z}_i^f$	We sample the offspring haplotypes using Algorithm 1 in Bates, Sesia, Sabatti, and Candès (2020). This algorithm unconditionally samples a full haplotype $\boldsymbol{Z}_i^m$ or $\boldsymbol{Z}_i^f$ according to the hidden Markov model described in Appendix C.1. It depends on the genetic distances $\boldsymbol{r}$ and de novo mutation rate ϵ. We sample $r_j \sim Unif(c, d)$ and set $r_k = \infty$ for $k = 37, 62, 86, 112$ so that the instruments are unconditionally independent. From these haplotypes we choose a subset $\mathcal{J}_g \subset \mathcal{J}$ to be instruments.	$\epsilon = 10^{-8}$ $c = 0$ $d = 0.75$ $\mathcal{J}_g = \{25, 50, 75, 100, 125\}$

D_i	The exposure follows a linear structural equation model $$D_i = \gamma^\intercal \boldsymbol{Z}_i + \theta^m C_i^m + \theta^f C_i^f + \theta^c C_i + \nu_i$$ where $\nu_i \sim \mathcal{N}(0, 0.7)$. We choose γ so that it is zero everywhere except for γ_{24}, γ_{49}, γ_{74}, γ_{99} and γ_{124} which represent causal variants. The parameters are chosen so that $Var(D_i) = 1$.	$\theta^m = \theta^f = \sqrt{0.3}$ $\theta^c = \sqrt{0.75}$ $\gamma_j = \sqrt{0.1}$ for $j = 24, 49, 74, 99, 124$.
Y_i	The outcome follows a linear structural equation model $$Y_i = \beta D_i + \delta^\intercal \boldsymbol{Z}_i + \phi^m C_i^m + \phi^f C_i^f + \phi^c C_i + \upsilon_i$$ where $\upsilon_i \sim \mathcal{N}(0, 0.7)$. We choose δ so that it is zero everywhere except for δ_{23}, δ_{27}, δ_{48}, δ_{52}, δ_{73}, δ_{77}, δ_{98}, δ_{102}, δ_{123} and δ_{127} which represent pleiotropic variants. The parameters are chosen so that $Var(Y_i) = 1$.	$\beta = 0$ $\phi^m = \phi^f = \sqrt{0.3}$ $\phi^c = \sqrt{0.75}$ $\delta_j = \sqrt{0.05}$ for $j = 23, 27, 48, 52, 73, 77, 98, 102, 123, 127$.

Theorem 5.1 implies that a sufficient adjustment set for this simulation is

$$(\boldsymbol{M}^{mf}_{\mathcal{B}_g}, \boldsymbol{F}^{mf}_{\mathcal{B}_g}, \boldsymbol{Z}_{\mathcal{B}}) \tag{C.9}$$

where

$$\mathcal{B} = \mathcal{J} \setminus \{24, 25, 26, 49, 50, 51, 99, 74, 75, 76, 99, 100, 101, 124, 125, 126\}$$

and

$$\mathcal{B}_g = \mathcal{B} \cup \{25, 50, 75, 100, 125\}.$$

Bibliography

Acuna-Hidalgo, Rocio, Joris A. Veltman, and Alexander Hoischen (2016).
"New insights into the generation and role of de novo mutations in health and disease".
In: *Genome Biology* 17.1, pp. 1–19.

Andrews, Donald W K and Xiaoxia Shi (2013).
"Inference based on conditional moment inequalities".
In: *Econometrica* 81.2, pp. 609–666.

Andrews, Donald W K and Gustavo Soares (2010).
"Inference for parameters defined by moment inequalities using generalized moment selection".
In: *Econometrica* 78.1, pp. 119–157.

Angrist, Joshua D. and Guido W. Imbens (1995).
"Two-stage least squares estimation of average causal effects in models with variable treatment intensity".
In: *Journal of the American Statistical Association* 90.430, p. 431.

Aronow, Peter M. and Donald K.K. Lee (2013).
"Interval estimation of population means under unknown but bounded probabilities of sample selection".
In: *Biometrika* 100.1, pp. 235–240.

Balke, Alexander and Judea Pearl (1997).
"Bounds on treatment effects from studies with imperfect compliance".
In: *Journal of the American Statistical Association* 92.439, pp. 1171–1176.

Bareinboim, Elias, Jin Tian, and Judea Pearl (2014).
"Recovering from selection bias in causal and statistical inference".
In: *Proceedings of The Twenty-Eighth Conference on Artificial Intelligence*,
Pp. 339–341.

Bates, Stephen, Matteo Sesia, Chiara Sabatti, and Emmanuel Candes (2020).
"Causal inference in genetic trio studies".
In: *arXiv*.
arXiv: 2002.09644.

Bates, Stephen, Matteo Sesia, Chiara Sabatti, and Emmanuel Candès (2020).
"Causal inference in genetic trio studies".

In: *Proceedings of the National Academy of Sciences of the United States of America* 117.39, pp. 24117–24126.
arXiv: 2002.09644.

Belmont, John W. et al. (2005).
"A haplotype map of the human genome".
In: *Nature* 437.7063, pp. 1299–1320.

Benjamini, Yoav and Ruth Heller (2008).
"Screening for partial conjunction hypotheses".
In: *Biometrics* 64.4, pp. 1215–1222.

Benonisdottir, Stefania and Augustine Kong (2022).
"The genetics of participation: method and analysis".
In: *bioRxiv*, p. 2022.02.11.480067.

Berger, Roger L. and Dennis D. Boos (1994).
"P values maximized over a confidence set for the nuisance parameter".
In: *Journal of the American Statistical Association* 89.427, pp. 1012–1016.

Bherer, Claude, Christopher L. Campbell, and Adam Auton (2017).
"Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales".
In: *Nature Communications* 8.

Bowden, Jack, George Davey Smith, and Stephen Burgess (2015).
"Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression".
In: *International Journal of Epidemiology* 44.2, pp. 512–525.

Bowden, Jack, George Davey Smith, et al. (2016).
"Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator".
In: *Genetic Epidemiology* 40.4, pp. 304–314.

Bowden, Jack, Fabiola Del Greco M, et al. (2016).
"Assessing the suitability of summary data for two-sample mendelian randomization analyses using MR-Egger regression: The role of the I^2 statistic".
In: *International Journal of Epidemiology* 45.6, pp. 1961–1974.

Bowden, Jack, Fabiola Del Greco M, et al. (2017).
"A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization".
In: *Statistics in Medicine* 36.11, pp. 1783–1802.

Boyd, Andy et al. (2013).
"Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children".

In: *International Journal of Epidemiology* 42, pp. 111–127.

Boyle, Evan A, Yang I Li, and Jonathan K Pritchard (2017).
"An expanded view of complex traits: From polygenic to omnigenic".
In: *Cell* 169.7, pp. 1177–1186.

Bretz, Frank, Torsten Hothorn, and Peter Westfall (2016).
Multiple Comparisons Using R.
Chapman and Hall/CRC.

Broman, Karl W. and James L. Weber (2000).
"Characterization of human crossover interference".
In: *American Journal of Human Genetics* 66.6, pp. 1911–1926.

Brookhart, M. Alan and Sebastian Schneeweiss (2007).
"Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results".
In: *International Journal of Biostatistics* 3.1, pp. 1–19.

Brumpton, Ben et al. (2020).
"Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses".
In: *Nature Communications* 11.1, pp. 1–13.

Burgess, Stephen, Adam Butterworth, and Simon G. Thompson (2013).
"Mendelian randomization analysis with multiple genetic variants using summarized data".
In: *Genetic Epidemiology* 37.7, pp. 658–665.

Burgess, Stephen and Jeremy A. Labrecque (2018).
"Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates".
In: *European Journal of Epidemiology* 33.10, pp. 947–952.
arXiv: 1804.05545.

Burgess, Stephen, Robert A. Scott, et al. (2015).
"Using published data in Mendelian randomization: A blueprint for efficient identification of causal risk factors".
In: *European Journal of Epidemiology* 30.7, pp. 543–552.

Caliebe, Amke et al. (2022).
"Including diverse and admixed populations in genetic epidemiology research".
In: *Genetic Epidemiology*, pp. 1–25.

Cardon, Lon R. and Lyle J. Palmer (2003).
"Population stratification and spurious allelic association".
In: *Lancet* 361.9357, pp. 598–604.

Cassim, Shemana et al. (2019).

"Patient and carer perceived barriers to early presentation and diagnosis of lung cancer: A systematic review".
In: *BMC Cancer* 19.1, pp. 1–14.

Chattopadhyay, Ambarish, Christopher H. Hase, and José R. Zubizarreta (2020).
"Balancing vs modeling approaches to weighting in practice".
In: *Statistics in Medicine* 39.24, pp. 3227–3254.

Chen, Lina et al. (2008).
"Alcohol Intake and Blood Pressure: a Systematic Review Implementing a Mendelian Randomization Approach".
In: *PLoS Medicine* 5.3, e52.

Chernozhukov, Victor, Han Hong, and Ellie Tamer (2007).
"Estimation and confidence regions for parameter sets in econometric models".
In: *Econometrica* 75.5, pp. 1243–1284.

Chernozhukov, Victor, Sokbae Lee, and Adam M Rosen (2013).
"Intersection bounds: Estimation and inference".
In: *Econometrica* 81.2, pp. 667–737.
arXiv: 0907.3503.

Cinelli, Carlos and Chad Hazlett (2020).
"Making sense of sensitivity: extending omitted variable bias".
In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82.1, pp. 39–67.

Cornfield, Jerome et al. (1959).
"Smoking and lung cancer: recent evidence and a discussion of some questions".
In: *Journal of the National Cancer Institute* 22, pp. 173–203.

Crump, Richard K. et al. (2009).
"Dealing with limited overlap in estimation of average treatment effects".
In: *Biometrika* 96.1, pp. 187–199.

Curnow, R.N. (1972).
"The multifactorial model for the inheritance of liability to disease and its implications for relatives at risk".
In: *Biometrics* 28.4, pp. 931–946.

Davey Smith, George (2001).
"Reflections on the limitations to epidemiology".
In: *Journal of Clinical Epidemiology* 54.4, pp. 325–331.

— (2006).
"Capitalising on Mendelian randomization to assess the effects of treatments".
In: *JLL Bulletin*.

Davey Smith, George and Shah Ebrahim (2003).

"'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease?"
In: *International Journal of Epidemiology* 32.1, pp. 1–22.

Davey Smith, George, Smith Michael, et al. (2020).
"Mendel 's laws , Mendelian randomization and causal inference in observational data: substantive and nomenclatural issues".
In: *European Journal of Epidemiology* 35.2, pp. 99–111.

Davies, Neil M., Matt Dickson, et al. (2018).
"The causal effects of education on health outcomes in the UK Biobank".
In: *Nature Human Behaviour* 2.2, pp. 117–125.

Davies, Neil M., David Gunnell, et al. (2013).
"Physicians' prescribing preferences were a potential instrument for patients' actual prescriptions of antidepressants".
In: *Journal of Clinical Epidemiology* 66.12, pp. 1386–1396.

Davies, Neil M., Laurence J. Howe, et al. (2019).
"Within family Mendelian randomization studies".
In: *Human Molecular Genetics* 28.R2, R170–R179.

Deming, W. Edwards and Frederick F. Stephan (1940).
"On a least squares adjustment of a sampled frequency table when the expected marginal totals are known".
In: *The Annals of Mathematical Statistics* 11.4, pp. 427–444.

Didelez, Vanessa and Nuala Sheehan (2007).
"Mendelian randomization as an instrumental variable approach to causal inference".
In: *Statistical Methods in Medical Research* 16.4, pp. 309–330.

Diemer, Elizabeth W. et al. (2020).
"Application of the instrumental inequalities to a Mendelian randomization study with multiple proposed instruments".
In: *Epidemiology* 31.1, pp. 65–74.

Ding, Peng, Xinran Li, and Luke W Miratrix (2017).
"Bridging finite and super population causal inference".
In: *Journal of Causal Inference* 5.2, pp. 1–8.

Dorn, Jacob and Kevin Guo (2022).
"Sharp sensitivity analysis for inverse propensity weighting via quantile balancing".
In: *Journal of the American Statistical Association*, pp. 1–36.
arXiv: 2102.04543.

Duarte, Guilherme et al. (2021).
"An automated approach to causal inference in discrete settings".
In: *arXiv*, pp. 1–53.

arXiv: 2109.13471.

Egger, Matthias, George Davey Smith, and Chris Minder (1997).
"Bias in meta-analysis detected by a simple, graphical test".
In: *The BMJ* 315.629.

Falconer, D. S. (1965).
"The inheritance of liability to certain diseases, estimated from the incidence among relatives".
In: *Annals of Human Genetics* 29.1, pp. 51–76.

Feinstein, Alvan R (1988).
"Scientific standards in epidemiologic studies of the menace of daily life".
In: *Science* 242, pp. 1257–1263.

Fisher, R. A. (1918).
"The Correlation Between Relatives on the Supposition of Mendelian Inheritance."
In: *Transactions of the Royal Society of Edinburgh* 52.2, pp. 399–433.

Fisher, Ronald A. (1925).
Statistical methods for research workers.
Edinburgh: Oliver & Boyd.

— (1935).
The design of experiments.
Edinburgh: Oliver & Boyd,
P. 257.

— (1951).
"Statistical methods in genetics".
In: *Heredity* 6, pp. 1–12.

Fisher, Ronald Aylmer (1926).
"The Arrangement of Field Experiments".
In: *Journal of the Ministry of Agriculture* 33, pp. 503–513.

Fraser, Abigail et al. (2013).
"Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort".
In: *International Journal of Epidemiology* 42, pp. 97–110.

Fry, Anna et al. (2017).
"Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population".
In: *American Journal of Epidemiology* 186.9, pp. 1026–1034.

Gastwirth, Joseph L., Abba M. Krieger, and Paul R. Rosenbaum (1998).
"Dual and simultaneous sensitivity analysis for matched pairs".
In: *Biometrika* 85.4, pp. 907–920.

Giacomini, Raffaella and Toru Kitagawa (2021).

"Robust Bayesian inference for set-identified models".
In: *Econometrica* 89.4, pp. 1519–1556.

Gopalan, Shyamalika et al. (2022).
"Human genetic admixture through the lens of population genomics".
In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1852.
arXiv: 2109.12190.

Gray, Richard and Keith Wheatley (1991).
"How to avoid bias when comparing bone marrow transplantation with chemotherapy".
In: *Bone Marrow Transplantation* 7.3, pp. 9–12.

Griffith, Gareth J. et al. (2020).
"Collider bias undermines our understanding of COVID-19 disease risk and severity".
In: *Nature Communications* 11.1, pp. 1–12.

Guo, Zijian et al. (2018).
"Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting".
In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 80.4, pp. 793–815.
arXiv: 1603.05224.

Haldane, John B S (1919).
"The combination of linkage values and the calculation of distances between the loci of linked factors".
In: *Journal of Genetics* 8.29, pp. 299–309.

Hartwig, Fernando P, Linbo Wang, et al. (2022).
"Average causal effect estimation via instrumental variables: the no simultaneous heterogeneity assumption".
In: *arXiv*, pp. 1–36.

Hartwig, Fernando Pires, George Davey Smith, and Jack Bowden (2017).
"Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption".
In: *International Journal of Epidemiology* 46.6, pp. 1985–1998.

Hartwig, Fernando Pires, Neil Martin Davies, and George Davey Smith (2018).
"Bias in Mendelian randomization due to assortative mating".
In: *Genetic Epidemiology* 42.7, pp. 608–620.

Heckman, James J and Ganesh Karapakula (2019).
"The Perry Preschoolers at Late Midlife: A Study in Design-Specific Inference".
In: *National Bureau of Economic Research Working Paper Series* No. 25888.6, pp. 14–21.

Heckman, James J. (1979).
"Sample selection bias as a specification error".

In: *Econometrica* 47.1, pp. 153–161.

Hemani, Gibran, Jack Bowden, and George Davey Smith (2018).
"Evaluating the potential role of pleiotropy in Mendelian randomization studies".
In: *Human Molecular Genetics* 27.R2, R195–R208.

Hemani, Gibran, Jie Zheng, et al. (2018).
"The MR-base platform supports systematic causal inference across the human phenome".
In: *eLife* 7, pp. 1–29.

Hernán, Miguel A and James M Robins (2006).
"Instruments for causal inference: an epidemiologist's dream?"
In: *Epidemiology*, pp. 360–372.

— (2020).
Causal inference: what if.
Boca Raton: Chapman & Hall/CRC.

Heyne, H. O. et al. (2023).
"Mono- and biallelic variant effects on disease at biobank scale".
In: *Nature* 613.7944, pp. 519–525.

Holland, Paul W. (1986).
"Statistics and causal inference".
In: *Journal of the American Statistical Association* 81.396, pp. 945–960.

Horvitz, D G and DJ Thompson (1952).
"A generalization of sampling without replacement from a finite universe".
In: *Journal of the American Statistical Association* 44.260, pp. 663–685.

Howe, Laurence J., Daniel J. Lawson, et al. (2019).
"Genetic evidence for assortative mating on alcohol consumption in the UK Biobank".
In: *Nature Communications* 10.1.

Howe, Laurence J., Michel G. Nivard, et al. (2022).
"Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects".
In: *Nature Genetics* 54.5, pp. 581–592.

Hu, Yingyao and Susanne M. Schennach (2008).
"Instrumental variable treatment of nonclassical measurement error models".
In: *Econometrica* 76.1, pp. 195–216.

Huang, Jonathan Yinhao (2021).
"Representativeness is not representative: addressing major inferential threats in the UK Biobank and other big data repositories".
In: *Epidemiology* 32.2, pp. 189–193.

Hughes, Rachael A. et al. (2019).

"Selection bias when estimating average treatment effects using one-sample instrumental variable analysis".
In: *Epidemiology* 30.3, pp. 350–357.

Ichimura, Hidehiko (1993).
"Semiparametric least squares (SLS) and weighted SLS estimation of single-index models".
In: *Journal of Econometrics* 58.1-2, pp. 71–120.

Imbens, Guido W and Donald B Rubin (2015).
Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.
Cambridge: Cambridge University Press.

Imbens, Guido W. (2014).
"Instrumental variables: An econometrician's perspective".
In: *Statistical Science* 29.3, pp. 323–358.
arXiv: 1410.0163.

— (2020).
"Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics".
In: *Journal of Economic Literature* 58.4, pp. 1129–1179.
arXiv: 1907.07271.

Imbens, Guido W. and Joshua D. Angrist (1994).
"Identification and Estimation of Local Average Treatment Effects".
In: *Econometrica* 62.2, p. 467.

Imbens, Guido W. and Charles F. Manski (2004).
"Confidence intervals for partially identified parameters".
In: *Econometrica* 72.6, pp. 1845–1857.

Jin, Jin et al. (2021).
"Mendelian randomization analysis using multiple biomarkers of an underlying common exposure".
In: *bioRxiv*, p. 2021.02.05.429979.

Kang, Hyunseung, Laura Peck, and Luke Keele (2018).
"Inference for instrumental variables: A randomization inference approach".
In: *Journal of the Royal Statistical Society. Series A: Statistics in Society* 181.4, pp. 1231–1254.
arXiv: 1606.04146.

Kang, Hyunseung, Anru Zhang, et al. (2016a).
"Instrumental Variables Estimation With Some Invalid Instruments and Its Application To Mendelian Randomization".
In: *Journal of the American Statistical Association* 111.513, pp. 132–144.

— (2016b).

"Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization".
In: *Journal of the American Statistical Association* 111.513, pp. 132–144.
arXiv: 1401.5755.

Katan, Martjin B. (1986).
"Apolipoprotein E isoforms, serum cholesterol, and cancer".
In: *International Journal of Epidemiology* 33.1, p. 9.

Kitagawa, Toru (2015).
"A Test for Instrument Validity".
In: *Econometrica* 83.5, pp. 2043–2063.

Klein, Roger W. and Richard H. Spady (1993).
"An efficient semiparametric estimator for binary response models".
In: *Econometrica* 61.2, pp. 387–421.

Kolesár, Michal et al. (2015).
"Identification and Inference With Many Invalid Instruments".
In: *Journal of Business & Economic Statistics* 33.4, pp. 474–484.

Kong, Augustine et al. (2018).
"The nature of nurture: Effects of parental genotypes".
In: *Science* 359, pp. 424–428.

Lai, Benjamin et al. (2022).
"Causal relationships between gout and hypertension : a bidirectional Mendelian randomisation study with coarsened exposures".
In: *Research Square*, pp. 1–22.

Lander, Eric S and Nicholas J Schork (1994).
"Genetic dissection of complex traits".
In: *Science* 265, pp. 2037–2048.

Lauritzen, Steffen L. and Nuala A. Sheehan (2003).
"Graphical models for genetic analyses".
In: *Statistical Science* 18.4, pp. 489–514.

Lawlor, Debbie A., Kate Tilling, and George Davey Smith (2017).
"Triangulation in Aetiological Epidemiology".
In: *International Journal of Epidemiology* nil.nil, dyw314.

Lawlor, Debbie A., Kate Tilling, and George Davey Smith (2016).
"Triangulation in aetiological epidemiology".
In: *International Journal of Epidemiology* 45.6, pp. 1866–1886.

Lawn, Rebecca B. et al. (2019).
"Schizophrenia risk and reproductive success: A Mendelian randomization study".
In: *Royal Society Open Science* 6.3.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart (2011).
"Weight trimming and propensity score weighting".
In: *PLoS ONE* 6.3.

Lee, Sang Hong, Michael E. Goddard, et al. (2012).
"A better coefficient of determination for genetic profile analysis".
In: *Genetic Epidemiology* 36.3, pp. 214–224.

Lewbel, Arthur (2000).
"Identification of the binary choice model with misclassification".
In: *Econometric Theory* 16.4, pp. 603–609.

— (2019).
"The identification zoo: meanings of identification in econometrics".
In: *Journal of Economic Literature* 57.4, pp. 835–903.

Liu, Zhonghua et al. (2022).
"Mendelian randomization mixed-scale treatment effect robust identification and estimation for causal inference".
In: *Biometrics*.

Locke, A. E. et al. (2015).
"Genetic studies of body mass index yield new insights for obesity biology".
In: *Nature* 518.7538, pp. 197–206.

Lower, G. M. et al. (1979).
"N-Acetyltransferase Phenotype and Risk in Urinary Bladder Cancer: Approaches in Molecular Epidemiology. Preliminary Results in Sweden and Denmark".
In: *Environmental Health Perspectives* 29.nil, pp. 71–79.

Lu, Haidong et al. (2022).
"Toward a clearer definition of selection bias when estimating causal effects".
In: *Epidemiology* 33.5, pp. 699–706.

Lyall, Donald M. et al. (2017).
"Association of body mass index with cardiometabolic disease in the UK biobank: a Mendelian randomization study".
In: *JAMA Cardiology* 2.8, pp. 882–889.

Manski, Charles F. (2003).
Partial identification of probability distributions.
New York: Springer-Verlag,
Pp. 1–175.

Marshall, John (2016).
"Coarsening bias: How coarse treatment measurement upwardly biases instrumental variable estimates".
In: *Political Analysis* 24.2, pp. 157–171.

Martins-Silva, Thais et al. (2019).
"Assessing causality in the association between attention-deficit/hyperactivity disorder and obesity: a Mendelian randomization study".
In: *International Journal of Obesity*, pp. 2500–2508.

Menni, Cristina et al. (2020).
"Real-time tracking of self-reported symptoms to predict potential COVID-19."
In: *Nature medicine*.

Millwood, Iona Y. et al. (2019).
"Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China".
In: *The Lancet* 393.10183, pp. 1831–1842.

Miratrix, L. W., S. Wager, and J. R. Zubizarreta (2018).
"Shape-constrained partial identification of a population mean under unknown probabilities of sample selection".
In: *Biometrika* 105.1, pp. 103–114.

Mitchell, Ruth E. et al. (2022).
"Strategies to investigate and mitigate collider bias in genetic and Mendelian randomization studies of disease progression".
In: *medRxiv*, pp. 1–33.

Molinari, Francesca (2020).
"Microeconometrics with partial identification".
In: *Handbook of Econometrics*.
Vol. 7.
Elsevier B.V.,
Pp. 355–486.
arXiv: 2004.11751.

Morton, Newton E (1955).
"Sequential tests for the detection of linkage".
In: *American Journal of Human Genetics* 7.3, pp. 277–318.

Mourifié, Ismael and Yuanyuan Wan (2017).
"Testing local average treatment effect assumptions".
In: *Review of Economics and Statistics* 99.2, pp. 305–313.

Munafò, Marcus R. et al. (2018).
"Collider scope: When selection bias can substantially influence observed associations".
In: *International Journal of Epidemiology* 47.1, pp. 226–235.

Nadeau, Joseph H. (2017).
"Do gametes woo? Evidence for their nonrandom union at fertilization".
In: *Genetics* 207.2, pp. 369–387.

Nevo, Aviv (2003).
"Using weights to adjust for sample selection when auxiliary information is available".
In: *Journal of Business and Economic Statistics* 21.1, pp. 43–52.

Newey, W K and S Stouli (2022).
"Heterogeneous coefficients, control variables and identification of multiple treatment effects".
In: *Biometrika* 109.3, pp. 865–872.
arXiv: 2009.02314.

Neyman, Jerzy (1990).
"On the application of probability theory to agricultural experiments. Essay on principles. Section 9."
In: *Statistical Science* 5.4, pp. 465–480.

Otto, Sarah P. and Bret A. Payseur (2019).
"Crossover interference: Shedding light on the evolution of recombination".
In: *Annual Review of Genetics* 53, pp. 19–44.

Pasman, Joëlle A. et al. (2018).
"GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal influence of schizophrenia liability".
In: *Nature Neuroscience* 21.9, pp. 1161–1170.

Patterson, Nick, Alkes L. Price, and David Reich (2006).
"Population Structure and Eigenanalysis".
In: *PLoS Genetics* 2.12, e190.

Pearl, Judea (1988).
Probabilistic reasoning in intelligent systems: networks of plausible inference.
San Mateo, California: Morgan Kaufmann,
P. 552.

— (1995).
"Causal diagrams for empirical research".
In: *Biometrika* 82.4, pp. 669–688.

— (2000).
Causality: models, reasoning, and inference.
Cambridge: Cambridge University Press.

— (2009).
Causality.
2nd ed.
New York: Cambridge University Press,
P. 478.

Pirastu, Nicola et al. (2021).
"Genetic analyses identify widespread sex-differential participation bias".

In: *Nature Genetics* 53.5, pp. 663–671.

Pitman, E. J. G. (1937).

"Significance Tests Which May Be Applied To Samples From Any Populations".

In: *Supplement to the Journal of the Royal Statistical Society* 4.1, pp. 119–130.

Plump, Andrew and George Davey Smith (2019).

"Identifying and validating new drug targets for stroke and beyond: can Mendelian randomization help?"

In: *Circulation* 140.10, pp. 831–835.

Price, Alkes L. et al. (2008).

"Long-Range LD Can Confound Genome Scans in Admixed Populations".

In: *American Journal of Human Genetics* 83.1, pp. 132–135.

Rekaya, Romdhane et al. (2016).

"Analysis of binary responses with outcome-specific misclassification probability in genome-wide association studies".

In: *Application of Clinical Genetics* 9, pp. 169–177.

Richardson, Thomas S and James M Robins (2013a).

"Single world intervention graphs: A primer".

In: *Second UAI Workshop on Causal Structural Learning.*

Bellevue, Washington.

Richardson, Thomas S., Robin J. Evans, and James M. Robins (2011).

"Transparent parametrizations of models for potential outcomes".

In: *Bayesian Statistics* 9, pp. 569–610.

Richardson, Tom G., Eleanor Sanderson, et al. (2020).

"Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study".

In: *The BMJ* 369, pp. 1–12.

Richardson, Tom S and James M Robins (2013b).

"Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

Richmond, Rebecca C. et al. (2019).

"Investigating causal relations between sleep traits and risk of breast cancer in women: Mendelian randomisation study".

In: *British Medical Journal* 365, pp. 1–12.

Rose, Sherri and Mark J van der Laan (2008).

"Simple optimal weighting of cases and controls in case-controls studies".

In: *The International Journal of Biostatistics* 4.1.

Rosenbaum, Paul (2021).

Replication and evidence factors in observational studies.

CRC Press.

Rosenbaum, Paul R (2002a).
Observational Studies.
2nd.
New York: Springer-Verlag,
P. 375.

— (2004).
Randomization inference with an instrumental variable.

— (2010).
"Evidence factors in observational studies".
In: *Biometrika* 97.2, pp. 333–345.

— (2017).
Observation & Experiment.
2nd.
Cambridge, Mass.: Harvard University Press,
P. 374.

— (1987).
"Sensitivity analysis for certain permutation inferences in matched observational studies".
In: *Biometrika* 74.1, pp. 13–26.

— (2002b).
"Covariance adjustment in randomized experiments and observational studies".
In: *Statistical Science* 17.3, pp. 286–327.

Rosenbaum, Paul R. and Donald B. Rubin (1983).
"The central role of the propensity score in observational studies for causal effects".
In: *Biometrika* 70.1, pp. 41–55.

Rosenberger, William F., Diane Uschner, and Yanying Wang (2019).
"Randomization: The forgotten component of the randomized clinical trial".
In: *Statistics in Medicine* 38.1, pp. 1–12.

Rubin, Donald B (1974).
"Estimating causal effects of treatment in randomized and nonrandomized studies".
In: *Journal of Educational Psychology* 66.5, pp. 688–701.

— (1980).
"Comment: 'Randomization analysis of experimental data: The Fisher randomization test'".
In: *Journal of the American Statistical Association* 75.371, pp. 591–593.

Sanderson, Eleanor et al. (2022).
"Mendelian randomization".
In: *Nature Reviews Methods Primers* 2.6.

Sargan, John D (1958).

"The estimation of economic relationships using instrumental variables".
In: *Econometrica* 26.3, pp. 393–415.

Sargan, John D (1988).
Lectures on advanced econometric theory.
New York and Oxford: Blackwell,
P. 176.

Scharfstein, Daniel O, Andrea Rotnitzky, and James M Robins (1999).
"Adjusting for nonignorable drop-out using semiparametric nonresponse models".
In: *Journal of the American Statistical Association* 94.448, pp. 1096–1120.

Schennach, Susanne M. (2020).
Mismeasured and unobserved variables.
Vol. 7.
Elsevier B.V.,
Pp. 487–565.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015).
"Biological insights from 108 schizophrenia-associated genetic loci".
In: *Nature* 511.7510, pp. 421–427.

Shapiro, Alexander (1991).
"Asymptotic analysis of stochastic programs".
In: *Annals of Operations Research* 30, pp. 169–186.

Shapiro, Alexander, Darinka Dentcheva, and Andrzej Ruszczyński (2009).
Lectures on Stochastic Programming: Modeling and Theory.
Society for Industrial, Applied Mathematics, and the Mathematical Programming Society.

Signorovitch, James E. et al. (2012).
"Matching-adjusted indirect comparisons: A new tool for timely comparative effectiveness research".
In: *Value in Health* 15.6, pp. 940–947.

Sirugo, Giogio, Scott M Williams, and Sarah A Tishkoff (2019).
"The Missing Diversity in Human Genetic Studies".
In: *Cell* 177.1, pp. 26–31.

Smeden, Maarten van, Timothy L. Lash, and Rolf H.H. Groenwold (2020).
"Reflection on modern methods: Five myths about measurement error in epidemiological research".
In: *International Journal of Epidemiology* 49.1, pp. 338–347.

Smith, Louisa H. (2020).
"Selection Mechanisms and Their Consequences: Understanding and Addressing Selection Bias".
In: *Current Epidemiology Reports* 7.4, pp. 179–189.

Smith, Louisa H. and Tyler J. VanderWeele (2019).
"Bounding bias due to selection".
In: *Epidemiology* 30.4.
arXiv: 1810.13402.

Smith, Shannon, El H. Hay, et al. (2013).
"Genome wide association studies in presence of misclassified binary responses".
In: *BMC Genetics* 14.

Song, Suyong, Susanne M. Schennach, and Halbert White (2015).
"Estimating nonseparable models with mismeasured endogenous variables".
In: *Quantitative Economics* 6.3, pp. 749–794.

Spady, Richard H (2007).
"Semiparametric methods for the measurement of latent attitudes and the estimation of their behavioral consequences."

Spielman, R. S., R. E. McGinnis, and W. J. Ewens (1993).
"Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM)".
In: *American Journal of Human Genetics* 52.3, pp. 506–516.

Spirtes, Peter, Clark N Glymour, and Richard Scheines (2000).
Causation, prediction, and search.
MIT press.

Stangl, Anne L. et al. (2019).
"The Health Stigma and Discrimination Framework: A global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas".
In: *BMC Medicine* 17.1, pp. 18–23.

Stock, James H. and Francesco Trebbi (2003).
"Retrospectives: Who invented instrumental variable regression?"
In: *Journal of Economic Perspectives* 17.3, pp. 177–194.

Stolley, Paul D (1991).
"Re: "when genius errs: R. A. Fisher and the lung cancer controversy"".
In: *American Journal of Epidemiology* 133.5, pp. 416–425.

Stoye, Jörg (2009).
"More on confidence intervals for partially identified parameters".
In: *Econometrica* 77.4, pp. 1299–1315.

Stuart, Elizabeth A. et al. (2011).
"The use of propensity scores to assess the generalizability of results from randomized trials".
In: *Journal of the Royal Statistical Society. Series A: Statistics in Society* 174.2, pp. 369–386.

Sun, Yi Qian et al. (2019).

"Adiposity and asthma in adults: A bidirectional Mendelian randomisation analysis of the HUNT Study".
In: *Thorax*, pp. 1–7.

Swanson, Sonja A. et al. (2015).
"Bounding the per-protocol effect in randomized trials: An application to colorectal cancer screening".
In: *Trials* 16.1, pp. 1–11.

Swerdlow, Daniel I. et al. (2016).
"Selecting instruments for Mendelian randomization in the wake of genome-wide association studies".
In: *International Journal of Epidemiology* 45.5, pp. 1600–1616.

Taubes, Gary (1995).
"Epidemiology faces its limits".
In: *Science* 269, pp. 164–169.

Tchetgen, Eric Tchetgen, Bao Luo Sun, and Stefan Walter (2021).
"The GENIUS Approach to Robust Mendelian Randomization Inference".
In: *Statistical Science* 36.3, pp. 443–464.
arXiv: 1709.07779.

Thomas, Duncan C. and David V. Conti (2004).
"Commentary: The concept of 'Mendelian randomization'".
In: *International Journal of Epidemiology* 33.1, pp. 21–25.

Thompson, Caroline A and Onyebuche A Arah (2014).
"Selection bias modeling using observed data augmented with imputed record-level probabilities".
In: *Annals of Epidemiology* 24.10, pp. 747–753.

Thompson, Elizabeth A (2000).
"Statistical inference from genetic data on pedigrees".
In: *NSF-CBMS Regional Conference Series in Probability and Statistics*.
Vol. 6,
Pp. 1–169.

Tian, Haodong et al. (2022).
"Relaxing parametric assumptions for non-linear Mendelian randomization using a doubly-ranked stratification method".
In: *bioRxiv*, pp. 1–36.

Tudball, Matthew J, Jack Bowden, et al. (Aug. 2021).
"Mendelian randomization with coarsened exposures".
In: *Genetic Epidemiology* 45, pp. 338–350.
arXiv: /doi.org/10.1002/gepi.22376 [https:].

Tudball, Matthew J, Rachael A Hughes, et al. (2022).
"Sample-constrained partial identification with application to selection bias".
In: *Biometrika* 109.3.

Vogelezang, Suzanne et al. (2020).
"Novel loci for childhood body mass index and shared heritability with adult cardiometabolic traits".
In: *PLoS Genetics*, pp. 1–26.

Wang, Gao, Abhishek Sarkar, et al. (2020).
"A simple new approach to variable selection in regression, with application to genetic fine mapping".
In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 82.5, pp. 1273–1300.

Wang, Jingshu and Art B. Owen (2019).
"Admissibility in Partial Conjunction Testing".
In: *Journal of the American Statistical Association* 114.525, pp. 158–168.
eprint: 1508.00934.

Wang, Jingshu, Qingyuan Zhao, et al. (2021).
"Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments".
In: *PLoS Genetics* 17.6, pp. 1–24.

Wang, Linbo and Eric Tchetgen Tchetgen (2018).
"Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables".
In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 80.3, pp. 531–550.
arXiv: 1611.09925.

Wang, Qin, Tom G Richardson, et al. (2022).
"A phenome-wide bidirectional Mendelian randomization analysis of atrial fibrillation".
In: *International Journal of Epidemiology* March, pp. 1153–1166.

Wang, Wei and Shabbir Ahmed (2008).
"Sample average approximation of expected value constrained stochastic programs".
In: *Operations Research Letters* 36.5, pp. 515–519.

Wheatley, Keith and Richard Gray (2004).
"Commentary: Mendelian randomization - An update on its use to evaluate allogeneic stem cell transplantation in leukemia".
In: *International Journal of Epidemiology* 33.1, pp. 15–17.

Williamson, Elizabeth J. et al. (2020).
"Factors associated with COVID-19-related death using OpenSAFELY".

In: *Nature* 584.7821, pp. 430–436.

Wright, John et al. (2013).
"Cohort profile: The born in bradford multi-ethnic family cohort study".
In: *International Journal of Epidemiology* 42.4, pp. 978–991.

Wright, Philip G (1928).
The Tariff on Animal and Vegetable Oils.
New York: MacMillan,
P. 347.

Wright, Sewall (1920).
"The relative importance of heredity: determining the piebald pattern of guinea pigs".
In: *Proceedings of the National Academy of Sciences* 6.6, pp. 320–332.

— (1923).
"The theory of path coefficients: a reply to Niles' criticism".
In: *Genetics* 8.3, pp. 239–255.

Young, Alexander L. et al. (2022).
"Mendelian imputation of parental genotypes improves estimates of direct genetic effects".
In: *Nature Genetics* 54, pp. 897–905.

Zelen, Marvin (1977).
"A New Design for Randomized Clinical Trials".
In: *New England Journal of Medicine* 300, pp. 1242–1245.

Zhang, Yao and Qingyuan Zhao (2022).
"What is a randomization test?"
In: *arXiv*, pp. 1–26.

Zhao, Anqi, Youjin Lee, et al. (2022).
"Evidence Factors From Multiple, Possibly Invalid, Instrumental Variables".
In: *The Annals of Statistics* 50.3, nil.

Zhao, Qingyuan, Dylan S. Small, and Bhaswar B. Bhattacharya (2019).
"Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap".
In: *Journal of the Royal Statistical Society: Series B* 81.3, pp. 1–27.

Zhao, Qingyuan, Jingshu Wang, et al. (2020).
"Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score".
In: *Annals of Statistics* 48.3, pp. 1742–1769.

Zollner, Linda et al. (2022).
"Mendelian Randomization Analysis of the Relationship Between Native American Ancestry and Gallbladder Cancer Risk".
In: *medRxiv*, pp. 1–27.