Almost exact Mendelian randomization

Matthew J Tudball[1], George Davey Smith[1], and Qingyuan Zhao[2]

[1]MRC Integrative Epidemiology Unit, University of Bristol
[2]Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge

August 31, 2022

Abstract

Mendelian randomization (MR) is an observational design based on the random transmission of genes from parents to offspring. However, this inferential basis is typically only implicit or used as an informal justification. As parent-offspring data becomes more widely available, we 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. We 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. We then propose a randomization test for causal hypotheses in the MR design by using precisely the exogenous randomness in meiosis and fertilization. We 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. We demonstrate via simulation that propensity scores obtained from the underlying meiosis model can form powerful test statistics. Besides transparency and conceptual appeals, our 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 linkage disequilibrium with pleiotropic variants. We conclude with a negative and positive control analysis in the Avon Longitudinal Study of Parents and Children using our R package `almostexactmr` (`https://github.com/matt-tudball/almostexactmr`).

1 Introduction

1.1 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 heridity", 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 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 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). Within this framework, there has been considerable recent methodological work to replace the exclusion restriction with more plausible assumptions, typically by placing structure on the sparsity (Kang, Zhang, et al. 2016) or distribution of pleiotropic effects across individual genetic variants (Bowden, Davey Smith, and Burgess 2015; Zhao et al. 2020; Kolesár et al. 2015).

1.2 Towards an almost exact inference for MR

As parent-offspring trio data becomes more widely available, it is increasingly feasible to perform MR within families, as originally intended. There has been some recent methodological and applied development for within-family designs (Davies 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 regression 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 an observational design is that it has an explicit inferential basis, randomness in meiosis and fertilization, which has been thoroughly studied and modelled in genetics since 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 identifying causal genetic variants (Bates et al. 2020) and was implicit in earlier genetic linkage analysis (Morton 1955) and the transmission disequilibrium test (Spielman, McGinnis, and Ewens 1993). Lauritzen and Sheehan (2003) attempted to represent meiosis models using graphs; however, they were concerned with computatational advantages of graphical models and did not consider their potential for causal inference.

The idea of exact hypothesis testing 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 embraced the idea of randomization testing. 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 et al. 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).

1.3 Our contributions

In this article, we propose a statistical framework that enables us to use meiosis models as the "reasoned basis" for inference in MR by unifying several ideas mentioned above. The randomization test we propose is *almost exact* in the sense that the test has exactly the nominal size if the meiosis and fertilization model is correct.

Our first contribution is a theoretical description of MR (and the assumptions therein) via the language of causal directed acyclic graphs (DAGs) (Pearl 2009). These graphical tools allow us to visualize and dissect the assumptions imposed on the biological processes involved in heredity. 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 2013), we identify sufficient confounder adjustment sets to eliminate these sources of bias. Our results provide important theoretical insights into a trade-off between reducing pleiotropy-induced bias and increasing statistical power.

For statistical inference, we propose 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 and Rubin 1983; 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 et al. 2020). We show how the hidden Markov model for meiosis and fertilization implied by Haldane (1919) greatly simplifies the sufficient adjustment sets and computation of the randomization test. In essence, our proposal extends existing randomization inference techniques for instrumental variables to allow testing based on biological randomness in reproduction (i.e. Mendelian randomization).

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 an accurate phenotype model to dramatically improve its power. Furthermore, the hidden Markov model based on Haldane's original formulation implies a propensity score for each 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. Second, 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-ethnic samples), assortative mating, dynastic effects and pleiotropy by linkage.

We demonstrate these advantages with a simulation study and real data example in the Avon Longitudinal Study of Parents and Children (ALSPAC). The simulation study first confirms that our almost exact test produces uniformly-distributed p-values under the null and then 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, backdoor paths could exist to produce a false rejection of the null. We provide evidence that our almost exact test closes these paths. The positive control is the effect of child's BMI on itself plus some zero-mean noise. We also compare our results with the results from a "typical" MR analysis unconditional on any parental or offspring haplotypes.

2 Background

2.1 Causal inference preliminaries

This section lays out some standard notation and assumptions in causal inference. Readers looking for an introduction to causal inference concepts, including causal graphical models, single world intervention graphs, randomization inference, and instrumental variables, can consult Appendix A. We express our causal assumptions and model via causal graphs, then demonstrate that randomization inference 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.

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 resulting incidence of cardiovascular disease. Individual i's *potential outcomes* (also called *counterfactuals*) corresponding to exposure level $D_i = d$ are given by $Y_i(d)$. The collection of potential outcomes for the sample is given by $\mathcal{F} = \{(Y_i(0), Y_i(1)) \colon i = 1, 2, \ldots, N\}$.

We make the *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), such that $Y_i(d) \perp\!\!\!\perp D_j$ for all $i \neq j$ and d in the support of D_i. We also assume *no hidden versions of the same treatment*. A violation of this assumption could occur if $D_i \in \{0, 1\}$ were a binary measure indicating abstinence ($D_i = 0$) or some alcohol consumption ($D_i = 1$). The effect of alcohol consumption on cardiovascular disease is likely to exhibit a dose-response relationship, meaning that the potential outcome $Y_i(1)$ is not well-defined since it could take multiple distinct values depending on the unobserved amount of consumption. The previous two assumptions are sometimes jointly referred to as the *stable unit treatment value assumption* (Rubin 1980). We also make the *consistency* assumption (Hernán and Robins 2020) which states that the observed outcome corresponds to the potential outcome at the realized exposure level $Y_i = Y_i(D_i)$.

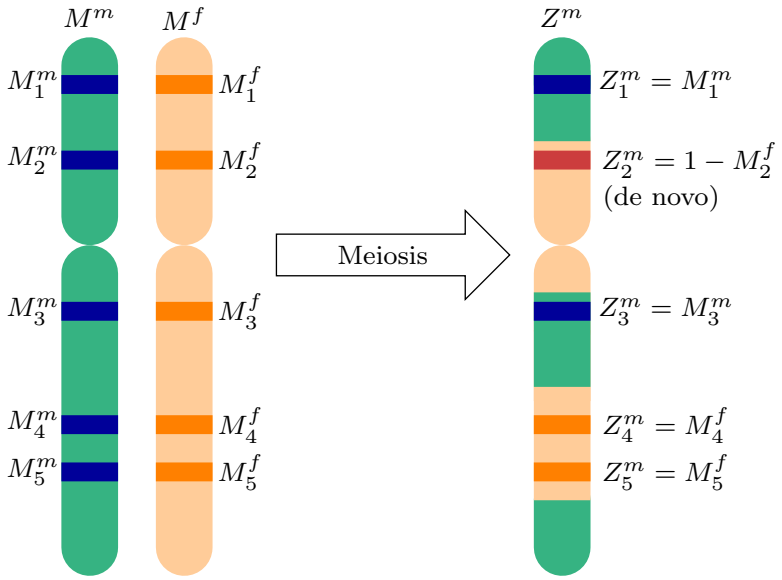
2.2 Genetic preliminaries

Before we proceed, it is instructive to provide a basic overview of the relevant concepts in genetics, with a focus on modelling the processes involved in genetic inheritance, namely *meiosis* and *fertilization*. For a thorough exposition on statistical models for meiosis and 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. Each chromosome is a doubled strand of helical DNA comprised 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 the 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* (Bates et al. 2020) and the two haplotypes forming a homologous pair of chromosomes is called a *genotype*.

Meiosis is a type of cell division that results in reproductive 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 *crossover*. The recombined chromosomes are then further divided and separated into gametes. Since recombinations 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 results in *genetic linkage*. Fertilization is the

Figure 1: Illustration of the meiotic process for five sites on a chromosome.



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

In genetic trio studies we observe the haplotypes of the mother, father and their child at p SNPs on a single chromosome, where $\mathcal{J} = \{1, 2, \ldots, p\}$ is the set of SNP indices. We will denote the haplotypes as follows:

$$\begin{aligned}
&\text{Individual's haplotypes:} && \boldsymbol{Z}^m = (Z_1^m, \ldots, Z_p^m) \in \{0,1\}^p, && \boldsymbol{Z}^f = (Z_1^f, \ldots, Z_p^f) \in \{0,1\}^p;\\
&\text{Mother's haplotypes:} && \boldsymbol{M}^m = (M_1^m, \ldots, M_p^m) \in \{0,1\}^p, && \boldsymbol{M}^f = (M_1^f, \ldots, M_p^f) \in \{0,1\}^p;\\
&\text{Father's haplotypes:} && \boldsymbol{F}^m = (F_1^m, \ldots, F_p^m) \in \{0,1\}^p, && \boldsymbol{F}^f = (F_1^f, \ldots, F_p^f) \in \{0,1\}^p,
\end{aligned}$$

where the superscript m (or f) indicates a maternally (or paternally) inherited haplotype. Furthermore, denote $\boldsymbol{M}_j^{mf} = (M_j^m, M_j^f)$ as the mother's haplotypes at site j and similarly for $\boldsymbol{F}_j^{mf}$ and $\boldsymbol{Z}_j^{mf}$. The offspring's genotype at site $j \in \mathcal{J}$ 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 1 illustrates how an offspring's maternally-inherited haplotype $\boldsymbol{Z}^m$ at five sites on a chromosome are related to the mother's haplotypes $\boldsymbol{M}^m$ and $\boldsymbol{M}^f$. At site $j \in \mathcal{J}$ in a gamete produced by meiosis, the allele is inherited from either the mother's m haplotype or f haplotype (ignoring 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 homogeneous Poisson process. Haldane's model is described in Appendix B in detail and can simplify our method considerably (Section 3.5). Nonetheless, our "almost exact" MR framework is modular and does not rely on a specific meiosis model. In fact, 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 assume that there is a small probability of independent mutations. This is formalized in the next assumption.

Assumption 1 (Haldane's model)**.** Given that $U_j^m = u_j^m \in \{m, f\}$ and fertilization occurs (this is represented as $S = 1$ in Section 3), each Z_j^m is equal to $M_j^{(u_j^m)}$ unless an independent mutation occurs. More specifically,

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

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). Unless it is necessary to compute the exact randomization distribution under a recombination model, for practical purposes it often suffices to treat $\epsilon = 0$ (i.e., no mutations).

This meiosis model assumes the absence of *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%. 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 (Davies et al. 2019). While we sidestep this discussion for now, we return to it in Section 6.

3 Almost exact Mendelian randomization

3.1 A causal model for Mendelian inheritance

Returning to the alcohol and cardiovascular disease example in Section 2.1, observational studies suggest that moderate alcohol consumption confers reduced risk relative to abstinence or heavy consumption (Millwood et al. 2019). This could indicate systematic differences among people with different drinking patterns (confounding) rather than a causal effect. There is a genetic variant in the *ALDH2* gene which regulates acetaldehyde metabolism. In some populations, an allele of *ALDH2* produces a protein that is inactive in metabolising acetaldehyde, causing discomfort while drinking and thereby reducing consumption. We might like to use the random allocation of variant copies of *ALDH2* during meiosis and fertilization to make causal inference about the downstream effect of alcohol consumption on cardiovascular disease, however, we need to carefully clarify the conditions under which this inference would be valid. To this end, we construct a very general causal model in this section 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. Technically speaking, we would like to apply the randomization test described in Appendix A.3 to MR.

Figure 2 shows a working example of our causal model on a chromosome with just $p = 3$ 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 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 3). By definition, a joint distribution *factorizes* according to the DAG in

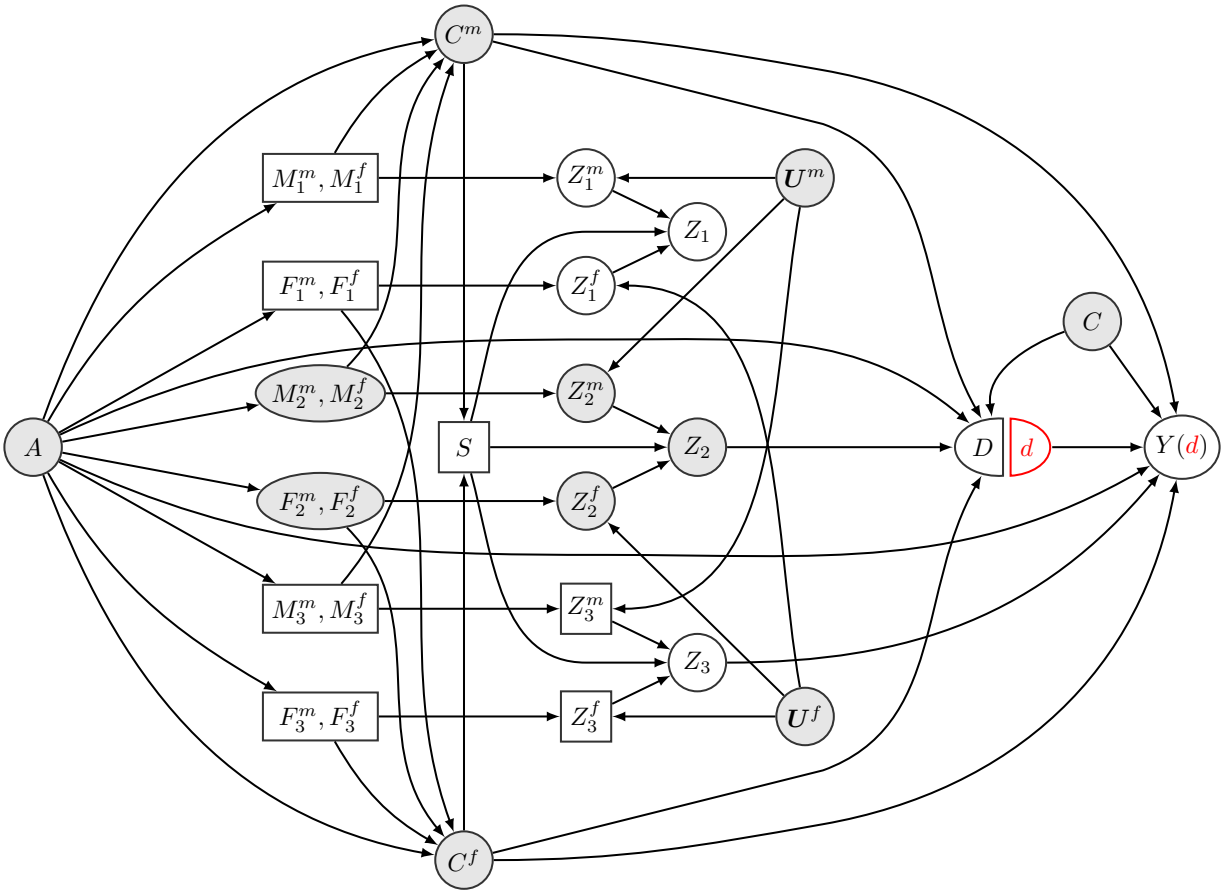


Figure 2: The single world intervention graph for a working example of a chromosome with $p = 3$ variants. Transparent nodes are observed and grey nodes are unobserved. Square nodes are the confounders being conditioned on in Proposition 2. A is ancestry; $\boldsymbol{M}^f = (M_1^f, M_2^f, M_3^f)$ is 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 2 if its density function can be decomposed as (let f be a generic symbol for density function)

$$
\begin{aligned}
& f(\text{all variables}) \\
=& f(A)f(\boldsymbol{U}^m)f(\boldsymbol{U}^f)f(C) && \text{(Exogenous variables)} \\
& f(\boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f \mid A) && \text{(Population stratification, Section 3.1.1)} \\
& f(C^m \mid A, \boldsymbol{M}^m, \boldsymbol{M}^f) f(C^f \mid A, \boldsymbol{F}^m, \boldsymbol{F}^f) && \text{(Parental phenotypes, Section 3.1.2)} \\
& f(\boldsymbol{Z}^m \mid \boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{U}^m) f(\boldsymbol{Z}^f \mid \boldsymbol{F}^m, \boldsymbol{F}^f, \boldsymbol{U}^f) && \text{(Meiosis, Section 3.1.3)} \\
& f(S \mid C^m, C^f) && \text{(Assortative mating, Section 3.1.3)} \\
& f(\boldsymbol{Z} \mid \boldsymbol{Z}^m, \boldsymbol{Z}^f, S) && \text{(Fertilization, Section 3.1.3)} \\
& f(D \mid A, \boldsymbol{Z}, C^m, C^f, C) f(Y(d) \mid A, \boldsymbol{Z}, C^m, C^f, C) && \text{(Offspring phenotypes, dynastic effects,} \\
& && \text{confounding, Section 3.1.4)}
\end{aligned}
$$

Next, we describe each term on the right hand side above and its corresponding subgraph and biological process. To simplify the discussion, we assume all DAGs in this article are faithful, so conditional independence between random variables is equivalent to d-separation in the DAG.

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). This can introduce spurious associations between genetic variants and traits (Lander and Schork 1994).

We represent population structure via the node A in the subgraph in Figure 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. This is formalized in the assumption below.

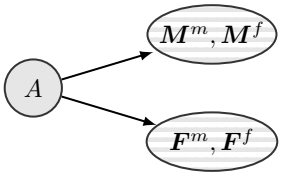
Assumption 2. The parental haplotypes $\boldsymbol{M}^m$, $\boldsymbol{M}^f$, $\boldsymbol{F}^m$, and $\boldsymbol{F}^f$ depend on the latent population structure A, so

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

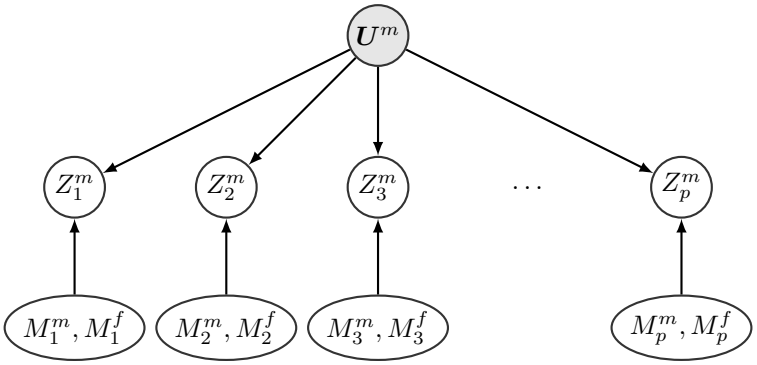
The node A may also capture any linkage disequilibrium in the parental haplotypes. That is, because the parental haplotypes are determined by the same process as the grandparental haplotypes and so on, recombination introduces dependence among nearby genetic variants. The distribution of A and the distribution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

3.1.2 Parental phenotypes

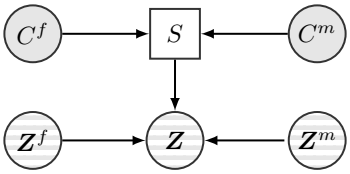
We impose no assumptions on the nature and the distribution of the parental phenotypes C^m and C^f. They can depend arbitrarily on the parental haplotypes $\boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f$ and the population structure A, once again because our proposal for almost exact MR conditions on the parental haplotypes.



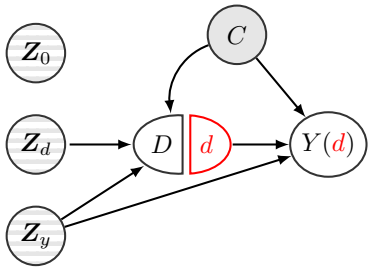
(a) Population structure (Section 3.1.1).



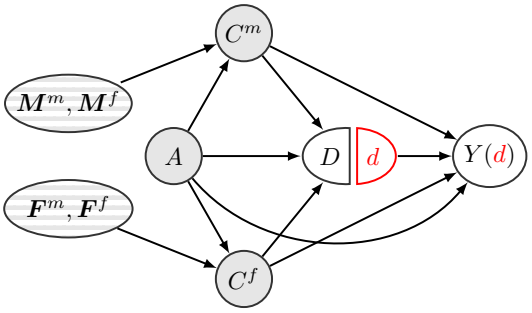
(b) Meiotic recombination of the mother's haplotypes (Section 2.2).



(c) Assortative mating (Section 3.1.3).



(d) Offspring phenotypes. (Section 3.1.4).



(e) Dynastic effects (Section 3.1.2).

Figure 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.

Assumption 3. The parental phenotypes C^m and C^f are descendants of the latent population structure A and the corresponding parental haplotypes (i.e. $(\boldsymbol{M}^m, \boldsymbol{M}^f)$ for C^m and $(\boldsymbol{F}^m, \boldsymbol{F}^f)$ for C^f). In other words, we allow the following dependence:

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

3.1.3 Offspring genotypes

There are two biological processes involved in the genesis of the offspring's genotype: meiosis (gamete formation) and fertilization. The meiotic process is briefly reviewed in Section 2.2, and the key Assumption 1 can be represented by the causal diagram in Figure 3b (for the mother). A crucial assumption underlying our almost exact inference is the exogeneity of the meiosis indicators $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$. This is reflected in Figures 2 and 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 4. 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 different models have been proposed for 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 2.2, we describe the classical model of Haldane (1919) which assumes $\boldsymbol{U}^m$ follows a Poisson process. This model has been used by Bates et al. (2020) to locate causal variants. We will see in Section 3.5 that such Markovian structure greatly simplifies 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 a SNP on the *ADH1B* gene related to alcohol consumption is more likely to be shared among spouses relative to 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 3c, where the mating indicator $S \in \{0, 1\}$ is a common child of the parental phenotypes C^m and C^f ($S = 1$ means mating). In any MR study, we necessarily condition on $S = 1$, otherwise the offspring would not exist. This is formalized in Figure 3c by the arrows from S to $\boldsymbol{Z}$. In particular, we may define the offspring's genotype $\boldsymbol{Z}$ as

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

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.

Our causal model implies that $\boldsymbol{Z}^m \perp\!\!\!\perp \boldsymbol{Z}^f \mid (\boldsymbol{M}^m, \boldsymbol{M}^f, \boldsymbol{F}^m, \boldsymbol{F}^f, S = 1)$, however, this is not necessarily a benign assumption. Indeed, there is empirical evidence that gametes may pair up non-randomly (Nadeau 2017), which could be represented by arrows from $\boldsymbol{Z}^m$ and $\boldsymbol{Z}^f$ to S. This is an example of transmission ratio distortion, which we discuss later in Section 6. For now, we simply note that we must assume the absence of this phenomenon.

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 the consistency assumption (11) before.

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 population structure, parental phenotypes and other environmental factors, it is not randomized.

Assumption 5. There may be unmeasured confounders between the exposure and outcome, so 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 as MR uses unconfounded variation (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 6. 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 (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 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 6, 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 3d provides the graphical representation of Assumption 6. 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). The view that pleiotropy is widespread, if not universal, is implied in Fisher's infinitesimal model (Fisher 1918) and supported by recent human genetic studies (Boyle, Li, and Pritchard 2017).

Dynastic effects, sometimes called *genetic nurture* (Kong et al. 2018), is a phenomenon characterized by parental phenotypes exerting a direct influence on the offspring's phenotypes. This is depicted in Figure 3e, where paths emanate from the parental haplotypes $\boldsymbol{M}^m, \boldsymbol{M}^f$ and $\boldsymbol{F}^m, \boldsymbol{F}^f$ to the parental phenotypes C^m and C^m and on to the offspring phenotypes D and Y.

3.2 Conditions for identification

With the causal model outlined in Section 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 3a), assortative mating (Figure 3c), and dynastic effects (Figure 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 (Spielman, McGinnis, and Ewens 1993; Bates et al. 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 1. *Under the causal graphical model described in Section 3.1, the offspring's 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:*

$$\begin{aligned} Z_j^m &\perp\!\!\!\perp (A, C^m, C^f, C) \mid (M_j^m, M_j^f, S = 1), \\ Z_j &\perp\!\!\!\perp (A, C^m, C^f, C) \mid (M_j^m, M_j^f, F_j^m, F_j^f, S = 1). \end{aligned} \tag{2}$$

However, the conditional independence (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, as recognized by Bates et al. (2020) in the context of mapping causal variants. 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 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 $d \in \mathcal{D}$.

Simiarly, 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 6), the exclusion restriction is satisfied if and only if $j \notin \mathcal{J}_y$.

Returning to the example in Figure 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

Table 1: Some paths between Z_1 and $Y(d)$ in Figure 2.

Name of bias	Path	Blocking variable
Dynastic effect	$Z_1^m \leftarrow M_1^m, M_1^f \rightarrow C^m \rightarrow Y(d)$	(M_1^m, M_1^f)
Population stratification	$Z_1^m \leftarrow M_1^m, M_1^f \leftarrow A \rightarrow Y(d)$	(M_1^m, M_1^f)
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	$Z_1^m \leftarrow M_1^m, M_1^f \leftarrow C^m \rightarrow \boxed{S} \leftarrow C^f \leftarrow F_3^m, F_3^f \rightarrow Z_3^f \rightarrow Z_3 \rightarrow Y(d)$	(M_1^m, M_1^f) or Z_3^f or Z_3 or (F_3^m, F_3^f)
Nearly determined ancestry	$Z_1^m \leftarrow \boldsymbol{U}^m \rightarrow \boxed{Z_3^m} \leftarrow M_3^m, M_3^f \leftarrow A \rightarrow Y(d)$	(M_3^m, M_3^f)

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 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 2. *For the example in Figure 2, the following conditional independence relationships are true for all $d \in \mathcal{D}$:*

$$Z_1^m \perp\!\!\!\perp Y(d) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{V}_{\{3\}}^m = (\boldsymbol{M}_3^{mf}, Z_3^m), S = 1), \tag{3}$$
$$Z_1 \perp\!\!\!\perp Y(d) \mid (\boldsymbol{M}_1^{mf}, \boldsymbol{F}_1^{mf}, \boldsymbol{V}_{\{3\}} = (\boldsymbol{M}_3^{mf}, \boldsymbol{F}_3^{mf}, Z_3), S = 1). \tag{4}$$

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. To show $\boldsymbol{V} = (\boldsymbol{M}_1^{mf}, \boldsymbol{V}_{\{3\}}^m)$ is minimal for (3) and better understand the potential biases in MR studies, Table 1 lists several paths between Z_1^m and $Y(d)$ that are named after the key biological mechanism involved. The table only includes the maternal paths, but the same blocking also holds for the paternal paths. □

To our knowledge, the potential bias in Table 1 due to nearly determined ancestry has not yet been identified in the literature. This is a form of collider bias introduced because the ancestry

indicator U_j^m 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 very high probability. Due to genetic linkage, there is also a high probability that $U_1^m = m$.

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 1. *Suppose $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ is a full chromosome. Consider the general causal model for Mendelian randomization in Section 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 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. □

Since Proposition 1 ensures that, after conditioning on $\boldsymbol{M}_j^{mf}$, 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 as an instrument 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 2) but not independent of the set of causal variants (condition 1 of Theorem 2). We will work with a general $\mathcal{B}$ until Section 3.5 where we describe the structure of this set. It is straightforward to extend Theorem 1 to establish validity of the genotype Z_j at locus j as an instrumental variable. Details are omitted.

3.3 Hypothesis testing

This section describes our randomization-based approach to statistical inference in Mendelian randomization studies. We begin by describing an idealized exact setting where the randomization distribution is known. We then discuss the realistic setting where the randomization distribution must be approximated by a meiosis model.

We first describe the simplest case where we use a single genetic variant from the offspring's maternally-inherited haplotype as an instrument. In particular, define the *propensity score* for some instrument Z_{ij}^m at locus j of individual i as

$$\pi_{ij}^m = \mathbb{P}(Z_{ij}^m = 1 \mid \boldsymbol{M}_{ij}^{mf}, \boldsymbol{V}_{i\mathcal{B}}^m) \tag{5}$$

where $\mathcal{B} \subseteq \mathcal{J}$. 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 chosen to satisfy the conditions in Theorem 1.

Let us consider a model for the potential outcomes of the form

$$Y_i(d) = Y_i(0) + \beta d \text{ for all } d \in \mathcal{D} \text{ and } i = 1, \ldots, N. \tag{6}$$

Let $\mathcal{F} = \{Y_i(0) \colon i = 1, \ldots, N\}$ 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

$$H_0 \colon \beta = \beta_0, \; H_1 \colon \beta \neq \beta_0 \tag{7}$$

where β_0 is some hypothetical value of the causal effect. If the null hypothesis is true, then the model (6) implies that the potential outcome under no exposure $d = 0$ can be identified from the observed data since

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

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

Theorem 2 and the model (6) imply that we are testing the following conditional independence:

$$H_0 \colon Z_{ij}^m \perp\!\!\!\perp Q_i(\beta_0) \mid (\boldsymbol{M}_{ij}^{mf}, \boldsymbol{V}_{i\mathcal{B}}^m), \quad H_1 \colon Z_{ij}^m \not\perp\!\!\!\perp Q_i(\beta_0) \mid (\boldsymbol{M}_{ij}^{mf}, \boldsymbol{V}_{i\mathcal{B}}^m). \tag{8}$$

Suppose we have selected a test statistic $T(\boldsymbol{Z}_j^m \mid \mathcal{F})$ where possible 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

$$\begin{aligned} P(\boldsymbol{Z}_j^m \mid \mathcal{F}) \; &= \tilde{\mathbb{P}}(T(\tilde{\boldsymbol{Z}}_j^m \mid \mathcal{F}) \leq T(\boldsymbol{Z}_j^m \mid \mathcal{F})) \\ &= \sum_{\tilde{\boldsymbol{z}}^m \in \{0,1\}^N} I\{T(\tilde{\boldsymbol{z}}_j^m \mid \mathcal{F}) \leq T(\boldsymbol{Z}_j^m \mid \mathcal{F})\} \prod_{\tilde{z}_i^m \in \tilde{\boldsymbol{z}}^m} (\pi_{ij}^m)^{\tilde{z}_i} (1 - \pi_{ij}^m)^{1 - \tilde{z}_i}, \end{aligned} \tag{9}$$

where $I\{\cdot\}$ is the indicator function, $\tilde{\boldsymbol{Z}}_j^m$ denotes a random draw from the distribution (5) and $\tilde{\mathbb{P}}$ denotes probability with respect to the distribution (5). Given the propensity score and the null hypothesis, this p-value can be computed exactly by enumerating over all possible values of $\tilde{\boldsymbol{Z}}^m$ or approximated by drawing $\tilde{\boldsymbol{Z}}^m$ a finite number of times from $\boldsymbol{\pi}_j^m$; 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 (9) highlights that knowledge of the propensity score $\boldsymbol{\pi}_j^m$ is essential for performing 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, the analyst. Therefore, the best we can do is perform *almost exact* inference by replacing $\boldsymbol{\pi}_j^m$ with a reasonable model-based approximation. The model we use in this paper is Haldane's hidden Markov model described in Appendix B. As discussed in Section 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.

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. Theorem 3 in Appendix B 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 (9) via the proportion of $\tilde{t}_1, \ldots, \tilde{t}_K$ which are larger than t:

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

3.4 Choosing a test statistic

Our randomization test retains exact nominal size under the null hypothesis regardless of test statistic, however, we can often improve power by selecting a test statistic that better blocks the confounding paths between Z_j and $Q(\beta_0)$. We show in Section 4.2 that test statistics that do not control for confounders can have almost no power for certain null hypotheses. We propose a powerful test statistic that can be constructed by including a so-called "clever covariate" (Scharfstein, Rotnitzky, and Robins 1999; Rose and Laan 2008) in the test statistic

$$X_j^m = \frac{Z_j^m}{\pi_j^m} - \frac{1 - Z_j^m}{1 - \pi_j^m}$$

such that

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

This covariate exploits the "central role of the propensity score" (Rosenbaum and Rubin 1983) that

$$Y(d) \perp\!\!\!\perp Z_j^m \mid \pi_j^m.$$

where π_j^m is defined as in equation (5), provided $0 < \pi_j^m < 1$. Conditioning on π_j^m blocks all confounding paths between Z_j^m and $Y(d)$. Furthermore, π_j^m reduces to a single variable the adjustment set $(\boldsymbol{M}_j^{mf}, \boldsymbol{V}_{\mathcal{B}})$ which is potentially high dimensional and highly correlated.

Alternatively, we could improve power by constructing data-driven test statistics via flexible machine learning techniques such as neural networks, gradient boosting or random forests, although this may be computationally costly (Watson and Wright 2019).

3.5 Simplification via Markovian structure

Conditional independencies implied by Haldane's model for meiosis also allow us to greatly simplify the sufficient confounder adjustment set. Theorem 1 highlights a trade-off in choosing the adjustment variables $\boldsymbol{V}_{\mathcal{B}}$: 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.

The conditions in Theorem 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 other variants to block the pleiotropic paths, as illustrated in the working example in Figure 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 allows us to conclude $Z_j \perp\!\!\!\perp Z_l \mid \boldsymbol{M}_j^{mf}, \boldsymbol{F}_j^{mf}, \boldsymbol{V}_{\{k\}}$ for all $j < k < l$ if there are no mutations and $M_k^f \neq M_k^m$.

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 loci between b_1 and b_2, which of course contains the locus j of interest.

Theorem 2. *Consider the setting in Theorem 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*

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 2 imply the first two conditions in Theorem 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 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

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

Proposition 3. *Suppose* $M_j^m \neq M_j^f$ *for all* $j \in \mathcal{B}$. *Then, under the first two assumptions in Theorem 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 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. □

3.6 Multiple instruments

Proposition 3 allows us to formalize the intuition that genetic instruments across the genome can provide independent evidence about the causal effect of the exposure, if the right loci are conditioned on.

Corollary 1. Z_j^m *and* $Z_{j'}^m$ *are independent valid instruments conditional on* $(\boldsymbol{M}_j^{mf}, \boldsymbol{M}_{j'}^{mf}, \boldsymbol{V}_{\mathcal{B}}^m)$ *if*

1. *The first two assumptions of Theorem 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 1 says that any two instruments are valid and independent if they lie within separate partitions and each partition contains a causal variant of the exposure and does not contain any pleiotropic variants (i.e. with a direct effect on Y not through D). As a result of this corollary, we can combine the p-values using standard procedures to test the intersection or global null hypothesis (Bretz, Hothorn, and Westfall 2016).

One such procedure is called Fisher's method (Fisher 1925; Wang and Owen 2019). If $\{p_1, p_2, \ldots, p_k\}$ are a collection of independent p-values then, when all of the corresponding null hypotheses are true (or a single shared null hypothesis is true),

$$-2\sum_{j=1}^{k}\log(p_j) \sim \mathcal{X}^2_{2k},$$

where $\mathcal{X}^2_{2k}$ denotes the chi-squared distribution with $2k$ degrees of freedom. We use Fisher's method to aggregate our independent p-values in the applied example in Section 5.

As some instruments may violate the exclusion restriction, a more robust approach is to test the partial conjunction of the null hypotheses (Wang and Owen 2019). In practice, it may not be possible to separate closely linked instruments into partitions separated by a heterozygous variant, in which case the hypothesis (8) can be tested using $(Z^m_j, Z^m_{j'})$ jointly. Corollary 3 in Appendix C derives the joint randomization distribution of a collection of instruments.

4 Simulation

4.1 Setup

In this section we explore the properties of our almost exact test via simulation. The set up of the simulation is described in detail in Appendix D. To summarize, we consider a null effect of an exposure on an outcome (i.e. $\beta = 0$), both of which have variance one, using 5 genetic instruments on different chromosomes. The instruments are non-causal markers for nearby causal variants and there are also pleiotropic variants in linkage disequilibrium with the instruments. From the above setup we simulate a sample of 15,000 parent-offspring trios.

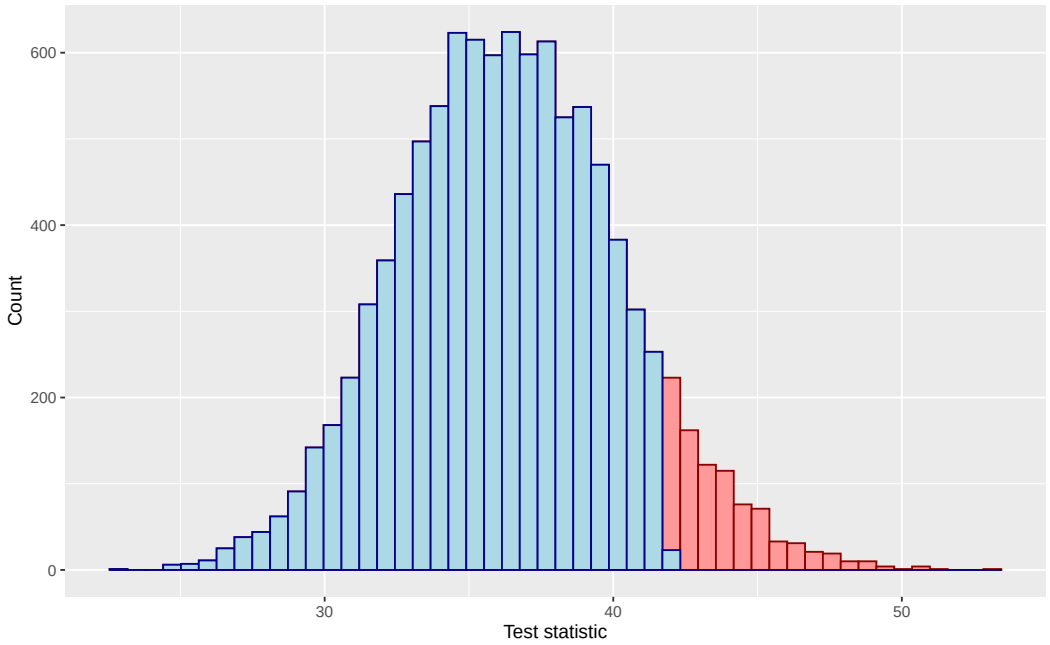
To make our setup more tangible, Table 2 shows the first 6 lines of observed and counterfactual data (in red) from the simulation for one of the instruments and corresponding parental haplotypes. We can see that individual 4 will provide almost no information for a test of the null hypothesis; both of 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 2 shows a counterfactual draw of each individual's instrument conditional on the adjustment set given in Equation (22) in Appendix D, 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 now inherits both minor alleles. Figure 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 2: First 6 lines 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 4: Histogram of 10,000 test statistics under the exact null hypothesis $H_0 : \beta = -0.3$



4.2 Power

In this section we simulate the power of our almost exact test using a correct adjustment set (see Equation (22) in Appendix D). 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 under the correct $\beta = 0$ in the left panels of Figure 5.

The histograms on the right side of Figure 5 depict the distribution of p-values under an alternative hypothesis $H_1 : \beta = 0.5$. The power to reject this hypothesis varies significantly across 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 6 expands upon the previous figure by plotting a power curve for test statistic 1 and 2.

We can see that test statistic 1 has power almost equal to 0 between $\beta_0 = 0$ and $\beta_0 = 1$. This occurs because the simple two-stage least squares estimator unconditional on the adjustment set is upward biased, with an Anderson-Rubin 95% confidence interval of 0.64–0.89. We have minimal power to reject null hypotheses in that region unless we condition on the confounders in the test statistic, because the resampled instruments retain their correlation with the confounders.

Test statistic 2, which conditions on the confounders via a clever covariate, 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 is interesting to note that using test statistic 2 is not always more powerful than test statistic 1.

5 Applied example

5.1 Preliminaries

In this section we illustrate our approach using a negative control and a positive control. The negative control is the effect of child's BMI at age 7 on mother's BMI pre-pregnancy. Dynastic effects could induce a spurious correlation between child's BMI-associated variants and their mother's BMI pre-pregnancy. This opens the backdoor path seen in Figure 3e. Closing this backdoor path is crucial for reliable causal inference. The positive control is the effect of child's BMI at age 7 on itself, plus some mean-zero noise. We vary the proportion of the outcome that is attributable to noise to assess the power of our test.

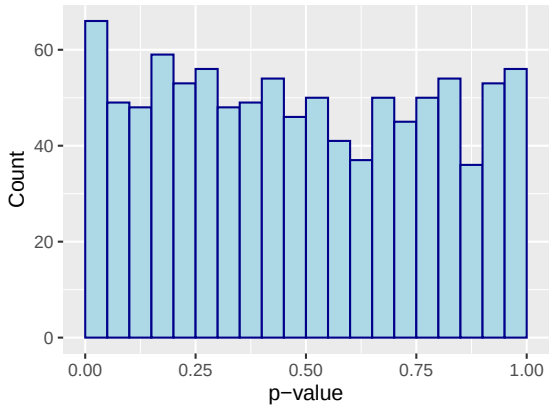
Our data consist of 6,222 mother-child duos from the Avon Longitudinal Study of Parents and Children (ALSPAC). ALSPAC is a longitudinal cohort initially comprising pregnant women resident in Avon, UK with expected dates of delivery from 1 April 1991 to 31 December 1992. The initial sample consisted of 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. 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/`).

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. Around each instrument, we condition on all variants which are more than 500 kilobases away.

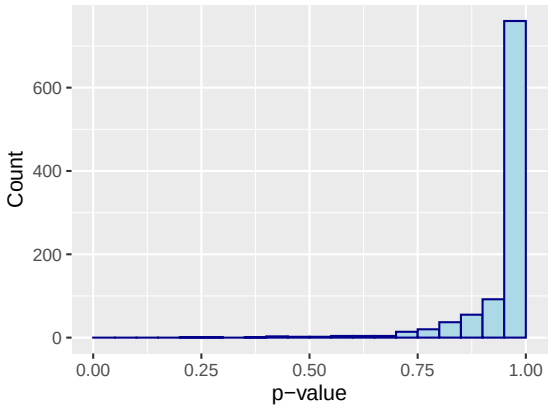
5.2 Data processing

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

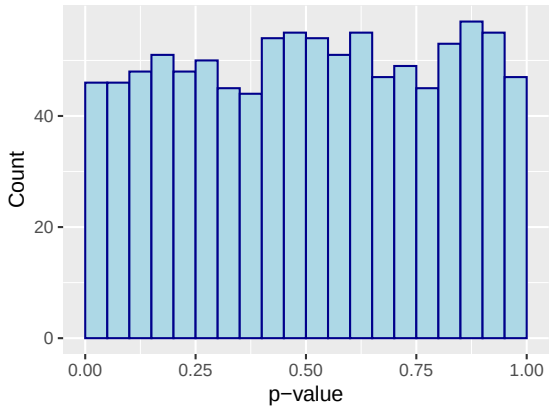
Figure 5: Histograms of 1,000 p-values for several null hypotheses and test statistics. Test statistic 1 is the F-statistic from a linear regression of the adjusted outcome on the instruments. Test statistic 2 is similar but includes the propensity scores for each instrument as covariates. Test statistic 3 includes only the parental genotypes for each instrument as covariates.



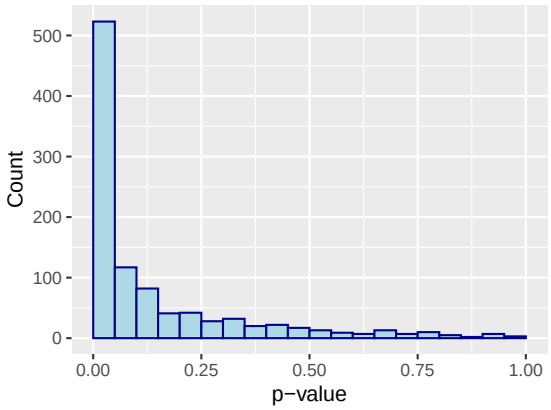
(a) $H_0 : \beta = 0$ and test statistic 1



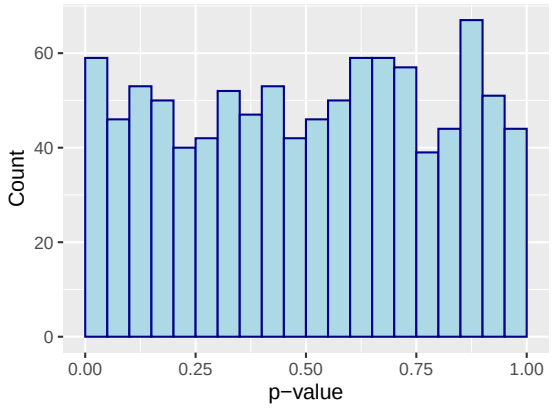
(b) $H_0 : \beta = 0.5$ and test statistic 1



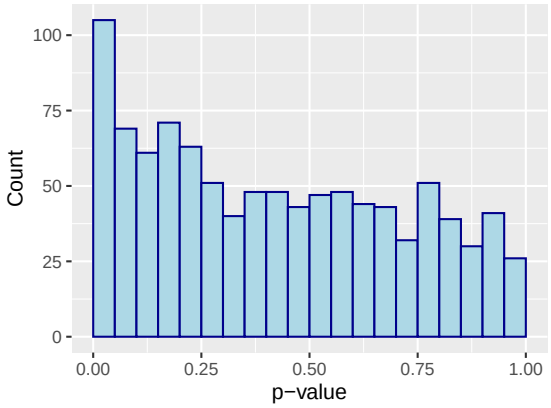
(c) $H_0 : \beta = 0$ and test statistic 2



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

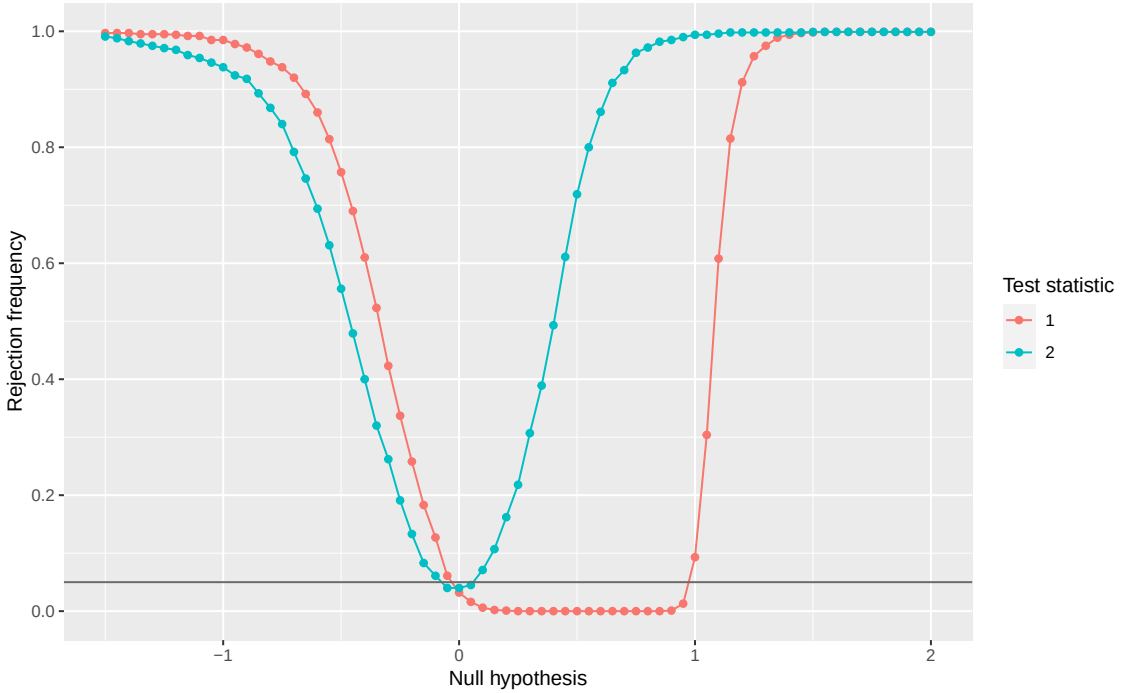


(e) $H_0 : \beta = 0$ and test statistic 3



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

Figure 6: Power curves for two choices of test statistic. Test statistic 1 is the F-statistic from a naive regression of the adjusted outcome on the instruments. Test statistic 2 is similar but includes the propensity scores for each instrument as covariates. Each point on the figure is the rejection frequency over 1,000 replications.



5.3 Results

Table 3 shows the negative control results and Table 4 shows the positive control results across all instruments. 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 against the null. The aggregated p-values for the positive control range from 0.03 (when 10% of the simulated outcome is noise) to 0.16 (when 50% of the simulated outcome is noise). This is weak evidence against the null, resulting from insufficiently strong instruments.

We can also compare the results in Tables 3 and 4 with a typical two-stage least squares (2SLS) regression 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 simulated outcome is noise) to 4.5×10^{-11} (when 50% 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 6

Table 3: 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 4: Results from the ALSPAC positive control example

Instrument (rsID)	Chromosome	Proximal 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

6 Discussion

Our test represents an almost exact approach to within-family MR, however, Section 5 demonstrates that power may be limited relative to typical Mendelian randomization analyses in unrelated individuals. Since our test can leverage the precise amount of power available in a single meiosis, this suggests that Mendelian randomization in unrelated individuals is drawing power from elsewhere, most likely many meioses across multiple 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 MR in unrelated individuals must), 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 valuable for the MR literature to discuss the extent to which Mendelian inheritance across multiple generations is driving the power behind existing results.

We must also return to the problem of transmission ratio distortion (TRD) discussed in Section 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 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.

7 Acknowledgements

The authors thank Kate Tilling, Rachael A Hughes, Jack Bowden, Gibran Hemani, Neil M Davies, Ben Brumpton, and Nianqiao Ju for their helpful feedback. In addition, the authors are extremely grateful to all the families who took part in the ALSPAC cohort, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. Ethical approval for our applied example was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The UK Medical Research Council and Wellcome (grant number 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. This publication is the work of the authors who will serve as guarantors for the contents of this paper. This research was supported in part by the Wellcome Trust (grant number 220067/Z/20/Z) and EPSRC (grant number EP/V049968/1). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

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. ISSN: 1474760X. DOI: 10.1186/s13059-016-1110-1.

Bates, Stephen et al. (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. ISSN: 10916490. DOI: 10.1073/pnas.2007743117. arXiv: 2002.09644.

Belmont, John W. et al. (2005). "A haplotype map of the human genome". In: *Nature* 437.7063, pp. 1299–1320. ISSN: 00280836. DOI: 10.1038/nature04226.

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. ISSN: 20411723. DOI: 10.1038/ncomms14994.

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. ISSN: 14643685. DOI: 10.1093/ije/dyv080.

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. DOI: 10.1093/ije/dys064.

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. DOI: 10.1016/j.cell.2017.05.038.An.

Bretz, Frank, Torsten Hothorn, and Peter Westfall (2016). *Multiple Comparisons Using R*. [] Chapman and Hall/CRC. DOI: 10.1201/9781420010909.

Broman, Karl W. and James L. Weber (2000). "Characterization of human crossover interference". In: *American Journal of Human Genetics* 66.6, pp. 1911–1926. ISSN: 00029297. DOI: 10.1086/302923.

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. ISSN: 20411723. DOI: 10.1038/s41467-020-17117-4.

Cardon, Lon R. and Lyle J. Palmer (2003). "Population stratification and spurious allelic association". In: *Lancet* 361.9357, pp. 598–604. ISSN: 01406736. DOI: 10.1016/S0140-6736(03)12520-2.

Davey Smith, George (2001). "Reflections on the limitations to epidemiology". In: *Journal of Clinical Epidemiology* 54.4, pp. 325–331. ISSN: 08954356. DOI: 10.1016/S0895-4356(00)00334-6.

— (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. ISSN: 03005771. DOI: 10.1093/ije/dyg070.

Davey Smith, George 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. ISSN: 1573-7284. DOI: 10.1007/s10654-020-00622-7.

Davies, Neil M. et al. (2019). "Within family Mendelian randomization studies". In: *Human Molecular Genetics* 28.R2, R170–R179. ISSN: 14602083. DOI: 10.1093/hmg/ddz204.

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. ISSN: 09622802. DOI: 10.1177/0962280206077743.

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. DOI: 10.1017/s0080456800012163.

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. ISSN: 03005771. DOI: 10.1093/ije/dyp379.

Fisher, Ronald Aylmer (1926). "The Arrangement of Field Experiments". In: *Journal of the Ministry of Agriculture* 33, pp. 503–513. DOI: 10.23637/ROTHAMSTED.8V61Q.

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. DOI: 10.1093/ije/dys066.

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.

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 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. ISSN: 10982272. DOI: 10.1002/gepi.22138.

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.

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. ISSN: 14602083. DOI: 10.1093/hmg/ddy163.

Hernán, Miguel A and James M Robins (2020). *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.

Holland, Paul W. (1986). "Statistics and causal inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960. ISSN: 1537274X. DOI: 10.1080/01621459.1986.10478354.

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. ISSN: 20411723. DOI: 10.1038/s41467-019-12424-x.

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. ISSN: 1061-4036. DOI: 10.1038/s41588-022-01062-7.

Imbens, Guido W and Donald B Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9781139025751.

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. ISSN: 1467985X. DOI: 10.1111/rssa.12353. arXiv: 1606.04146.

Kang, Hyunseung, Anru Zhang, et al. (2016). "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. DOI: 10.1080/01621459.2014.994705.

Katan, Martjin B. (1986). "Apolipoprotein E isoforms, serum cholesterol, and cancer". In: *International Journal of Epidemiology* 33.1, p. 9. ISSN: 03005771. DOI: 10.1093/ije/dyh312.

Kolesár, Michal et al. (2015). "Identification and Inference With Many Invalid Instruments". In: *Journal of Business & Economic Statistics* 33.4, pp. 474–484. DOI: 10.1080/07350015.2014.978175.

Kong, Augustine et al. (2018). "The nature of nurture: Effects of parental genotypes". In: *Science* 359, pp. 424–428. DOI: 10.1101/219261.

Lander, Eric S and Nicholas J Schork (1994). "Genetic dissection of complex traits". In: *Science* 265, pp. 2037–2048. ISSN: 15461718. DOI: 10.1038/ng0496-355.

Lauritzen, Steffen L, Phillip A Dawid, et al. (1990). "Independence properties of directed markov fields". In: *Networks* 20.5, pp. 491–505. DOI: 10.1002/net.3230200503.

Lauritzen, Steffen L. and Nuala A. Sheehan (2003). "Graphical models for genetic analyses". In: *Statistical Science* 18.4, pp. 489–514. ISSN: 08834237. DOI: 10.1214/ss/1081443232.

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. DOI: 10.1289/ehp.792971.

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. ISSN: 1474547X. DOI: 10.1016/S0140-6736(18)31772-0.

Morton, Newton E (1955). "Sequential tests for the detection of linkage". In: *American Journal of Human Genetics* 7.3, pp. 277–318.

Nadeau, Joseph H. (2017). "Do gametes woo? Evidence for their nonrandom union at fertilization". In: *Genetics* 207.2, pp. 369–387. ISSN: 19432631. DOI: 10.1534/genetics.117.300109.

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. ISSN: 15452948. DOI: 10.1146/annurev-genet-040119-093957.

Patterson, Nick, Alkes L. Price, and David Reich (2006). "Population Structure and Eigenanalysis". In: *PLoS Genetics* 2.12, e190. DOI: 10.1371/journal.pgen.0020190.

Pearl, Judea (2009). *Causality*. 2nd ed. New York: Cambridge University Press, p. 478.

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. DOI: 10.2307/2984124.

Richardson, Tom S and James M Robins (2013). "Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality".

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 R (2004). *Randomization inference with an instrumental variable*.

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. DOI: 10.1017/CBO9780511810725.016.

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. ISSN: 10970258. DOI: 10.1002/sim.7901.

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. ISSN: 1537274X. DOI: 10.1080/01621459.1980.10477512.

Sanderson, Eleanor et al. (2022). "Mendelian randomization". In: *Nature Reviews Methods Primers* 2.6. DOI: 10.1038/s43586-021-00092-5.

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.

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. ISSN: 00029297.

Taubes, Gary (1995). "Epidemiology faces its limits". In: *Science* 269, pp. 164–169.

Thomas, Duncan C. and David V. Conti (2004). "Commentary: The concept of 'Mendelian randomization'". In: *International Journal of Epidemiology* 33.1, pp. 21–25. ISSN: 03005771. DOI: 10.1093/ije/dyh048.

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. ISBN: 0940600498.

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. DOI: 10.1371/journal.pgen.1008718.

Wang, Jingshu and Art B. Owen (2019). "Admissibility in Partial Conjunction Testing". In: *Journal of the American Statistical Association* 114.525, pp. 158–168. ISSN: 1537274X. DOI: 10.1080/01621459.2017.1385465. arXiv: 1508.00934.

Watson, David S. and Marvin N. Wright (2019). "Testing conditional independence in supervised learning algorithms". In: *arXiv*. arXiv: 1901.09917.

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. ISSN: 03005771. DOI: 10.1093/ije/dyg313.

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.

Zhao, Qingyuan 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. DOI: 10.1214/19-AOS1866.

A Introduction to causal inference

This section introduces some basic concepts in causal inference, including the potential outcomes framework (Neyman 1990; Rubin 1974), randomization inference (Fisher 1935; Rubin 1980), instrumental variables (Wright 1923; Imbens and Rubin 2015), and single world intervention graphs (Richardson and Robins 2013), which are essential for describing our methodology. Interested readers can find a more thorough coverage of these topics in other texts (Imbens and Rubin 2015; Hernán and Robins 2020).

A.1 Treatment assignment and potential outcomes

In the typical setup of a randomized experiment with non-compliance, we have a sample of N individuals indexed by $i = 1, 2, \ldots, N$ and each individual is randomly assigned to receive a binary treatment $Z_i \in \{0, 1\}$. The common convention is that $Z_i = 1$ denotes assignment to an experimental treatment and $Z_i = 0$ a control treatment. However, individuals might not comply with their assigned treatment, and we denote the treatment that the individual actually takes as $D_i \in \{0, 1\}$. Finally, we observe an outcome variable Y_i for each individual. As the treatment uptake D_i is not randomized, there may exist a confounding variable C_i that is a common cause of both D_i and Y_i.

Individual i has two *potential outcomes* (also called *counterfactuals*) of her treatment uptake, $D_i(0)$ and $D_i(1)$. If she is randomized to the experimental (or control) treatment, she will take treatment $D_i(1)$ (or $D_i(0)$). For example, some individuals will take the experimental treatment regardless of their assigned treatment, so $D_i(1) = D_i(0) = 1$. Each individual also has four potential outcomes $Y_i(z, d)$ from the experiment corresponding to each combination of treatment assignment $z \in \{0, 1\}$ and treatment uptake $d \in \{0, 1\}$. Similarly, we may define the potential outcome with

Table 5: Observed data from a hypothetical experiment for a LDL cholesterol-lowering drug.

i	Z_i	D_i	$D_i(1)$	$D_i(0)$	Y_i	$Y_i(1)$	$Y_i(0)$
1	1	1	1	?	120	120	?
2	1	1	1	?	120	120	?
3	1	0	0	?	75	?	75
4	0	0	?	1	165	165	?
5	0	0	?	0	135	?	135
6	0	0	?	0	105	?	105

just D being intervened on by $Y_i(d) = Y_i(Z_i, d)$, where the assigned treatment takes its "natural" value Z_i.

In defining these potential outcomes, we have implicitly made the *no interference* assumption which states that individual j's treatment is independent of individual i's outcome when $i \neq j$ (Rubin 1980). To simplify the exposition, we further make the *exclusion restriction* assumption in this section. That is, we assume that the treatment assignment has no causal effect on the outcome except via treatment uptake, so

$$Y_i(1, d) = Y_i(0, d) = Y_i(d) \text{ for } d \in \{0, 1\}. \tag{10}$$

Let $\mathcal{F} = \{(D_i(1), D_i(0), Y_i(0), Y_i(1)),\ i = 1, \ldots, N\}$ denote the collection of potential outcomes for all the individuals.

We define a *causal effect* of the treatment as a contrast of potential outcomes. When the treatment is binary, the causal effect for individual i is given by $\beta_i = Y_i(1) - Y_i(0)$, the difference in individual i's outcomes between the two possible treatments. However, inference for the individual treatment effect β_i is difficult because we do not observe both potential outcomes of the same individual simultaneously. This has been famously described as the "fundamental problem of causal inference" (Holland 1986). Indeed, we only observe the potential outcome corresponding to the treatment that is actually received, such that

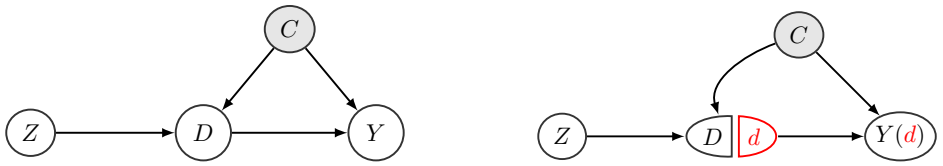
$$D_i = \begin{cases} D_i(1) & \text{if } Z_i = 1, \\ D_i(0) & \text{if } Z_i = 0; \end{cases} \text{ and } Y_i = \begin{cases} Y_i(1) & \text{if } D_i = 1, \\ Y_i(0) & \text{if } D_i = 0. \end{cases} \tag{11}$$

The above equation is sometimes called the *consistency* assumption since it ensures that the observed outcomes and potential outcomes are consistent with one another (Hernán and Robins 2020).

From this perspective, causal inference can be regarded as a missing data problem. Consider a simple hypothetical experiment in Table 5 consisting of $N = 6$ individuals, 3 of whom are randomized to take an experimental LDL cholesterol-lowering drug and 3 of whom are randomized to take a placebo. However, not everyone adheres to the assigned treatment. The outcome variable is LDL cholesterol measured in grams per litre (mg/dL). As discussed above, we can only observe the potential outcomes corresponding to the observed treatment assignment and uptake; all other potential outcomes are missing.

A.2 Causal graphical models

The setting above can be described by a directed acyclic graph (DAG) as shown in Figure 7a. Below we will use some basic concepts in DAG models such as Markov properties and d-separation, which



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

Figure 7: Causal diagram of the example in Appendix A

are described in detail in other texts (Lauritzen, Dawid, et al. 1990; Pearl 2009; Hernán and Robins 2020).

We will use single world intervention graphs (SWIG) to unify the counterfactual and graphical descriptions of the causal inference problem (Richardson and Robins 2013). The SWIG representation of our setting above is given in Figure 7b. Since we are interested in the causal effect of an intervention on D, the SWIG splits that node into two halves: D, representing the randomized treatment, 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.

It has been shown that SWIGs define a graphical model for the potential outcomes (Richardson and Robins 2013), so we can apply d-separation to obtain conditional independence between counterfactuals. For example, Figure 7b implies *exchangeability* (or *ignorability*),

$$Z_i \perp\!\!\!\perp Y_i(d) \text{ for all } d \in \{0, 1\}. \tag{12}$$

However, $D_i \perp\!\!\!\perp Y_i(d)$ is generally not true due to the confounder C_i.

A.3 Randomization inference for instrumental variables

To construct an exact randomization test, the key idea is to base the inference precisely on the randomness introduced by the experimenter. To this end, we must characterize the treatment assignment mechanism.

Let $\boldsymbol{Z} = (Z_1, \ldots, Z_N)^\intercal$ denote the N-vector of treatment assignments. To simplify the exposition, we will assume that the experiment is completely randomized, such that a fixed number of individuals N_t are assigned to the experimental treatment and $N_c = N - N_t$ are assigned to the control treatment. The same method below can be applied to more sophisticated assignment mechanisms (such as the ones we describe later for within-family MR). Let $\Omega = \{(z_1, \ldots, z_N) \in \{0, 1\}^N : \sum_{i=1}^N z_i = N_t\}$ 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

$$\mathbb{P}(\boldsymbol{Z} = \boldsymbol{z} \mid \mathcal{F}) = \begin{cases} \binom{N}{N_t}^{-1}, & \text{for all } \boldsymbol{z} \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

To illustrate randomization inference, consider the hypothetical experiment in Table 5. Suppose we are interested in evaluating the effectiveness of this drug at lowering LDL cholesterol. However, although the drug is initially randomly assigned, the treatment uptake is not randomized. In particular, non-compliance may be driven by a confounder $C_i \in \{0, 1\}$. This might be an underlying comorbidity such that those with $C_i = 1$ have a higher baseline outcome $Y_i(0)$ but experience negative side effects from the experimental treatment. Due to the side effects, individuals with the

comorbidity (such as $i = 3$ in Table 5) may be inclined to switch to the control treatment when they are assigned to the experimental drug. Due to the systematic shift of high baseline individuals from the experimental treatment to the control treatment, a simple intention-to-treat estimate (by regressing Y_i on Z_i) will underestimate the causal effect.

To address unobserved confounding such as systematic non-compliance, one approach is to use an instrumental variable. An instrumental variable induces unconfounded variation in the treatment without otherwise affecting the outcome. In our example, the randomized treatment assignment $Z_i \in \{0, 1\}$ is a good instrument for treatment uptake D_i, because it will change the outcome Y_i only through D_i by the exclusion restriction assumption (10). Furthermore, it is independent of the underlying comorbidity status and the counterfactual outcomes, as shown by (12).

Randomization inference for instrumental variables (Rosenbaum 2004; Kang, Peck, and Keele 2018) tests sharp null hypotheses of the form

$$H_0 : Y_i(d) - Y_i(0) = \beta_0 d, \text{ for all } d \in \{0, 1\}.$$

This implies a constant additive treatment effect β_0 across individuals. Under this hypothesis and the consistency assumption (11), the baseline potential outcome can be written in terms of the observable data (Z_i, D_i, Y_i) as

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

which is termed as the "adjusted response" by Rosenbaum (2004). Therefore, when the null hypothesis is true, the randomization of Z_i, namely (12), implies that

$$Z_i \perp\!\!\!\perp Y_i - \beta_0 D_i.$$

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

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

The randomization test then rejects H_0 at significance level α, if the p-value

$$P(\boldsymbol{Z} \mid \mathcal{F}) = \tilde{\mathbb{P}}(T(\tilde{\boldsymbol{Z}} \mid \mathcal{F}) \leq T(\boldsymbol{Z} \mid \mathcal{F}))$$

is less than or equal to α. Here $\tilde{\boldsymbol{Z}}$ is an independent copy of $\boldsymbol{Z}$ and $\tilde{\mathbb{P}}$ means that the probability is taken over $\tilde{\boldsymbol{Z}}$ according to the randomization distribution (13). In plain terms, we are asking: if we re-ran the experiment many times under the null hypothesis (i.e. Z_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 α, then we have little confidence in the null hypothesis.

This p-value has size α in the sense that

$$\mathbb{P}(P(\boldsymbol{Z} \mid \mathcal{F}) \leq \alpha \mid H_0) = \alpha.$$

for any significance level $0 \leq \alpha \leq 1$ and test statistic $T(\cdot \mid \mathcal{F})$. For continuously distributed test statistics the proof relies on the idea that $T(\tilde{\boldsymbol{Z}} \mid \mathcal{F}) \overset{d}{=} T(\boldsymbol{Z} \mid \mathcal{F})$ under H_0 which means that

$P(\boldsymbol{Z} \mid \mathcal{F})$ is the cumulative distribution of $T(\tilde{\boldsymbol{Z}} \mid \mathcal{F})$ at $T(\boldsymbol{Z} \mid \mathcal{F})$. Since cumulative distributions are uniformly distributed the result follows.

Next, we illustrate the randomization test using the hypothetical experiment in Table 5 and the null hypothesis $H_0 : Y_i(0) = Y_i(1)$ for all i (i.e. $\beta_0 = 0$). For the realized experiment in Table 5, the difference in outcomes between the experimental and placebo groups is $(120 + 120)/2 - (75 + 165 + 135 + 105)/4 = 0$. In other words, average LDL cholesterol appears to be identical in the experimental and control arms. However, it is unclear whether this should be interpreted as evidence of a null causal effect. As discussed above, it is possible that individuals with high baseline outcomes are more inclined to switch from the experimental treatment to the control treatment.

Our observed test statistic for this experiment is $T(\boldsymbol{Z} \mid \mathcal{F}) = (120 + 120 + 75)/3 - (165 + 135 + 105)/3 = -30$. Since we know the missing potential outcomes under the null hypothesis and we know the mechanism by which treatment was randomly assigned, we can also consider the results of counterfactual experiments. Table 6 shows a counterfactual experiment that could have occurred with missing potential outcomes imputed under the null. The counterfactual treatment assignment is given by $\tilde{Z}_i$ to distinguish it from the factual Z_i. We can compute the difference in outcomes from this counterfactual experiment, equal to $T(\tilde{\boldsymbol{Z}} \mid \mathcal{F}) = 420/3 - 300/3 = 40$. Indeed, we could enumerate the counterfactual results from all 20 equally possible experiments, shown in Figure 8. The bars highlighted in red are comprised of 4 counterfactual experiments with an average outcome difference less than or equal to that observed in our actual experiment. Therefore, under the null hypothesis, the one-sided probability of observing a result more extreme than our observed result is 4/20 or 20%.

Table 6: Imputed data from a counterfactual experiment under the exact null hypothesis

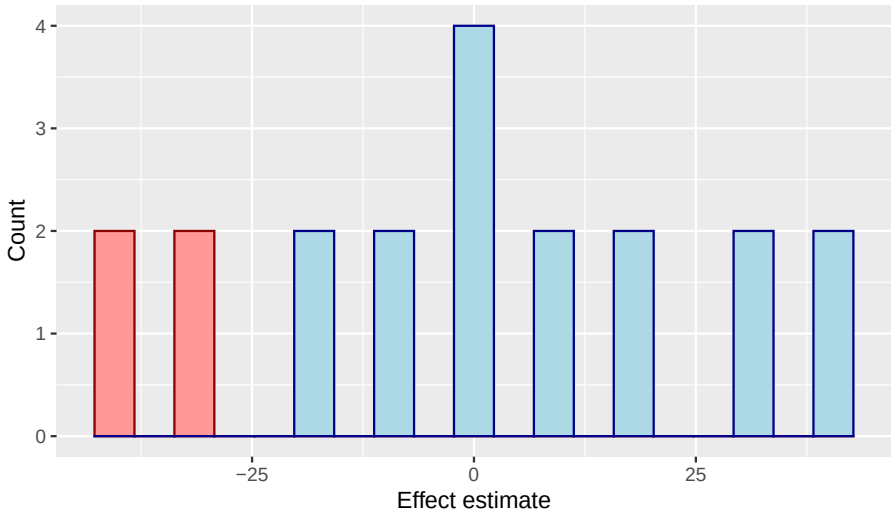
i	$\tilde{Z}_i$	D_i	$D_i(1)$	$D_i(0)$	Y_i	$Y_i(1)$	$Y_i(0)$
1	1	1	1	?	120	120	120
2	0	1	1	?	120	120	120
3	0	0	0	?	75	75	75
4	1	0	?	1	165	165	165
5	1	0	?	0	135	135	135
6	0	0	?	0	105	105	105

B 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 et al. 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

Figure 8: Histogram of outcome differences for the exact null hypothesis



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

$$\mathbb{P}(Z_j^m = M_j^{(u_j^m)} \mid U_j^m = u_j^m) = 1 - \epsilon \tag{14}$$

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 9. This graph differs from the more general structure shown in Figure 3b in that each meiosis indicator U_j^m depends only on the previous indicator U_{j-1}^m. Figure 10 embeds the hidden Markov model within the complete causal model used throughout Section 3.2.

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 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

$$\mathbb{P}(Z_j^m = 1 \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) = \sum_{u \in \{m,f\}} \mathbb{P}(Z_j^m = 1 \mid U_j^m = u) \mathbb{P}(U_j^m = u \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m). \tag{15}$$

It is therefore more convenient to consider the conditional probability of U_j^m. We state the following theorem:

Theorem 3. *Using the conditional independence properties implied by Figure 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

$$\begin{aligned} \beta_p^m(u_p^m) &= 1 \\ \beta_k^m(u_k^m) &= \sum_{u \in \{m,f\}} \beta_{k+1}^m(u) \mathbb{P}(U_{k+1}^m = u \mid U_k^m = u_k^m) \mathbb{P}(Z_{k+1}^m = z_{k+1}^m \mid U_{k+1}^m = u), \;\; k = 1, \dots, p-1, \end{aligned}$$

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 3 can be simplified further.

Corollary 2. *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 3 can equivalently be written as*

$$\begin{aligned} &\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \\ \propto \; &\Bigg[\sum_{u \in \{m,f\}} \tilde{\beta}_{h-1}^m(u) \mathbb{P}(U_{h-1}^m = u \mid U_j^m = m) \Bigg] \Bigg[\sum_{u \in \{m,f\}} \mathbb{P}(U_j^m = m \mid U_l^m = u)\, \tilde{\alpha}_l^m(u) \Bigg]. \end{aligned}$$

where

$$\begin{aligned} \tilde{\alpha}_{b_1+1}^m(u_{b_1+1}^m) &= \mathbb{P}(Z_{b_1+1}^m = z_{b_1+1}^m \mid U_{b_1+1}^m = u_{b_1+1}^m) \mathbb{P}(U_{b_1+1}^m = u_{b_1+1}^m \mid U_{b_1}^m = u_{b_1}^m) \\ \tilde{\alpha}_k^m(u_k^m) &= \sum_{u \in \{m,f\}} \mathbb{P}(Z_k^m = z_k^m \mid U_k^m = u_k^m) \mathbb{P}(U_k^m = u_k^m \mid U_{k-1}^m = u) \tilde{\alpha}_{k-1}^m(u), \;\; k = b_1 + 2, \dots, p; \end{aligned}$$

and

$$\begin{aligned}\tilde{\beta}^m_{b_2-1}(u^m_{b_2-1}) &= \mathbb{P}(U^m_{b_2} = u^m_{b_2} \mid U^m_{b_2-1} = u^m_{b_2-1})\mathbb{P}(Z^m_{b_2} = z^m_{b_2} \mid U^m_{b_2} = u^m_{b_2}) \\ \beta^m_k(u^m_k) &= \sum_{u\in\{m,f\}} \tilde{\beta}^m_{k+1}(u)\mathbb{P}(U^m_{k+1} = u \mid U^m_k = u^m_k)\mathbb{P}(Z^m_{k+1} = z^m_{k+1} \mid U^m_{k+1} = u),\ k = 1,\dots,b_2-2.\end{aligned}$$

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 3.

Corollary 3. *Suppose we have a collection of instruments $\mathcal{J} = \{j_1, j_2, \dots, j_r\}$ such that $l < j_1 < j_2 < \dots < j_r < h$. Then the propensity score can be written as*

$$\mathbb{P}(U^m_{j_1} = u^m_{j-1}, U^m_{j_2} = u^m_{j_2}, \dots, U^m_{j_r} = u^m_{j_r} \mid \boldsymbol{Z}^m_{\mathcal{B}} = \boldsymbol{z}^m_{\mathcal{B}}) \tag{16}$$

$$= \mathbb{P}(U^m_{j_1} = u^m_{j_1} \mid \boldsymbol{Z}^m_{\mathcal{B}} = \boldsymbol{z}^m_{\mathcal{B}}) \prod_{k=2}^{r} \mathbb{P}(U^m_{j_k} = u^m_{j_k} \mid U^m_{j_{k-1}} = u^m_{j_{k-1}}, \boldsymbol{Z}^m_{\mathcal{B}} = \boldsymbol{z}^m_{\mathcal{B}}) \tag{17}$$

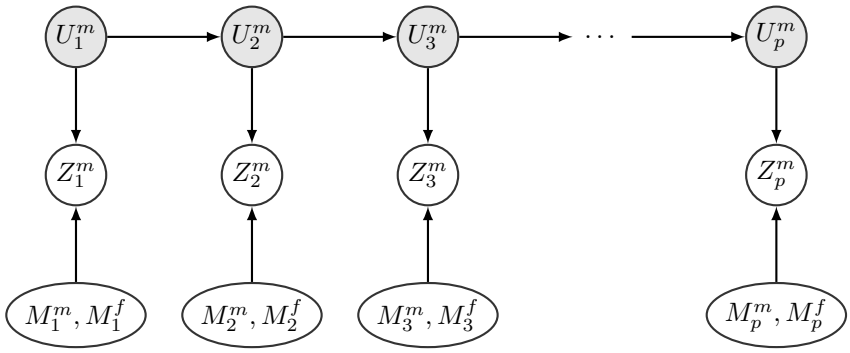
The first propensity score $\mathbb{P}(U^m_{j_1} = m \mid \boldsymbol{Z}^m_{\mathcal{B}} = \boldsymbol{z}^m_{\mathcal{B}})$ takes the form in Theorem 3 and

$$\mathbb{P}(U^m_{j_k} = m \mid U^m_{j_{k-1}} = u^m_{j_{k-1}}, \boldsymbol{Z}^m_{\mathcal{B}} = \boldsymbol{z}^m_{\mathcal{B}}) \tag{18}$$

$$\propto \mathbb{P}(U^m_{j_k} = m \mid U^m_{j_{k-1}} = u^m_{j_{k-1}}) \Big[\sum_{u\in\{m,f\}} \beta^m_{h-1}(u)\mathbb{P}(U^m_{h-1} = u \mid U^m_{j_k} = m) \Big] \tag{19}$$

where $\beta^m_{h-1}(u)$ is the backward weight defined in Theorem 3.

Figure 9: Graphical representation of Haldane's hidden Markov model

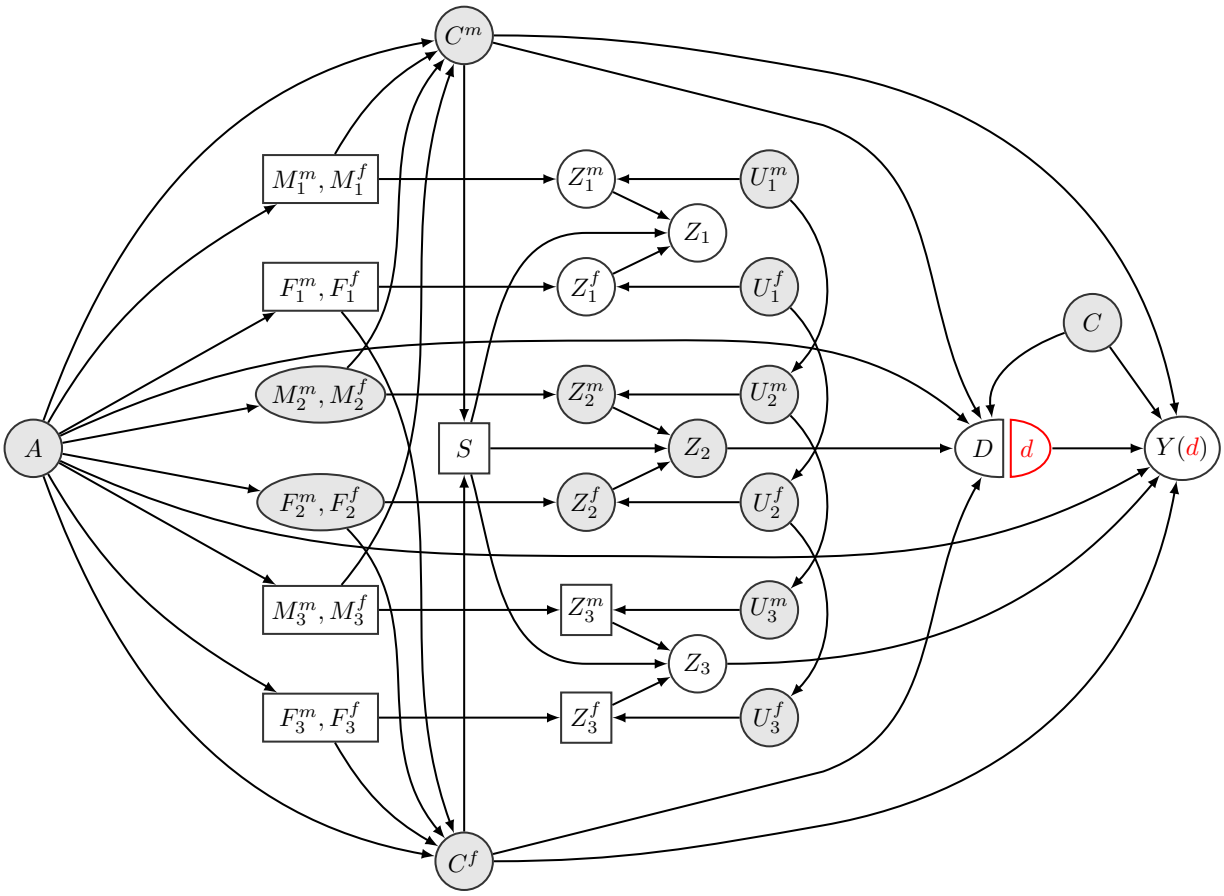


C Technical proofs

Proposition 1

Proof. From Assumption 1 we know that, conditional on $(M^m_j, M^f_j, F^m_j, F^f_j)$, Z^m_j and Z^f_j only depend on $\boldsymbol{U}^m$ and $\boldsymbol{U}^f$, respectively, and exogenous mutation events. By (1), $Z_j = Z^m_j + Z^f_j$ given that $S = 1$ (fertilization occurs). Finally, by Assumption 4, 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. □

Figure 10: Haldane's hidden Markov model embedded in our full causal model



Theorem 3

Proof. The conditional probability of U_j^m can be factorized as

$$
\begin{aligned}
&\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \\
&\propto \mathbb{P}(U_j^m = m, \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) \\
&= \mathbb{P}(\boldsymbol{Z}_{h:p}^m = \boldsymbol{z}_{h:p}^m \mid U_j^m = m)\mathbb{P}(U_j^m = m, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m) \\
&= \left[\sum_{u\in\{m,f\}} \mathbb{P}(\boldsymbol{Z}_{(j+1):p}^m = \boldsymbol{z}_{(j+1):p}^m, U_{h-1}^m = u \mid U_j^m = m)\right]\left[\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m, U_l^m = u, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m)\right] \\
&= \left[\sum_{u\in\{m,f\}} \mathbb{P}(\boldsymbol{Z}_{(j+1):p}^m = \boldsymbol{z}_{(j+1):p}^m \mid U_{h-1}^m = u)\mathbb{P}(U_{h-1}^m = u \mid U_j^m = m)\right] \\
&\quad \left[\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m \mid U_l^m = u)\mathbb{P}(U_l^m = u, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m)\right] \\
&= \left[\sum_{u\in\{m,f\}} \beta_{h-1}^m(u)\mathbb{P}(U_{h-1}^m = u \mid U_j^m = m)\right]\left[\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m \mid U_l^m = u)\alpha_l^m(u)\right].
\end{aligned}
$$

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 2

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 3 we begin by factorising the conditional probability of U_j^m.

$$\mathbb{P}(U_j^m = m \mid \boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m) = \frac{\mathbb{P}(\boldsymbol{Z}_{h:p}^m = \boldsymbol{z}_{h:p}^m \mid U_j^m = m)\mathbb{P}(U_j^m = m, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m)}{\mathbb{P}(\boldsymbol{Z}_{\mathcal{B}}^m = \boldsymbol{z}_{\mathcal{B}}^m)}. \tag{20}$$

Since $b_1 < j$ we are concerned with simplifying the second probability in the numerator of equation (20).

$$\begin{aligned}&\mathbb{P}(U_j^m = m, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m) \\ =\ &\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m, U_{b_1}^m = u, \boldsymbol{Z}_{1:l}^m = \boldsymbol{z}_{1:l}^m) \\ =\ &\sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m, \boldsymbol{Z}_{(b_1+1):l}^m = \boldsymbol{z}_{(b_1+1):l}^m \mid U_{b_1}^m = u)\mathbb{P}(U_{b_1}^m = u, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \\ =\ &\mathbb{P}(U_j^m = m, \boldsymbol{Z}_{(b_1+1):l}^m = \boldsymbol{z}_{(b_1+1):l}^m \mid U_{b_1}^m = m)\mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \\ =\ &\mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m \mid U_{j-1}^m = u)\mathbb{P}(U_{j-1}^m = u, \boldsymbol{Z}_{(b_1+1):(j-1)}^m = \\ &\qquad \boldsymbol{z}_{(b_1+1):(j-1)}^m \mid U_{b_1}^m = m) \\ =\ &\mathbb{P}(U_{b_1}^m = m, \boldsymbol{Z}_{1:b_1}^m = \boldsymbol{z}_{1:b_1}^m) \sum_{u\in\{m,f\}} \mathbb{P}(U_j^m = m \mid U_{j-1}^m = u)\tilde{\alpha}_{j-1}^m(u).\end{aligned}$$

where

$$\begin{aligned}\tilde{\alpha}^m_{b_1+1}(u^m_{b_1+1}) &= \mathbb{P}(Z^m_{b_1+1} = z^m_{b_1+1} \mid U^m_{b_1+1} = u^m_{b_1+1})\mathbb{P}(U^m_{b_1+1} = u^m_{b_1+1} \mid U^m_{b_1} = m)\\ \tilde{\alpha}^m_k(u^m_k) &= \sum_{u\in\{m,f\}} \mathbb{P}(Z^m_k = z^m_k \mid U^m_k = u^m_k)\mathbb{P}(U^m_k = u^m_k \mid U^m_{k-1} = u)\tilde{\alpha}^m_{k-1}(u),\\ &\quad \text{for } k = b_1+2,\ldots,j-1.\end{aligned}$$

We now factorize the denominator of equation (20).

$$\mathbb{P}(\boldsymbol{Z}^m_{\mathcal{B}} = \boldsymbol{z}^m_{\mathcal{B}}) = \mathbb{P}(\boldsymbol{Z}_{(b_1+1):l} = \boldsymbol{z}_{(b_1+1):l}, \boldsymbol{Z}_{h:p} = \boldsymbol{z}_{h:p} \mid U^m_{b_1} = m)\mathbb{P}(U^m_{b_1} = m, \boldsymbol{Z}_{1:b_1} = \boldsymbol{z}_{1:b_1}).$$

Substituting these simplified expressions back in equation (20) we obtain

$$\begin{aligned}&\mathbb{P}(U^m_j = m \mid \boldsymbol{Z}^m_{\mathcal{B}} = \boldsymbol{z}^m_{\mathcal{B}})\\ =\ &\frac{\mathbb{P}(\boldsymbol{Z}^m_{h:p} = \boldsymbol{z}^m_{h:p} \mid U^m_j = m)\mathbb{P}(U^m_{b_1} = m, \boldsymbol{Z}^m_{1:b_1} = \boldsymbol{z}^m_{1:b_1})\sum_{u\in\{m,f\}}\mathbb{P}(U^m_j = m \mid U^m_{j-1} = u)\tilde{\alpha}^m_{j-1}(u)}{\mathbb{P}(\boldsymbol{Z}_{(b_1+1):l} = \boldsymbol{z}_{(b_1+1):l}, \boldsymbol{Z}_{h:p} = \boldsymbol{z}_{h:p} \mid U^m_k = m)\mathbb{P}(U^m_{b_1} = m, \boldsymbol{Z}_{1:b_1} = \boldsymbol{z}_{1:b_1})}\\ =\ &\frac{\mathbb{P}(\boldsymbol{Z}^m_{h:p} = \boldsymbol{z}^m_{h:p} \mid U^m_j = m)\sum_{u\in\{m,f\}}\mathbb{P}(U^m_j = m \mid U^m_{j-1} = u)\tilde{\alpha}^m_{j-1}(u)}{\mathbb{P}(\boldsymbol{Z}_{(b_1+1):l} = \boldsymbol{z}_{(b_1+1):l}, \boldsymbol{Z}_{h:p} = \boldsymbol{z}_{h:p} \mid U^m_{b_1} = m)}.\end{aligned} \tag{21}$$

which does not depend on $\boldsymbol{Z}^m_{1:k}$. □

D Simulation description

Table 7: 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 et al. (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 B. 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 1 implies that a sufficient adjustment set for this simulation is

$$(\boldsymbol{M}_{\mathcal{B}_g}^{mf}, \boldsymbol{F}_{\mathcal{B}_g}^{mf}, \boldsymbol{Z}_{\mathcal{B}}) \tag{22}$$

where

$$\mathcal{B} = \mathcal{J} \setminus \{24, 25, 26, 49, 50, 51, 99, 74, 75, 76, 99, 100, 101, 124, 125, 126\}$$

and

$$\mathcal{B}_g = \mathcal{B} \cup \{25, 50, 75, 100, 125\}.$$