Leveraging family data to design Mendelian Randomization that is provably robust to population stratification

Nathan LaPierre[1*†], Boyang Fu[1*], Steven Turnbull[2], Eleazar Eskin[1,3,4], and Sriram Sankararaman[1,3,4†]

1. Department of Computer Science; 2. Department of Statistics; 3. Department of Computational Medicine; and 4. Department of Human Genetics, UCLA, Los Angeles CA

* These authors contributed equally to this work.

† Email corresponding authors at nathanl2012@gmail.com and sriram@cs.ucla.edu

Abstract

Mendelian Randomization (MR) has emerged as a powerful approach to leverage genetic instruments to infer causality between pairs of traits in observational studies. However, the results of such studies are susceptible to biases due to weak instruments as well as the confounding effects of population stratification and horizontal pleiotropy. Here, we show that family data can be leveraged to design MR tests that are provably robust to confounding from population stratification, assortative mating, and dynastic effects. We demonstrate in simulations that our approach, MR-Twin, is robust to confounding from population stratification and is not affected by weak instrument bias, while standard MR methods yield inflated false positive rates. We then conduct an exploratory analysis of MR-Twin and other MR methods applied to 121 trait pairs in the UK Biobank dataset. Our results suggest that confounding from population stratification can lead to false positives for existing MR methods, while MR-Twin is immune to this type of confounding, and that MR-Twin can help assess whether traditional approaches may be inflated due to confounding from population stratification.

Introduction

Mendelian Randomization (MR) is a widely-used analytical tool that uses genetic variants ("genetic instruments") to determine whether one trait (the "exposure") has a causal effect on another (the "outcome"). With the availability of massive biobank datasets such as the UK Biobank (Bycroft et al., 2018), MR analyses have become increasingly powerful and have been used to identify causal relationships between numerous pairs of traits (Hemani et al., 2018; Wade et al., 2018; Lyall et al., 2017; Haycock et al., 2017; Haase et al., 2012). The validity of MR rests on three key assumptions (Lawlor et al., 2008): (i) that the genetic instrument is significantly associated with the exposure; (ii) that the genetic instrument is independent of confounders of the exposure-outcome relationship; (iii) that the genetic instrument affects the outcome only through its effect on the exposure.

Unfortunately, the latter two assumptions are often violated in practice, due to several factors including horizontal pleiotropy, population stratification (and related phenomena such as assortative mating and dynastic effects), and batch effects. Even when these assumptions are met, the weak effects of typical genetic instruments on the exposure coupled with spurious correlation between genetic instruments and confounders (Burgess et al., 2011) can bias the results of MR analyses ("weak instrument bias"). The problem of population stratification has been extensively studied in the Genome-Wide Association Study (GWAS) literature, and approaches for mitigating its effects have been developed, including the usage of Principal Components Analysis (PCA) and Linear Mixed Models (LMMs) (Price et al., 2010). These approaches have generally been found to be effective at reducing the confounding introduced by population stratification (Price et al., 2010).

However, recent studies have demonstrated that, with sample sizes as large as those found in modern biobanks, even a small amount of residual population stratification can cause a considerable amount of bias (Cook et al., 2020; Brumpton et al., 2020; Berg et al., 2019; Haworth et al., 2019), and may even cause false positives in MR analysis (Haworth et al., 2019; Cinelli et al., 2022). In addition, while the confounding effects of population stratification are well-known, less attention has been directed towards confounding from other phenomena such as (cross-trait) assortative mating and dynastic effects, which can also cause MR false positives (Brumpton et al., 2020; Hartwig et al., 2018). Recent work has demonstrated that cross-trait assortative mating is widespread and

substantially inflates genetic correlation estimates between many trait pairs (Border et al., 2022).

It has recently been proposed that family-based genetic datasets could be used in MR studies to avoid confounding from population stratification (Brumpton et al., 2020; Pingault et al., 2018). A recent suite of methods have been developed for this purpose, and were shown to reduce the bias from this type of confounding (Brumpton et al., 2020). However, like other MR methods, these methods are susceptible to weak instrument bias, which can be substantial for small family-based datasets (Brumpton et al., 2020). In this paper, we introduce MR-Twin, a test for causal effects between pairs of traits that is able to leverage family-based genetic data to provably control for population stratification and utilize publicly-available summary statistics estimated in large biobank datasets to achieve power competitive with top existing methods for the same sample size. We develop versions of MR-Twin for trio, parent-child duo, and sibling data, evaluate MR-Twin's ability to control false positives due to population stratification and weak instrument bias, and compare it with existing methods.

Results

Methods overview

In a Mendelian Randomization (MR) analysis, we wish to determine whether one phenotype (the "exposure") has a causal effect on another phenotype (the "outcome") using genetic instrumental variables, which can be either single nucleotide polymorphims (SNPs), a polygenic score, or other genetic features. Under the assumption that the genetic instruments are associated with the exposure and are independent of the outcome given the exposure, the MR effect estimate of the exposure on the outcome will be valid even if there are unobserved confounders of the exposure-outcome relationship. The independence assumption, however, is often violated by population stratification (Figure 1A) or horizontal pleiotropy, as these phenomena cause the genetic instruments to be correlated with the outcome through pathways other than those through the exposure.

MR-Twin is a method that uses family-based genetic data to construct a test for whether the exposure has an effect on the outcome that is immune to confounding from population structure. It is based on the key idea that the genotypes of observed individuals are independent of population structure given the genotypes of the individuals' parents, since the mechanisms by which

genetic information is passed from parents to offspring are known (Figure 1B). In other words, conditioned on the parental genotypes, population structure provides no additional information about the distribution of the offspring's genotypes. Thus, by conditioning on the parental genotypes, confounding from population stratification can be avoided (Figure 1C), along with confounding from other phenomena such as cross-trait assortative mating and dynastic effects which operate through the parental genotypes (see Figure 1 of (Brumpton et al., 2020)).

We now outline the algorithm in the context of a trio design in which we have genetic data on the parents and the offspring. Let $\mathbf{X}$ and O denote the genotypes and outcome phenotype values respectively for some individual, and let $(\mathbf{X}_n; O_n)_{n=1}^N$, denote these across N trios. Also let **P1** and **P2** denote the genotypes of the parents of the individual with genotypes $\mathbf{X}$, and let $\mathbf{A} := (\mathbf{P1}, \mathbf{P2})$ refer to the set of parental genotypes. Let $\boldsymbol{Z}$ denote the set of external confounders measured on the same individual, which we define as the set of confounders that satisfy $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{A}$. Thus, population stratification is an external confounder (as are assortative mating and dynastic effects) while horizontal pleiotropy is not. The key idea is that we can formulate a hypothesis test of a causal effect conditional on the parental haplotypes $\boldsymbol{A}$. Bates et al. (Bates et al., 2020) show that such a test is also a test of the stronger null hypothesis of a causal effect conditional on $(\boldsymbol{A}, \boldsymbol{Z})$.

The way that this is accomplished is through a conditional randomization test, similar to the Digital Twin Test proposed by Bates et al in the context of GWAS (Bates et al., 2020; Candes et al., 2018). The idea is to sample so-called "digital twins" $\tilde{\mathbf{X}}$ from each set of parents $\mathbf{A}$ such that $\tilde{\mathbf{X}} \mid \mathbf{A}$ has the same distribution as $\mathbf{X} \mid \mathbf{A}$, which can easily be accomplished using the laws of Mendelian inheritance (Methods). We construct B such random samples across all trios, $(\tilde{\mathbf{X}}_n, O_n)_{n=1}^N$, and for each set b of twins we compute a test statistic $t_b = t((\tilde{\boldsymbol{X}}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$ representing the strength of association between the genetically-predicted exposure and the outcome. We also compute a test statistic for the true offspring of the trios, $t^* = t((\boldsymbol{X}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$.

We can then obtain a p-value for a non-zero causal effect of the exposure on the outcome, $p = \frac{1+\mathbf{1}\{t_b \geq t^*\}}{1+B}$. The set of B statistics derived from the digital twins represents a null distribution conditioned on the parental genotypes. If there is a true nonzero effect of the exposure on the outcome, we expect the statistic derived from the true offspring to be stronger than statistics derived from digital twins whose genotypes are randomly sampled from the parental genotypes. The test statistic and algorithm are explained in more detail in the Methods section.

MR-Twin controls for arbitrarily strong population stratification confounding in simulations

Algorithm 1 Simulate genotypes under population structure

procedure SIMGENO(F_{ST}, groups, N,M) ▷ F_{ST} is the fixation index, groups is the number of populations, N is number of samples, M is number of SNPs.
  Initialize the average MAF $\bar{\mathbf{f}}_j \overset{i.i.d.}{\sim} Unif(0.05, 0.5)$ for each SNP j.
  for k $\leq$ groups **do**
    $\mathbf{f}^k \sim Beta(\bar{\mathbf{f}}\frac{(1-F_{ST})}{F_{ST}}, (1-\bar{\mathbf{f}})\frac{(1-F_{ST})}{F_{ST}})$
    Generate genotype matrix $\boldsymbol{X}^{(k)}$ of population k such that $x_{ij}^k \sim Bin(2, \mathbf{f}_j^k)$ for each individual i and SNP j.
  end for
  $\boldsymbol{X} = [\boldsymbol{X}^{(1)}; \ldots; \boldsymbol{X}^{(groups)}]$ ▷ Stack the rows of each genotype matrix
  return Genotype matrix $\boldsymbol{X}$
end procedure

Algorithm 2 Simulate population-stratified phenotypes

procedure GETPHENO($\mathbf{X}$, U, h^2, α_E, γ_{ue}, γ_{uo}) ▷ $\mathbf{X}$ is the normalized genotype matrix, U is a vector with the fixed population label for each sample, h^2 is heritability of exposure E, α_E is effect of E on outcome O, γ_{ue} and γ_{uo} are fixed confounding effects of U on E and O.
  Generate genetic coefficient $\boldsymbol{\beta} \sim \mathcal{N}(0, h^2\mathbf{I}_M)$
  Compute $\sigma_{\epsilon_e}^2 = 1 - h^2$, $\sigma_{\epsilon_o}^2 = 1 - \alpha_E^2$
  Simulate $E = \boldsymbol{X}\boldsymbol{\beta} + \gamma_{ue}U + \epsilon_e$ where $\epsilon_e \sim \mathcal{N}(0, \sigma_{\epsilon_e}^2\boldsymbol{I})$
  Simulate $O = \alpha_E E + \gamma_{uo}U + \epsilon_o$ where $\epsilon_o \sim \mathcal{N}(0, \sigma_{\epsilon_o}^2\boldsymbol{I})$
  return (E,O)
end procedure

We compared the performance of MR-Twin to other MR methods via simulations consisting of two populations with allele frequency differences modeled according to the standard Balding-Nichols model (Balding and Nichols, 1995), following previous works (Ochoa and Storey, 2021; Conomos et al., 2016; Chen et al., 2015; Hubisz et al., 2009; Price et al., 2006; Pritchard et al., 2000). The procedure for simulating the genotypes is outlined in Algorithm 1. We use this algorithm to simulate "external" samples (non-trio data, e.g. from a biobank), as well as the parents for the trios. The offspring genotypes for the trios can then be easily sampled given the parental genotypes (Methods). For each sample, we retain the population label, a binary variable indicating which population each sample belongs to. Unless otherwise specified, each simulation had 50,000 (false positive rate simulations) or 100,000 (power simulations) external samples and 1,000 trio samples evenly split between two populations with fixation index $F_{ST} = 0.01$ and 100 SNPs, 50 of which

were causal for the exposure trait. Unless otherwise specified, the heritability of the exposure trait was set to $h^2 = 0.2$.

Next, we simulate both the exposure and outcome phenotypes following a linear model, as outlined in Algorithm 2. This model allows the population labels from the first step to have an effect on the exposure and outcome phenotypes, which models population stratification that violates the MR assumptions. We use this setting to assess the false positive rates (FPR) of methods under population stratification, allowing the effects of the population labels on the exposure and outcome phenotypes to range from 0 (no confounding) to 0.8 (substantial confounding). In a separate set of simulations to assess power, we set the confounding effect to 0 and varied the causal effect.

We performed 1000 simulation replicates under these settings, each time simulating a set of external and trio genotypes and phenotypes according to the chosen parameters, performing linear regression between each SNP and the exposure and outcome phenotypes, and using the resulting association statistics as input to each of the MR methods. We excluded SNPs with association p-values of above $0.05/M$ ($M = 100$) with the exposure phenotypes in the external data in order to limit weak instrument bias (Burgess et al., 2011). The methods we assessed include the trio mode of MR-Twin, standard inverse-variance weighted (IVW) MR (Burgess et al., 2013), MR-Egger (Bowden et al., 2015), the Weighted Median Estimator (Bowden et al., 2016), the Mode-Based Estimator (Hartwig et al., 2017), and a method introduced by Brumpton et al (Brumpton et al., 2020) to use family data to control for confounding due to population stratification and other population-related effects. Brumpton et al provide a suite of methods for different family datasets, following previous work such as (Fulker et al., 1999); here we focus on the trio-based method they describe (Brumpton et al., 2020), and simply refer to that method as "Brumpton" below.

The trio mode of MR-Twin maintained a calibrated FPR irrespective of the strength of confounding (Figure 2A). Non-family-based methods such as IVW, Egger, Median, and Mode all displayed substantially inflated FPR in the face of confounding consistent with their sensitivity to potential residual population stratification. The Brumpton method also displayed slightly inflated FPR, which increased with the strength of the confounding effect, likely due to weak instrument bias (Brumpton et al., 2020). To mitigate the impact of weak instrument bias, we applied a common approach employed in MR studies (Burgess et al., 2011) that involves filtering out variants for which the F-statistic of the association signal is low ($F < 10$ following previous recommendations).

This rendered the FPR inflation negligible, but also rendered Brumpton substantially less powerful than MR-Twin, whereas the "unfiltered" mode had similar power to MR-Twin (Figure 2B). Results with confidence intervals are shown in Figure S10. We further investigated the weak instrument bias by running simulations with no SNP filtering based on external data and with increasing numbers of SNPs – settings expected to generate large numbers of weak instruments – and confirmed that Brumpton had greater FPR inflation in these settings while MR-Twin remained calibrated (Figure S11) and did not lose power (Figure S12).

The standard MR methods (IVW, Egger, Median, and Mode), when run on the external data, had substantially higher power than the family-based methods, MR-Twin and Brumpton (Figure 2B). We performed additional simulations to understand if the lower power of MR-Twin was due to the smaller number of trios as opposed to methodological limitations. When applied to the offspring in each trio (Figure 3), the standard MR methods still had substantially inflated FPR (Figure 3A) but similar power to MR-Twin and Brumpton (Figure 3B). We also evaluated the FPR and power of these methods under varying number of trios (Figure S1). We observed that increasing number of trios increased power for all methods, as expected, suggesting that the family-based methods can be expected to obtain increased power as more genetic family data are ascertained in the future. The relative power of the methods remained roughly consistent across these experiments.

We also evaluated the Area Under the Receiver Operating Characteristic Curve (AUC-ROC; Figures S5 and S6). Comparing the two main family-based approaches, MR-Twin generally had higher AUC than Brumpton (filtered). Predictably, the AUC of standard MR methods drops sharply when there is confounding, and MR-Twin outperforms these methods in most such cases, though Egger was a notable exception in our findings and remained competitive with family-based approaches even under confounded settings. Similarly to our findings in Figures 2 and 3, family-based methods are more competitive with standard MR methods when run on similar sample sizes. As an additional sensitivity analysis, we also assessed the FPR (Figure S7) and Power (Figure S8) of methods in settings where there are very few instruments or very low heritability. The trends were broadly similar to those seen in Figures 2 and 3. MR-Twin maintained a calibrated FPR in all settings, although it did suffer a loss of power when heritability was very low (Figure S8B and S8D).

We performed simulations increasing the magnitude of population structure as measured by

the F_{ST} (without necessarily increasing the confounding strength), and observed that increasing the population structure leads to further FPR inflation for standard MR methods (Figure S2). We observed inflated FPR for standard MR methods even when there is no confounding (stratification) for large values of F_{ST} (Figure S2) likely due to correlation or linkage disequilibrium among the genetic variants induced by population structure. The standard implementation of IVW and Egger (Yavorska and Burgess, 2017; Broadbent et al., 2020) allows the user to pass in a variant correlation matrix, which removed the FPR inflation with no stratification (Figure S2); other methods such as Median and Mode do not currently have this option.

Next we assessed the runtime of methods run on the trio data (Figure S3). Brumpton, along with non-trio based methods (e.g., IVW), had similar run times (<1-5 seconds per simulation replicate); for succinctness, only Brumpton is shown. MR-Twin (with 100 simulated digital twins) took roughly one minute per simulation replicate under the simulation settings described above, with time increasing to up to four minutes if the number of families or SNPs was increased. The number of digital twins to simulate for MR-Twin involves a trade-off between speed and stability of results. We assessed the stability of MR-Twin with different numbers of digital twins, with results shown in Figure S9. We interpreted these findings as indicating that 100 digital twins are likely stable enough for simulations for which many replicates are run and speed is a priority, but 1000 or more digital twins is recommended for one-off real data analysis. Therefore we simulated 100 digital twins in our simulations and 1000 in our real data analysis. We note, however, that while the MR-Twin runtime increases linearly with the number of digital twins simulated, the generation and statistic computation for the digital twins can be done in parallel, so many twins can be simulated efficiently given multiple compute cores or nodes. For clarity of results, we did not take advantage of this in our runtime assessment.

Finally, MR-Twin also enables users to use parent-child duo or sibling datasets (Methods). We assessed the performance of these modes versus the trio mode of MR-Twin (Figure S4). We found that the duo and sibling modes, while having lower FPR than most standard MR methods, did not maintain a calibrated FPR at high levels of confounding, which is expected since the precise sampling of offspring genotypes from parents is not possible when either or both of the parental genotypes are not available.

Application to trio data in the UK Biobank

In order to assess the results given by MR-Twin relative to other approaches in a real data context, we next applied MR-Twin and four other MR methods (IVW, Egger, Median, and the Brumpton et al. method (Brumpton et al., 2020)) to 144 real trait pairs in the UK Biobank (Bycroft et al., 2018). These consisted of all pairwise combinations of 12 metabolic, anthropometric, and socioeconomic traits that were widely measured among the UK Biobank participants (listed in Table S1). We isolated 955 White British genetic trios from the full UK Biobank dataset (Supplemental Materials) and used PLINK (Purcell et al., 2007) to run linear regression on the remaining unrelated White British individuals for these 12 traits, including the top 20 principal components (PCs), age, and sex as covariates. The genetic instruments selected for each analysis were the SNPs with genome-wide significant p-values ($< 5.0 \times 10^{-8}$) for the exposure trait, after linkage disequilibrium (LD) pruning was performed so that none of these instruments were in substantial LD with one another (Supplemental Materials). Ignoring the degenerate cases where the exposure and outcome were the same trait or where there were no significant SNPs for the exposure trait (as was the case for the Townsend Deprivation Index [TDI]), there were 121 usable trait pairs.

Table 1 shows the results for six selected trait pairs (excluding Median for brevity because it gave similar results to IVW), while Supplemental Table S2 shows the full set of analyses. (Brumpton was run with several different variant filtering settings to assess the impact of potential weak instrument bias (Supplemental Materials); results for all runs are given in Supplemental Table S2.) For Table 1, we selected six analyses: two positive controls representing causal effects that are true by definition (LDL Cholesterol → Total Cholesterol and Weight → Body Mass Index [BMI]), two negative controls that represent seemingly implausible effects (Glucose → TDI and Height → Body Fat), and two trait pairs with unclear or conflicting evidence (BMI → Diastolic Blood Pressure [DBP] and BMI → TDI). In particular, previous studies have identified a significant effect for BMI → DBP (Lyall et al., 2017) and for BMI → TDI in women (Tyrrell et al., 2016) with IVW analysis, although Egger analysis did not replicate the significant findings in either case (Lyall et al., 2017; Tyrrell et al., 2016).

All methods performed as expected on the controls, with highly significant p-values for positive controls and insignificant p-values for negative controls. For BMI → DBP, IVW and Brumpton

yielded significant results while Egger and MR-Twin did not. For BMI → TDI, IVW and Egger yielded significant results while Brumpton and MR-Twin did not. In general, IVW tended to yield much stronger p-values than other methods and the family-based methods (Brumpton and MR-Twin) tended to be conservative (Supplemental Table S2), in line with our simulation results. In particular, of the 121 usable trait pairs, IVW identified 78 as significant, Egger identified 56 as significant, Brumpton identified 20 as significant, and MR-Twin identified 19 as significant.

Discussion

We introduced MR-Twin, a method for testing causal effects between pairs of traits within a Mendelian Randomization (MR) framework, which is provably robust to confounding of any strength resulting from population stratification. Our primary contributions are the following: (i) developing a digital twin test, originally proposed by Bates et al (Bates et al., 2020) in the context of genetic association studies, for MR, coupled with a novel statistic for this test; (ii) demonstrating that, by leveraging trio data, our proposed framework is robust to confounding due to population stratification and to biases from the inclusion of genetic instruments with weak effects; (iii) extending our framework to the setting of sibling data, a setting not considered by Bates et al; (iv) conducting the first (to our knowledge) large-scale evaluation of the digital twin test framework in comparison with existing methods for MR. We demonstrated that existing MR methods, including those designed to correct for confounding resulting from horizontal pleiotropy, are prone to false positives when there is confounding from population stratification.

While population stratification was the focus of this paper, the MR-Twin framework also provides immunity to several other types of confounding effects. Theory dictates that MR-Twin is immune to confounding from familial effects such as assortative mating and dynastic effects since these effects operate through the parental genotypes (see Figure 1 of (Brumpton et al., 2020)), though we do not explicitly test this in this manuscript. As recently demonstrated, cross-trait assortative mating is pervasive and impacts many common genetic analyses (Border et al., 2022) including MR (Hartwig et al., 2018), so this represents another valuable aspect of MR-Twin even if population stratification is believed to be well-controlled in a particular study. In general, MR-Twin is immune to any confounder that is independent of the genotypes of offspring given the

genotypes of their parents. We note that when we refer to "immunity" we mean this in a theoretical sense – for instance, under the assumption that the model for Mendelian inheritance is correct. In our particular implementation, we assume that the genetic instruments have been selected to be roughly independent and thus we can sample digital twin genotypes from the parental genotypes using a binomial model. In practice, of course, genetic variants on the same chromosome are never perfectly independent, though with appropriate caution the dependence is weak enough that the effect on calibration should be negligible. More complex models of meiosis will also rely on other factors such as haplotype phasing accuracy.

In addition to population and familial effects, we highlight two under-appreciated sources of bias in MR studies, both of which MR-Twin avoids without requiring the user to modify any parameters or arguments. The first is weak instrument bias (Burgess et al., 2011), which can bias the effect estimate of standard MR methods, including the Brumpton approach (Brumpton et al., 2020). This accounts for the Brumpton method yielding inflated FPRs when the confounding effects were strong (Figure 2A). One of the most common ways to control for weak instrument bias is by filtering out variants with a weak association signal, often with a threshold of $F < 10$ for the association between a variant and the exposure trait. However, this procedure has been criticized (Burgess et al., 2011) and may not fully correct for weak instrument bias. Other MR methods may also be affected by this bias. In two-sample study designs, the direction of the bias is towards the null rather than the confounded exposure-outcome association estimate (Lawlor, 2016), but the bias remains.

Additionally, we found that standard MR methods (IVW, Egger, Median, Mode, etc) may have inflated FPR when there is population structure that induces correlation between variants, even in the absence of stratification (Figure S2). The reason for the induced correlation is that, even though the variants were simulated independently, they were correlated with one another through the population labels. For example, suppose we have two variants, $X1$ and $X2$ and a population label U. The causal diagram for these three variables is $X1 \leftarrow U \rightarrow X2$, so $X1$ and $X2$ are correlated. Our findings corroborate earlier findings that correlation between SNPs can cause calibration issues in MR methods (Burgess et al., 2013). This phenomenon should be taken into account when performing MR simulations or when applying MR to real datasets where variants may be correlated. In the latter case, users should obtain SNP correlations from an appropriately

population-matched (Peterson et al., 2019) and sufficiently large (Benner et al., 2017) reference panel.

MR-Twin avoids both of these issues, without requiring the user to specify a SNP correlation matrix or apply various approaches to mitigate weak instrument bias. First, both MR-Twin and Brumpton avoid the correlated-variant issue because they condition on parental genotypes, severing the link between the offspring genotypes and the population structure. Second, MR-Twin would not lose FPR calibration due to weak instrument bias, because this phenomenon has nothing to do with the aspects of the MR-Twin test that guarantee immunity from confounding due to population and familial effects (sampling digital twin genotypes conditioned on parental genotypes). Theoretically, it is possible that the bias in the MR effect estimate used in the MR-Twin statistic (Methods) could lower power, but because the MR effect estimate equally affects both the digital twin statistics and the true offspring statistics, a reduction in power seems unlikely and was not observed empirically (Figure S12).

There is extensive literature on family-based methods for avoiding confounding due to population structure in genome-wide association studies or linkage analysis (Weiner et al., 2017; Laird and Lange, 2006; Abecasis et al., 2000; Fulker et al., 1999; Thomson, 1995; Spielman et al., 1993). One prominent example is the transmission disequilibrium test (TDT) (Spielman et al., 1993) and the more-recent polygenic TDT (pTDT) (Weiner et al., 2017). Bates et al (Bates et al., 2020) compare the digital twin test (DTT) to the TDT and show that the DTT is a generalization of the TDT and highlight some of its benefits. Because it is not immediately obvious how to adapt the TDT and pTDT to MR, we do not evaluate their potential use in this context.

There are several considerations that come into play when applying the MR-Twin method, which we note here. First, the number of digital twins simulated involves a trade-off between speed and precision (Figure S9). While MR-Twin was slower than competing MR methods (Figure S3), it still ran in a few minutes or less per run on both simulated and real data analyses, justifying the use of a fairly large number of digital twins if possible. Consequently, we recommend 1000 or more digital twins for real data analysis, which should be computationally feasible and precise (and, again, parallelization can make this quite efficient). 100 digital twins is likely sufficient in simulations where there are many replicates and speed is the paramount concern. Second, the populations of the external and family datasets should be similar. This is natural for biobanks like the UK

Biobank, but can be more challenging when attempting to combine separate datasets. Third, care should be taken to ensure that the normalization method used and covariates controlled-for are similar in the external and trio datasets in order to avoid potential loss of power.

While the genetic trio offspring used in our UK Biobank analysis were all adults (as all participants in this dataset were aged 40-69 at collection time (Bycroft et al., 2018)), other trio datasets may contain young children. This is a potential issue because some commonly-analyzed traits such as height and weight may not have the same relationship in youths or adolescents as they do in adults, and variants that affect these traits may not yet have realized their full effect in the children yet. Dealing with such time-varying exposures in the context of mendelian randomization is an area of ongoing research (Labrecque and Swanson, 2019), and it is not clear how this would impact MR-Twin results. Even when the offspring of the trios are all adults, it may be difficult to adequately sample certain traits. For example, we were not able to perform MR analysis for complex traits such as heart disease, since none of the offspring in our sample had developed heart disease, largely because all offspring in our sample were aged 40-49.

We note a few trends seen across many trait pairs in the real data results, reflecting some practical considerations. First, all standard MR methods identified substantially more trait pairs than did either family-based approach. Given our simulation results showing a large power difference in the methods when run with different sample sizes (Figure 2) but similar power when run with the same sample size (Figure 3), along with the fact that the UK Biobank has many more unrelated individuals than trios, we believe that this difference is largely due to the difference in the available sample sizes between unrelated and trio data. The number of trios available as part of public datasets is currently relatively small, limiting the power of family- or trio-based methods, but future increases in the number of available trios will lead to increases in the power of MR-Twin and other family-based methods. Second, some traits pairs had quite different results when the exposure and outcome traits were switched. For example, none of the standard MR methods had significant p-values for DBP $\rightarrow$ Weight, but all were significant for Weight $\rightarrow$ DBP (Supplemental Table S2). This may be due to one causal direction being correct while the other is incorrect, but may also be affected by factors such as differences in the heritability and/or polygenicity between the two traits.

Several extensions to the methods presented here are also possible. While we explored contin-

uous traits in this paper, further work needs to be done to apply MR-Twin to binary phenotypes such as disease labels. First, a different statistic such as binary cross entropy (rather than our negative squared loss statistic) may be more appropriate. Second, the use of the external effect size estimates in the statistic may have to be modified, depending on the regression method used and the interpretation of the estimates. For example, it would be inappropriate to replace the effect size estimates in our statistic with odds ratios produced by logistic regression. Even for linear data, it is possible that a different statistic than the one we proposed would be more powerful in some situations. Finding most-powerful statistics for a given significance threshold is a direction for future work. Future work could also improve upon the sibling mode of MR-Twin by using population-based priors to infer parental genotypes with a greater level of accuracy, thereby obtaining superior control of false positives. This approach could in principle be developed for and applied to more extended pedigrees.

In the Digital Twin Test paper (Bates et al., 2020), Bates et al propose using a Hidden Markov Model (HMM) to simulate digital twins from the parental haplotypes, the latter being generated by phasing the parental genotypes. For the simplicity of avoiding this phasing step and due to the fact that genetic instruments in MR studies are usually selected to be roughly independent (Burgess et al., 2013), we used a simpler method for simulating digital twins using binomial draws from the parental genotypes (Methods). However, the variants used may not be independent even if they appear to be (Burgess et al., 2013), or one may wish to include correlated variants to increase power. Extending MR-Twin to perform the HMM-based digital twin simulation could therefore increase power.

Finally, a pre-print from Tudball *et al.* proposes a randomzation-based approach to MR (Tudball et al., 2022) that, while being conceptually similar, differs from MR-Twin in a few practical aspects. First, Tudball *et al.* (Tudball et al., 2022) do not discuss the use of external summary statistics to increase power, whereas this is a core part of the MR-Twin approach (as well as in the Digital Twin Test of Bates *et al.* (Bates et al., 2020)). Second, Tudball *et al.* develop family-based propensity scores for individual SNPs and suggest aggregating them with Fisher's method or another p-value aggregation method, which is substantially different from our proposed sum-of-squares statistic over all SNPs (Methods). Finally, Tudball *et al.* employ the HMM-based digital twin simulation model, while (as discussed above) we employ the simpler binomial model. Nevertheless, the broad

conceptual similarities between the two methods highlight the promise of randomization-based approaches to make MR findings more robust and the value of continued development to extend these approaches to more complex pedigrees.

Methods

The MR-Twin framework

We first introduce the standard Mendelian Randomization (MR) model, without any confounding. Suppose that for a collection of N individuals we obtain their genotypes at M SNPs, and a phenotypic measure for an exposure trait and an outcome trait. For a given individual n we denote the genotype vector as $\mathbf{X}_n$, the genotype at some SNP j as $\mathbf{X}_{nj}$, the exposure trait as E_n, and the outcome trait as O_n. Let $(\mathbf{X}_n, E_n, O_n)_{n=1}^N$ denote the collection of these genotypes and traits over all N individuals, where $(\mathbf{X}_n)$ is an $(N \times M)$ matrix and (E_n) and (O_n) are $(N \times 1)$ vectors. Finally, let $\mathbf{X}$, E, and O refer to the genotype vector, exposure trait, and outcome trait for a generic individual.

MR uses the genetic "instruments" $\mathbf{X}$ to estimate the effect of an "exposure" trait E on an "outcome" trait O. This estimate is valid regardless of any confounder $\mathbf{U}$ of the association between E and O, assuming that the following conditions hold (Lawlor et al., 2008):

1. The genetic instrument $\mathbf{X}$ is significantly associated with the exposure trait E;
2. The genetic instrument $\mathbf{X}$ is independent of any variables (such as those in $\mathbf{U}$) that confound the relationship between E and the outcome trait O;
3. The genetic instrument $\mathbf{X}$ is not associated with O except due to its association with E.

The latter two criteria can be captured by the independence statement

$$\mathbf{X} \perp\!\!\!\perp O \mid E \tag{1}$$

Assuming these conditions hold, and assuming a linear model for the relationships between the genotypes and phenotypes (a typical assumption in MR analyses), we can test the null hypothesis that there is no direct causal effect of E on O,

$$H_0 : \beta_{EO} = 0 \tag{2}$$

where β_{EO} is not obtained by direct regression but rather via instrumental variables estimators such as the ratio estimator $\beta_{EO} = \beta_{XO}/\beta_{XE}$ (when a single instrument is used) or by two stage least squares or inverse-variance weighting (when multiple instruments are used) (Burgess et al., 2013).

However, in the case where we have residual population stratification, denoted $\mathbf{Z}$ (Figure 1A), this independence assumption is violated. This is because, using terminology from Pearl's graphical formalism (Pearl, 1995), $\mathbf{X} \leftarrow \mathbf{Z} \rightarrow O$ is a backdoor path between $\mathbf{X}$ and O, so the two are not marginally independent. Conditioning on E fails to block this backdoor path (i.e. see Figure 1A). Residual population stratification generally cannot be controlled for directly, though approaches such as Principal Components Analysis (PCA) and Linear Mixed Models (LMMs) have been used to reduce its effect (Price et al., 2010).

MR-Twin (Figure 1C) is a method that uses family-based genetic data to avoid this confounding. Suppose that, corresponding to each individual's genotypes $\mathbf{X}$, we also observe the genotypes $\mathbf{P1}$ and $\mathbf{P2}$ of their parents (we later relax the trio assumption to allow for parent-child duo or sibling data). Let $\mathbf{A} := (\mathbf{P1}, \mathbf{P2})$. According to the graphical criteria for d-separation developed by Pearl (Pearl, 1995), $\mathbf{A}$ d-separates $\mathbf{X}$ from $\mathbf{Z}$ (Figure 1B):

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{A} \tag{3}$$

This means that, assuming $\mathbf{X}$ does not affect some unmeasured variable which in turn affects O (i.e. no horizontal pleiotropy),

$$\mathbf{X} \perp\!\!\!\perp O \mid E, \mathbf{A} \tag{4}$$

thereby satisfying the MR conditions regardless of any residual population stratification.

As shown by Bates et al (Bates et al., 2020), the Digital Twin Test framework outlined in Algorithm 3 can be used to perform a hypothesis test conditioned on $\mathbf{A}$. The resulting test involves computing the test statistic $t^* = t(\boldsymbol{X}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$ (we give the statistic used in this paper in "MR-Twin Test Statistic Incorporating External Weights" below). To perform a test, we construct B

random samples ($\tilde{\mathbf{X}}_n$) where each $\tilde{\mathbf{X}}$ is a random sample given $\mathbf{A}$ with the same distribution as $\mathbf{X}$ given $\mathbf{A}$ (such a sample can be easily constructed using Mendelian inheritance; see "Generating Digital Twins" below). We refer to these samples as "digital twins". For each such sample b, we then compute $t_b = t((\tilde{\boldsymbol{X}}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$, representing a null distribution of genotypes conditioned on the parental genotypes. This, in turn, gives us a p-value for $t^* = \frac{1+\mathbf{1}t_b \geq t^*}{1+B}$, where B is the total number of permutations we perform. The MR-Twin test is therefore a kind of conditional randomization test (Bates et al., 2020; Candes et al., 2018).

Importantly, the proposed algorithm can leverage effect size estimates ($\hat{\boldsymbol{\beta}}$) from any external GWAS datasets (even GWAS datasets where such estimates might be biased due to population stratification) while providing valid tests. The proposed algorithm is robust to any external confounder satisfying Equation 3, such as population stratification, assortative mating, and dynastic effects.

Algorithm 3 Outline of MR-Twin

1. *Input:* Effect sizes for SNPs: $\hat{\boldsymbol{\beta}}$, trio data $\{(\boldsymbol{X}_n, \boldsymbol{A}_n, O_n)_{n=1}^N\}$
2. Compute the MR-Twin test statistic $t^* = t((\boldsymbol{X}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$
3. For b = 1 to B:
   (a) Sample digital twins $\tilde{\boldsymbol{X}}_n$ given their ancestors $\boldsymbol{A}_n$.
   (b) Compute the MR-Twin test statistic $t_b = t((\tilde{\boldsymbol{X}}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$
4. $p = \frac{1+\mathbf{1}\{t_b \geq t^*\}}{1+B}$

Output: p-value: p

Next, we detail the MR-Twin test statistic, digital twin generation algorithms, and formal proofs of the exchangeability of digital twins with each other and their real counterparts.

Conditional randomization test for mendelian randomization

The MR-Twin test is related to the digital twin test (Bates et al., 2020) and likewise is a kind of conditional randomization test (Candes et al., 2018). Like the digital twin test, MR-Twin leverages the fact that offspring genotypes are conditionally independent of "external" confounders such as population structure given the parental genotypes and uses a conditional randomization test to test the weaker, but equivalent, null hypothesis of no effect conditioned upon the parental genotypes.

Let $\mathbf{X}$ be a vector of offspring genotypes, and let $\mathbf{A}$ be the genotype vectors of the two parents

of the offspring. $\mathbf{A}$ may be directly observed, as in trio data, or inferred using parent-child duo or sibling data (see "Generating Digital Twins"). Let $\mathbf{Z}$ be one or more "external" confounders, defined (Bates et al., 2020) as

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{A} \tag{5}$$

Thus, population structure is an external confounder, while horizontal pleiotropic traits are not. We therefore have

$$\mathbf{X}|(\mathbf{Z}, \mathbf{A}) \stackrel{d}{=} \mathbf{X}|\mathbf{A} \tag{6}$$

Assuming that all confounders are external and that $\mathbf{X}$ is significantly associated with E, O is independent of $\mathbf{X}$ given $\mathbf{A}$ under the MR null hypothesis that E has no effect on O. This is because $\mathbf{X}$ would not have any effects on O mediated by E (since E does not affect O under the MR null hypothesis), and all paths not through E are blocked by conditioning on $\mathbf{A}$ as shown in Equation 6. We therefore want to test

$$\mathbf{X} \perp\!\!\!\perp O \mid \mathbf{A} \tag{7}$$

If this holds, then we cannot rule out that either $\mathbf{X}$ has no effect on E or E has no effect on O. We test this null hypothesis via a conditional randomization test (Candes et al., 2018).

In testing this null hypothesis, it is helpful to be able to leverage SNP effect sizes estimated from large, external datasets (such as publicly released summary statistics for resources like the UK Biobank (Bycroft et al., 2018)), as this will often yield more statistically significant variants and better effect size estimates than those generated using small genetic family datasets. We therefore note that the following property also holds:

$$\mathbf{X} \perp\!\!\!\perp \hat{\beta} \mid \mathbf{A} \tag{8}$$

where we use the shorthand $\hat{\beta}$ to refer to the estimated effect sizes of each SNP on the exposure and outcome traits.

We construct "digital twins" $\tilde{\mathbf{X}}$ sampled from the parental genotypes via Mendelian inheritance

(see "Generating Digital Twins") such that

$$\tilde{\mathbf{X}}|\mathbf{A} \stackrel{d}{=} \mathbf{X}|\mathbf{A} \tag{9}$$

Given equations 7, 8, and 9, we have the following under the null hypothesis:

$$\mathbf{X}|(\mathbf{A}, \hat{\beta}, O) \stackrel{d}{=} \mathbf{X}|\mathbf{A} \tag{10}$$

$$\tilde{\mathbf{X}}|(\mathbf{A}, \hat{\beta}, O) \stackrel{d}{=} \tilde{\mathbf{X}}|\mathbf{A} \tag{11}$$

It follows from equations 9, 10, and 11 that the digital twins are exchangeable under the null hypothesis:

$$\tilde{\mathbf{X}}|(\mathbf{A}, \hat{\beta}, O) \stackrel{d}{=} \mathbf{X}|(\mathbf{A}, \hat{\beta}, O) \tag{12}$$

Therefore, given some statistic $T = t((\boldsymbol{X}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$, where N is the number of families,

$$T|(\mathbf{A}, \hat{\beta}, O) \stackrel{d}{=} \tilde{T}|(\mathbf{A}, \hat{\beta}, O) \tag{13}$$

under the null, where $\tilde{T} = t((\tilde{\boldsymbol{X}}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}})$. We can then use the procedure outline in Algorithm 3 to obtain a p-value for this test statistic (Candes et al., 2018).

MR-Twin test statistic incorporating external weights

We construct a test statistic based on a negative sum of squares loss when using $\mathbf{X}$ to predict O via an MR estimate for the effect of E on O. First, we leverage the effect sizes from the external dataset of the genotype on the exposure trait $\hat{\beta}_{XE}$ to obtain the genetically-predicted exposure trait values:

$$\hat{E}_n = \sum_j \hat{\beta}_{XE,n} \mathbf{X}_{nj} \tag{14}$$

for each individual n and SNP j. We then compute the MR estimate for the effect of the exposure trait on the outcome trait, $\hat{\beta}_{EO}$. This estimate may be a conventional Inverse Variance Weighted

(IVW) estimate (Burgess et al., 2013) or various statistics designed to be robust to pleiotropy such as the Egger-based statistic (Bowden et al., 2015), the weighted median statistic (Bowden et al., 2016), or others. We then predict the outcome trait for each individual n as $\hat{O}_n = \hat{\beta}_{EO}\hat{E}_n$. Finally, we compute the negative squared error of these predictions $-\sum_n(\hat{O}_n - O_n)^2$, summed across all individuals. The full statistic is then

$$t((\boldsymbol{X}_n; O_n)_{n=1}^N; \hat{\boldsymbol{\beta}}) = -\sum_n((\hat{\beta}_{EO}\sum_j(\hat{\beta}_{XE,n}\mathbf{X}_{nj})) - O_n)^2 \tag{15}$$

Generating digital twins

We have assumed that trio data is available thus far for simplicity. However, the MR-Twin framework can also be used when parent-child duo data or sibling data are available. Here we discuss the algorithms used to generate digital twins given trio, parent-child duo, or sibling data.

Trio and duo modes

We assume that the SNPs used in the MR instrument are independent, a common assumption when multi-SNP instruments are used in MR (Burgess et al., 2013). Therefore, we separately sample the genotype of each SNP of the digital twin given the parent and/or offspring genotypes at that SNP. Let $(\mathbf{D}_n)$ be the $(N \times M)$ matrix of digital twin genotypes we will sample, corresponding to the true "offspring" genotypes in $(\mathbf{X}_n)$. Further, let n index some family and j index some SNP, such that $\mathbf{P1}_{nj}$ (for example) is the genotype for one parent in family n at SNP j. If we have both parents available, sampling $\mathbf{D}_{nj}$ is straightforward. Because the SNPs are considered independent, we do not need to know the parental haplotypes. If a parental genotype $\mathbf{P1}_{nj}$ is 0 or 2, respectively, then a 0 or 1, respectively, is inherited by $\mathbf{D}_{nj}$. If the parent genotype is 1, then either 0 or 1 is inherited with 50% probability each. $\mathbf{D}_{nj}$ inherits alleles from the two parents independently. This can be summarized as

$$\mathbf{D}_{nj} \sim Bern(\mathbf{P1}_{nj}/2) + Bern(\mathbf{P2}_{nj}/2) \tag{16}$$

where $Bern$ stands for the Bernoulli distribution, for each family n and SNP j.

If we only have one parent genotype available, then following Bates et al (Bates et al., 2020),

we fix the offspring's haplotype from the unobserved parent and only simulate a random draw from the observed parent's haplotype. If the observed parent is homozygous, then the allele inherited from that parent is fixed as well, so $\mathbf{D}_{nj} = \mathbf{X}_{nj}$. Otherwise, the allele inherited from this parent will be $Bern(0.5)$. In principle, 0.5 could be replaced with some value based on population allele frequencies. Similarly to the above, the model for the allele from the other parent can be written as $Bern(\mathbf{X}_{nj}/2)$. Thus, if the parent is a heterozygote, we have

$$\mathbf{D}_{nj} \sim Bern(1/2) + Bern(\mathbf{X}_{nj}/2) \tag{17}$$

Sibling mode

In the case where we observe sibling genotypes but not the genotypes of their parents, we assessed two potential approaches. In either case, the observed sibling information is used to infer the probabilities of digital twin genotypes based on the fact that the sibling genotypes give information about the probabilities of various parental genotypes. For instance, a child with a 2 genotype at a SNP guarantees that neither parent has a 0 genotype at that SNP, and makes it more likely that the parents have 2 genotypes than 1 genotypes. Most simply, if one sibling has a 2 genotype at a SNP and the other sibling has a 0, then the parents must both be heterozygotes. In all other cases, approximation is needed.

The first approach is straightforward and involves randomly drawing two haplotypes from the observed sibling haplotypes to generate a digital twin. This shuffling approach gives a rough approximation of the likelihood of digital twin genotypes given the information the observed siblings provide. The second approach, described in the Supplemental Materials, involves using the sibling data to infer a distribution over the possible parents, then performing a weighted random draw of digital siblings based on those parents. In practice, we found that the shuffling approach was faster and yielded lower FPR than the probabilistic approach while achieving similar power, so we used the shuffling approach for the results in this paper.

Software availability

The code implementing the MR-Twin package can be found at: https://github.com/nlapier2/MR-Twin. Scripts and instructions for repeating the experiments in this paper can be found at:

https://github.com/nlapier2/MRTwin-replication. This code is also available in the Supplemental Code files. Please note that UK Biobank genotypes are not publicly released, so those wishing to replicate the experiments will first have to get access to that data via https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 33127. B.F. and S.S. were supported in part by NIH R35GM125055 and NSF CAREER-1943497, III-2106908. E.E. and N.L. are funded by NSF award 2106908 and NIH awards U01HG011715 and R56HG010812. The authors would also like to thank Matthew J. Tudball for productive discussions on potential future work and our respective efforts to develop family-based Mendelian Randomization methods.

Author contributions: S.S. and E.E. conceived of and supervised the project. N.L., B.F., and S.S. developed the methods and wrote the manuscript. N.L., B.F., and S.T. wrote the software code and performed the analyses. All authors read and approved the final manuscript.

References

Abecasis GR, Cardon LR, and Cookson W. 2000. A general test of association for quantitative traits in nuclear families. *The American Journal of Human Genetics* **66**: 279–292.

Balding DJ and Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96**: 3–12.

Bates S, Sesia M, Sabatti C, and Candès E. 2020. Causal inference in genetic trio studies. *Proceedings of the National Academy of Sciences* **117**: 24117–24126.

564 Benner C, Havulinna AS, Järvelin MR, Salomaa V, Ripatti S, and Pirinen M. 2017. Prospects
565 of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide
566 association studies. *The American Journal of Human Genetics* **101**: 539–551.

567 Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang
568 X, Racimo F, Pritchard JK, et al.. 2019. Reduced signal for polygenic adaptation of height in
569 uk biobank. *eLife* **8**: e39725.

570 Border R, Athanasiadis G, Buil A, Schork AJ, Cai N, Young AI, Werge T, Flint J, Kendler KS,
571 Sankararaman S, et al.. 2022. Cross-trait assortative mating is widespread and inflates genetic
572 correlation estimates. *Science* **378**: 754–761.

573 Bowden J, Davey Smith G, and Burgess S. 2015. Mendelian randomization with invalid instru-
574 ments: effect estimation and bias detection through egger regression. *International Journal of*
575 *Epidemiology* **44**: 512–525.

576 Bowden J, Davey Smith G, Haycock PC, and Burgess S. 2016. Consistent estimation in mendelian
577 randomization with some invalid instruments using a weighted median estimator. *Genetic Epi-*
578 *demiology* **40**: 304–314.

579 Broadbent JR, Foley CN, Grant AJ, Mason AM, Staley JR, and Burgess S. 2020. Mendelianran-
580 domization v0.5.0: updates to an r package for performing mendelian randomization analyses
581 using summarized data. *Wellcome Open Research* **5**.

582 Brumpton B, Sanderson E, Heilbron K, Hartwig FP, Harrison S, Vie GÅ, Cho Y, Howe LD, Hughes
583 A, Boomsma DI, et al.. 2020. Avoiding dynastic, assortative mating, and population stratification
584 biases in mendelian randomization through within-family analyses. *Nature Communications* **11**:
585 1–13.

586 Burgess S, Butterworth A, and Thompson SG. 2013. Mendelian randomization analysis with
587 multiple genetic variants using summarized data. *Genetic Epidemiology* **37**: 658–665.

588 Burgess S, Thompson SG, and Collaboration CCG. 2011. Avoiding bias from weak instruments in
589 mendelian randomization studies. *International Journal of Epidemiology* **40**: 755–764.

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, et al.. 2018. The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**: 203–209.

Candes E, Fan Y, Janson L, and Lv J. 2018. Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**: 551–577.

Chen G, Yuan A, Shriner D, Tekola-Ayele F, Zhou J, Bentley AR, Zhou Y, Wang C, Newport MJ, Adeyemo A, et al.. 2015. An improved fst estimator. *PLOS One* **10**: e0135368.

Cinelli C, LaPierre N, Hill BL, Sankararaman S, and Eskin E. 2022. Robust mendelian randomization in the presence of residual population stratification, batch effects and horizontal pleiotropy. *Nature Communications* **13**: 1093.

Conomos MP, Reiner AP, Weir BS, and Thornton TA. 2016. Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics* **98**: 127–148.

Cook JP, Mahajan A, and Morris AP. 2020. Fine-scale population structure in the uk biobank: implications for genome-wide association studies. *Human Molecular Genetics* **29**: 2803–2811.

Fulker D, Cherny S, Sham P, and Hewitt J. 1999. Combined linkage and association sib-pair analysis for quantitative traits. *The American Journal of Human Genetics* **64**: 259–267.

Haase CL, Tybjærg-Hansen A, Ali Qayyum A, Schou J, Nordestgaard BG, and Frikke-Schmidt R. 2012. Lcat, hdl cholesterol and ischemic cardiovascular disease: a mendelian randomization study of hdl cholesterol in 54,500 individuals. *The Journal of Clinical Endocrinology & Metabolism* **97**: E248–E256.

Hartwig FP, Davey Smith G, and Bowden J. 2017. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology* **46**: 1985–1998.

Hartwig FP, Davies NM, and Davey Smith G. 2018. Bias in mendelian randomization due to assortative mating. *Genetic Epidemiology* **42**: 608–620.

Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, Carslake D, Hemani G, Paternoster L, Smith GD, et al.. 2019. Apparent latent structure within the uk biobank sample has implications for epidemiological analysis. *Nature Communications* **10**: 1–9.

Haycock PC, Burgess S, Nounu A, Zheng J, Okoli GN, Bowden J, Wade KH, Timpson NJ, Evans DM, Willeit P, et al.. 2017. Association between telomere length and risk of cancer and non-neoplastic diseases: a mendelian randomization study. *JAMA Oncology* **3**: 636–651.

Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, et al.. 2018. The mr-base platform supports systematic causal inference across the human phenome. *eLife* **7**: e34408.

Hubisz MJ, Falush D, Stephens M, and Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**: 1322–1332.

Labrecque JA and Swanson SA. 2019. Interpretation and potential biases of mendelian randomization estimates with time-varying exposures. *American Journal of Epidemiology* **188**: 231–238.

Laird NM and Lange C. 2006. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics* **7**: 385–394.

Lawlor DA. 2016. Commentary: Two-sample mendelian randomization: opportunities and challenges. *International Journal of Epidemiology* **45**: 908.

Lawlor DA, Harbord RM, Sterne JA, Timpson N, and Davey Smith G. 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**: 1133–1163.

Lyall DM, Celis-Morales C, Ward J, Iliodromiti S, Anderson JJ, Gill JM, Smith DJ, Ntuk UE, Mackay DF, Holmes MV, et al.. 2017. Association of body mass index with cardiometabolic disease in the uk biobank: a mendelian randomization study. *JAMA Cardiology* **2**: 882–889.

Ochoa A and Storey JD. 2021. Estimating fst and kinship for arbitrary population structures. *PLOS Genetics* **17**: e1009241.

Pearl J. 1995. Causal diagrams for empirical research. *Biometrika* **82**: 669–688.

Peterson RE, Kuchenbaecker K, Walters RK, Chen CY, Popejoy AB, Periyasamy S, Lam M, Iyegbe C, Strawbridge RJ, Brick L, et al.. 2019. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**: 589–603.

Pingault JB, O'reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, and Dudbridge F. 2018. Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* **19**: 566–580.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, and Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**: 904–909.

Price AL, Zaitlen NA, Reich D, and Patterson N. 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**: 459–463.

Pritchard JK, Stephens M, and Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al.. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**: 559–575.

Spielman RS, McGinnis RE, and Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *The American Journal of Human Genetics* **52**: 506–516.

Thomson G. 1995. Mapping disease genes: family-based association studies. *The American Journal of Human Genetics* **57**: 487.

Tudball MJ, Smith GD, and Zhao Q. 2022. Almost exact mendelian randomization. *arXiv preprint arXiv:2208.14035* .

Tyrrell J, Jones SE, Beaumont R, Astley CM, Lovell R, Yaghootkar H, Tuke M, Ruth KS, Freathy RM, Hirschhorn JN, et al.. 2016. Height, body mass index, and socioeconomic status: mendelian randomisation study in uk biobank. *BMJ* **352**: i582.

668 Wade KH, Carslake D, Sattar N, Davey Smith G, and Timpson NJ. 2018. Bmi and mortality in
669 uk biobank: revised estimates using mendelian randomization. *Obesity* **26**: 1796–1806.

670 Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, Samocha KE, Goldstein JI,
671 Okbay A, Bybjerg-Grauholm J, et al.. 2017. Polygenic transmission disequilibrium confirms that
672 common and rare variation act additively to create risk for autism spectrum disorders. *Nature*
673 *Genetics* **49**: 978–985.

674 Yavorska OO and Burgess S. 2017. Mendelianrandomization: an r package for performing mendelian
675 randomization analyses using summarized data. *International Journal of Epidemiology* **46**: 1734–
676 1739.

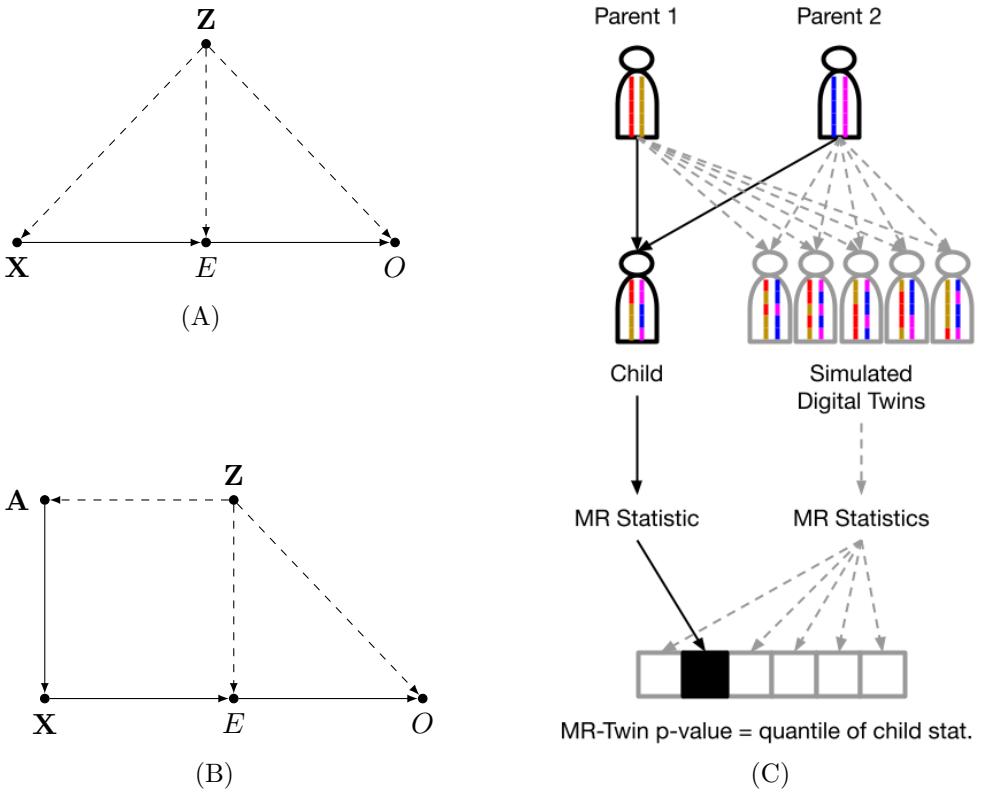
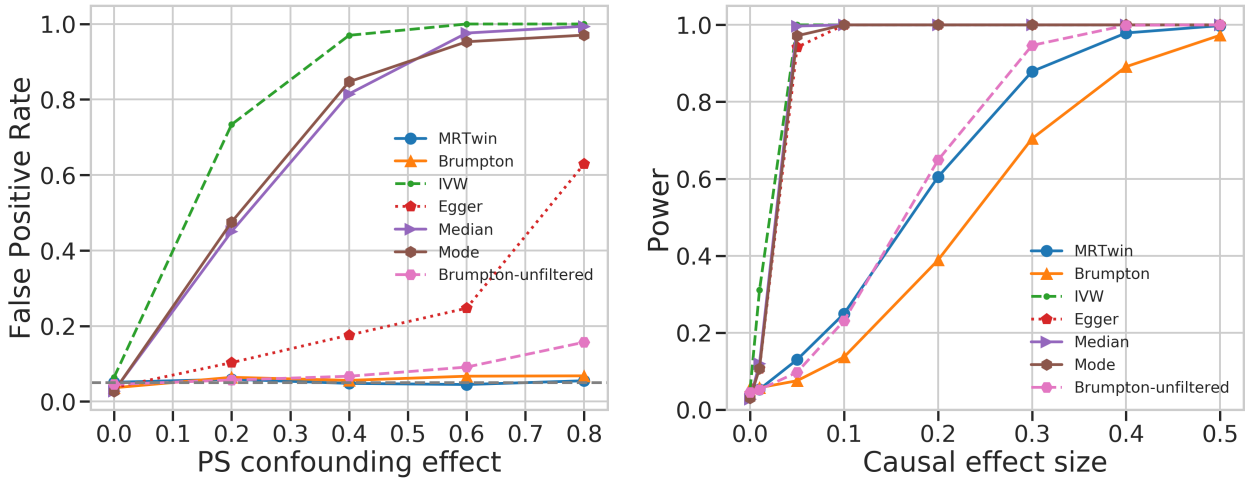
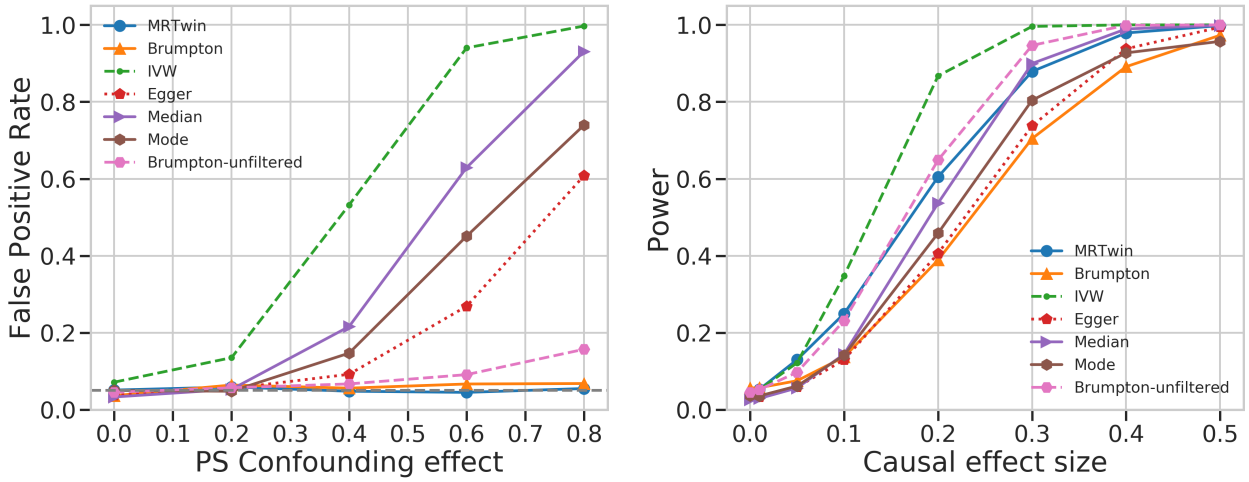


Figure 1: Illustrations of Mendelian Randomization assumptions and the MR-Twin framework. (A) Directed Acyclic Graph (DAG) depicting variables and their relationships in a typical Mendelian Randomization (MR) study, where X is the genotypic instrument, E is the exposure trait, and O is the outcome trait. An external confounder Z, such as population stratification, can cause violations of the MR assumptions. (B) If we have the parental haplotypes A, then X is independent of Z given A. (C) Illustration of the MR-Twin workflow. Digital twin genotypes are sampled from the parental genotypes. MR-Twin is a conditional randomization test, conditioned on A and therefore immune to confounding from Z, in which the p-value is computed based on the quantile of the true offspring's MR-Twin statistic compared to the digital twins' statistics.



(A) (B)

Figure 2: False Positive Rate (FPR) and Power comparison between various methods run on simulated data. (A) False positive rate (y-axis) under varying levels of confounding due to population stratification (PS), with the x-axis describing the magnitude of the confounding effect of population labels on the exposure and outcome trait. (B) Power (y-axis) as a function of the magnitude of the causal effect of the exposure on the outcome trait (x-axis) in a setting with no confounding. Results are averaged over 1000 simulation replicates.



(A) (B)

Figure 3: False Positive Rate (FPR) and Power comparison between various methods run on simulated trio data. This is similar to Figure 2 except that IVW, Egger, Median, and Mode are run on the offsprings of the trio dataset instead of the large "external" group of unrelated individuals, such that all methods have the same sample size. (A) False positive rate (y-axis) under varying levels of confounding due to population stratification (PS), with the x-axis describing the magnitude of the effect of the population labels on the exposure and outcome trait. (B) Power (y-axis) as a function of the causal effect size (x-axis). Results are averaged over 1000 simulation replicates.

	MR P-Values			MR-Twin P-Value
Traits	IVW	Egger	Brumpton	MR-Twin
LDL Chol. → Total Chol.	$< \mathbf{10^{-300}}$	$< \mathbf{10^{-300}}$	$\mathbf{1.64 \times 10^{-11}}$	$\leq \mathbf{9.99 \times 10^{-4}}$
Weight → BMI	$< \mathbf{10^{-300}}$	$< \mathbf{10^{-300}}$	$\mathbf{4.80 \times 10^{-6}}$	$\leq \mathbf{9.99 \times 10^{-4}}$
BMI → DBP	$\mathbf{2.24 \times 10^{-26}}$	5.64×10^{-1}	$\mathbf{3.46 \times 10^{-2}}$	2.69×10^{-1}
BMI → TDI	$\mathbf{1.18 \times 10^{-19}}$	$\mathbf{7.53 \times 10^{-3}}$	9.99×10^{-2}	8.79×10^{-2}
Glucose → TDI	1.54×10^{-1}	2.09×10^{-1}	6.61×10^{-1}	1.91×10^{-1}
Height → Body Fat	9.55×10^{-1}	9.83×10^{-2}	6.73×10^{-1}	5.09×10^{-1}

Table 1: Traditional MR results and MR-Twin results on selected trait pairs from the UK Biobank. Bold numbers are significant at $p < 0.05$. Note that $9.99 \times 10^{-4} = 1/1001$ is the minimum p-value for MR-Twin in this expriment, as 1000 digital twins were generated. Chol. = Cholesterol; BMI = Body Mass Index; DBP = Diastolic Blood Pressure; TDI = Townsend Deprivation Index.