Supplementary Note for: 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.

Simulation and software details

Standard Mendelian Randomization (MR) methods (IVW [1], Egger [2], Median [3], and Mode [4]) were run using the MendelianRandomization R package [5, 6] version 0.5.1. For all the MR methods, unless specified, we used the default options provided by the software. All regressions in the simulations were performed with the R *lm* function. For real data analysis, the Genome-Wide Association Study (GWAS) statistics were calculated using the PLINK software [7], version v1.90b6.6. We implemented MR-Twin in both R and Python; the Python version was used to generate the results in the paper except for the running time analysis (see below). We implemented the trio method from Brumpton et al [8] in R (referred to as "Brumpton" below). Brumpton and MR-Twin can use any summary-level MR statistic; we used IVW by default. R version 4.0.2 was used.

By default, the experiments of each simulation setting were run across 100 different seeds, and under each seed, 10 replicates were simulated, so in total, 1000 replicates were generated for each setting. For runtime analysis (Figure S3), Xeon(R) CPU E5-2670 compute nodes were used. For consistency with all the other MR methods implemented in R, we conducted the time complexity analysis using the R version of MR-Twin.

Replication instructions and details are available at https://github.com/nlapier2/MRTwin-replication.

UK Biobank data preprocessing and analysis

We first isolated the genetic trios from the UK Biobank data [9]. We filtered the data to only include people with self-reported white British ancestry who were not closely related, (e.g. no first, second, or third degree relatives), as defined by pairs of individuals who had a kinship coefficient $< (1/2)^{(9/2)}$ (following Hou and Burch et al [10]), leaving 291,274 people. We used the KING software [11] to estimate kinship coefficients for each pair of White British individuals in the dataset. Following the inference criteria set out by Manichaikul et al [11], we isolated pairs of individuals whose kinship coefficients (ϕ) were within $(\frac{1}{2^{3/2}}, \frac{1}{2^{5/2}})$ – these were the inferred parent-offspring pairs.

When an individual was involved in two or more parent-offspring pairs, we used the following procedure to identify trios and the children and parents for each of those trios. If an individual was found to be involved in multiple parent-offspring pairs, we considered this a family. Unrelated individuals within the family (as determined by the kinship coefficient) were considered parents. If more than two such parents were identified, the family was discarded. The family was also discarded if there were more than one children, if the parents had the same sex, or if the age gap between the parents and the child was less than 10. The remaining data yielded 955 trios, similar to previously-reported amounts [9].

We next performed MR analysis using Inverse Variance Weighting (IVW), Egger regression [2], the Weighted Median Estimator [3], the trio-based method introduced by Brumpton et al [8], and MR-Twin on 144 pairwise combinations of 12 traits (Table S1) in the UK Biobank [9]. To gather genetic instruments for each of the twelve traits, we performed a Genome-Wide Association Study (GWAS). First, we selected individuals with self-reported White British ancestry whose pairwise kinship values were all less than $0.5^{4.5}$, indicating no first, second, or third degree relatives were in the dataset. These individuals did not overlap with any of the individuals in the trio dataset. Phenotype values were standardized to have zero mean and unit variance. We then used PLINK

[7] to run linear regression on the unrelated White British individuals for these 12 traits, including the top 20 principal components (PCs), age, and sex as covariates.

For each exposure trait, we performed filtering of the SNPs as follows in order to select genetic instruments. We first removed all SNPs that did not reach genome-wide significance for the trait (p-value $< 5.0 \times 10^{-8}$). We then performed linkage disequilibrium (LD) pruning: for any pair of SNPs with $r^2 > 0.1$, the SNP with the less significant p-value was removed. The remaining SNPs were used as genetic instruments. The Townsend Deprivation Index (TDI) did not yield any significant instruments, so it was not used as an exposure trait. The remaining traits had 46 to 1502 genetic instruments, with 10 of the 12 having between 96 and 339 instruments (Table S1). 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 TDI), there were 121 usable trait pairs.

As in the simulations, we first applied Brumpton with all variants with association statistic $F < 10$ with the exposure in the trio data filtered out. However, this yielded zero remaining variants for 10 of the 12 exposure traits, so very few significant trait pairs were found. Consequently, we applied a p-value filter of $p > 0.05$ instead; the results for this setting are the results discussed in the main text. We also ran an "unfiltered" version of Brumpton without the marginal p-value filter, using all SNPs that were significant in the regressions on unrelated individuals. This method returned 66 significant trait pairs. However, since most of the significant trait pairs became insignificant when filtering out SNPs with $p > 0.05$ in the trio data, it is possible that weak instrument bias partially explains many of these associations. The results for all three filter settings are given in Supplementary Table S2. All other methods were run with default parameters. Results are shown in Supplementary Table S2 and are discussed in the main text.

Probabilistic approach to generating digital siblings

As discussed in the Methods section, we evaluated two approaches for generating digital twins for sibling data. We found that the haplotype-shuffling approach (Methods) was much faster and controlled false positive rate better than the probabilistic approach (Figure S4). Because the probabilistic approach may still be useful as a starting point for further research, we describe it here.

First, we re-introduce the notation described in the Methods. 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. In the sibling setting, instead of having a single $(N \times M)$ genotype matrix $(\mathbf{X}_n)$, we have a vector of such matrices, one for each of the k siblings:

$$\mathbf{S} := ((\mathbf{X}_n)^1, (\mathbf{X}_n)^2, ..., (\mathbf{X}_n)^k) \tag{1}$$

We define $\mathbf{S}_{nj}$ as the vector of genotypes for all siblings in family n at SNP j:

$$\mathbf{S}_{nj} := (\mathbf{X}_{nj}^1, \mathbf{X}_{nj}^2, ..., \mathbf{X}_{nj}^k) \tag{2}$$

For each family n and SNP j, we want to infer

$$P(\mathbf{D}_{nj} \mid \mathbf{P1}_{nj}, \mathbf{P2}_{nj}) \tag{3}$$

where $\mathbf{P1}_{nj}$ and $\mathbf{P2}_{nj}$ are the parental genotypes for family n at SNP j. Because we do not observe the parental genotypes, we cannot condition on them. However, we can take advantage of the fact that

$$\mathbf{D} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{P1}, \mathbf{P2} \tag{4}$$

to manipulate this probability such that we do not need to condition on the parents, as follows:

$$\begin{aligned}
P(\mathbf{D}_{nj} \mid \mathbf{P1}_{nj}, \mathbf{P2}_{nj}) &= P(\mathbf{D}_{nj} \mid \mathbf{P1}_{nj}, \mathbf{P2}_{nj}, \mathbf{S}_{nj}) && (5)\\
&= P(\mathbf{D}_{nj}, \mathbf{P1}_{nj}, \mathbf{P2}_{nj} \mid \mathbf{S}_{nj}) \div P(\mathbf{P1}_{nj}, \mathbf{P2}_{nj} \mid \mathbf{S}_{nj}) && (6)\\
&= P(\mathbf{P1}_{nj}, \mathbf{P2}_{nj} \mid \mathbf{D}_{nj}, \mathbf{S}_{nj}) \times P(\mathbf{D}_{nj} \mid \mathbf{S}_{nj}) \div P(\mathbf{P1}_{nj}, \mathbf{P2}_{nj} \mid \mathbf{S}_{nj}) && (7)
\end{aligned}$$

All of these expressions can be inferred from the data. We first discuss $P(\mathbf{D}_{nj} \mid \mathbf{S}_{nj})$. To specify this quantity, we sum over the possible parental genotypes as follows:

$$\begin{aligned} P(\mathbf{D}_{nj}|\mathbf{S}_{nj}) &= \sum_{g_1,g_2} P(\mathbf{D}_{nj}, \mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2|\mathbf{S}_{nj}) && (8) \\ &= \sum_{g_1,g_2} P(\mathbf{D}_{nj}|\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2, \mathbf{S}_{nj}) P(\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2|\mathbf{S}_{nj}) && (9) \\ &= \sum_{g_1,g_2} P(\mathbf{D}_{nj}|\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2) P(\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2|\mathbf{S}_{nj}) && (10) \end{aligned}$$

where g_1 and g_2 are the three possible genotypes of $\mathbf{P1}_{nj}$ and $\mathbf{P2}_{nj}$ (nine possible pairs). The last step above follows from the fact that $\mathbf{D}$ is independent of $\mathbf{S}$ given that the parent genotypes are known. As stated in the previous subsection,

$$\mathbf{D}_{nj} \sim Bern(\mathbf{P1}_{nj}/2) + Bern(\mathbf{P2}_{nj}/2) \qquad (11)$$

so we need to compute $P(\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2|\mathbf{S}_{nj})$. We can compute this via Bayes' formula:

$$P(\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2|\mathbf{S}_{nj}) = \frac{P(\mathbf{S}_{nj}|\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2) P(\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2)}{P(\mathbf{S}_{nj})} \qquad (12)$$

We assume that all parental genotypes have equal prior probability, so

$$P(\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2|\mathbf{S}_{nj}) \propto P(\mathbf{S}_{nj}|\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2) \qquad (13)$$

We assume that there are no maternal twins among the siblings, so they are independent samples from the parents. Thus,

$$P(\mathbf{S}_{nj}|\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2) = \prod_k P(\mathbf{X}^k_{nj}|\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2) \qquad (14)$$

Similarly to before,

$$P(\mathbf{X}^q_{nj}|\mathbf{P1}_{nj} = g_1, \mathbf{P2}_{nj} = g_2) \sim Bern(\mathbf{P1}_{nj}/2) + Bern(\mathbf{P2}_{nj}/2) \qquad (15)$$

Finally, we note that the quantities $P(\mathbf{P1}_{nj}, \mathbf{P2}_{nj}|\mathbf{D}_{nj}, \mathbf{S}_{nj})$ and $P(\mathbf{P1}_{nj}, \mathbf{P2}_{nj}|\mathbf{S}_{nj})$ can be inferred in the same manner as discussed in Equations 12 through 15. This completes the model specification. We can use this model to simulate digital siblings from $P(\mathbf{D}_{nj}|\mathbf{S}_{nj})$.

References

[1] S. Burgess, A. Butterworth, and S. G. Thompson, “Mendelian randomization analysis with multiple genetic variants using summarized data,” *Genetic Epidemiology*, vol. 37, no. 7, pp. 658–665, 2013.

[2] J. Bowden, G. Davey Smith, and S. Burgess, “Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression,” *International Journal of Epidemiology*, vol. 44, no. 2, pp. 512–525, 2015.

[3] J. Bowden, G. Davey Smith, P. C. Haycock, and S. Burgess, “Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator,” *Genetic Epidemiology*, vol. 40, no. 4, pp. 304–314, 2016.

[4] F. P. Hartwig, G. Davey Smith, and J. Bowden, “Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption,” *International Journal of Epidemiology*, vol. 46, no. 6, pp. 1985–1998, 2017.

[5] O. O. Yavorska and S. Burgess, “Mendelianrandomization: an r package for performing mendelian randomization analyses using summarized data,” *International Journal of Epidemiology*, vol. 46, no. 6, pp. 1734–1739, 2017.

[6] J. R. Broadbent, C. N. Foley, A. J. Grant, A. M. Mason, J. R. Staley, and S. Burgess, “Mendelianrandomization v0.5.0: updates to an r package for performing mendelian randomization analyses using summarized data,” *Wellcome Open Research*, vol. 5, 2020.

[7] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, “Plink: a tool set for whole-genome association and population-based linkage analyses,” *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

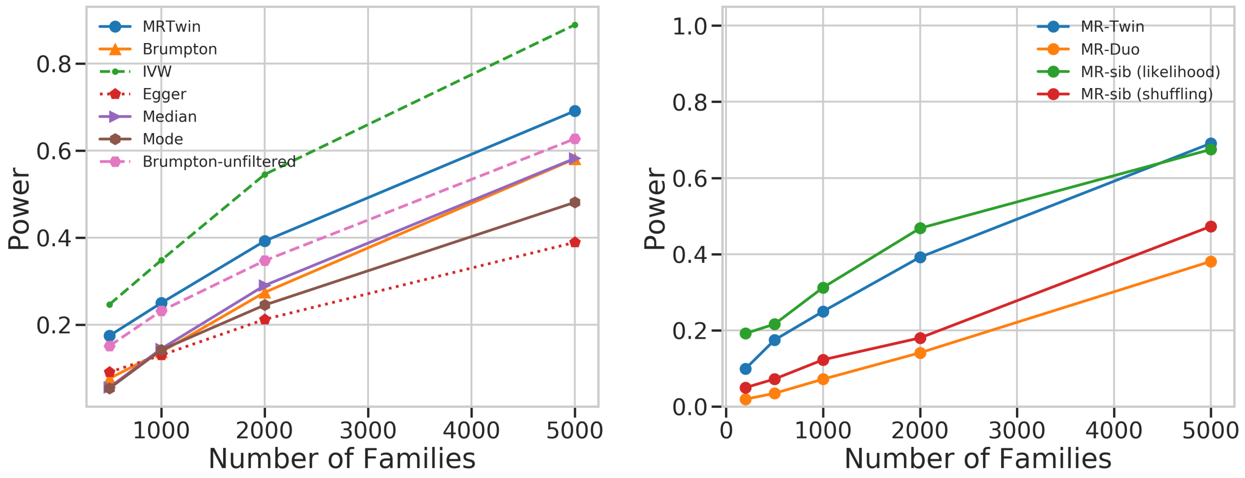
[8] B. Brumpton, E. Sanderson, K. Heilbron, F. P. Hartwig, S. Harrison, G. Å. Vie, Y. Cho, L. D. Howe, A. Hughes, D. I. Boomsma, *et al.*, “Avoiding dynastic, assortative mating, and population stratification biases in mendelian randomization through within-family analyses,” *Nature Communications*, vol. 11, no. 1, pp. 1–13, 2020.

[9] C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, *et al.*, "The uk biobank resource with deep phenotyping and genomic data," *Nature*, vol. 562, no. 7726, pp. 203–209, 2018.

[10] K. Hou, K. S. Burch, A. Majumdar, H. Shi, N. Mancuso, Y. Wu, S. Sankararaman, and B. Pasaniuc, "Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture," *Nature Genetics*, vol. 51, no. 8, pp. 1244–1251, 2019.

[11] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen, "Robust relationship inference in genome-wide association studies," *Bioinformatics*, vol. 26, no. 22, pp. 2867–2873, 2010.

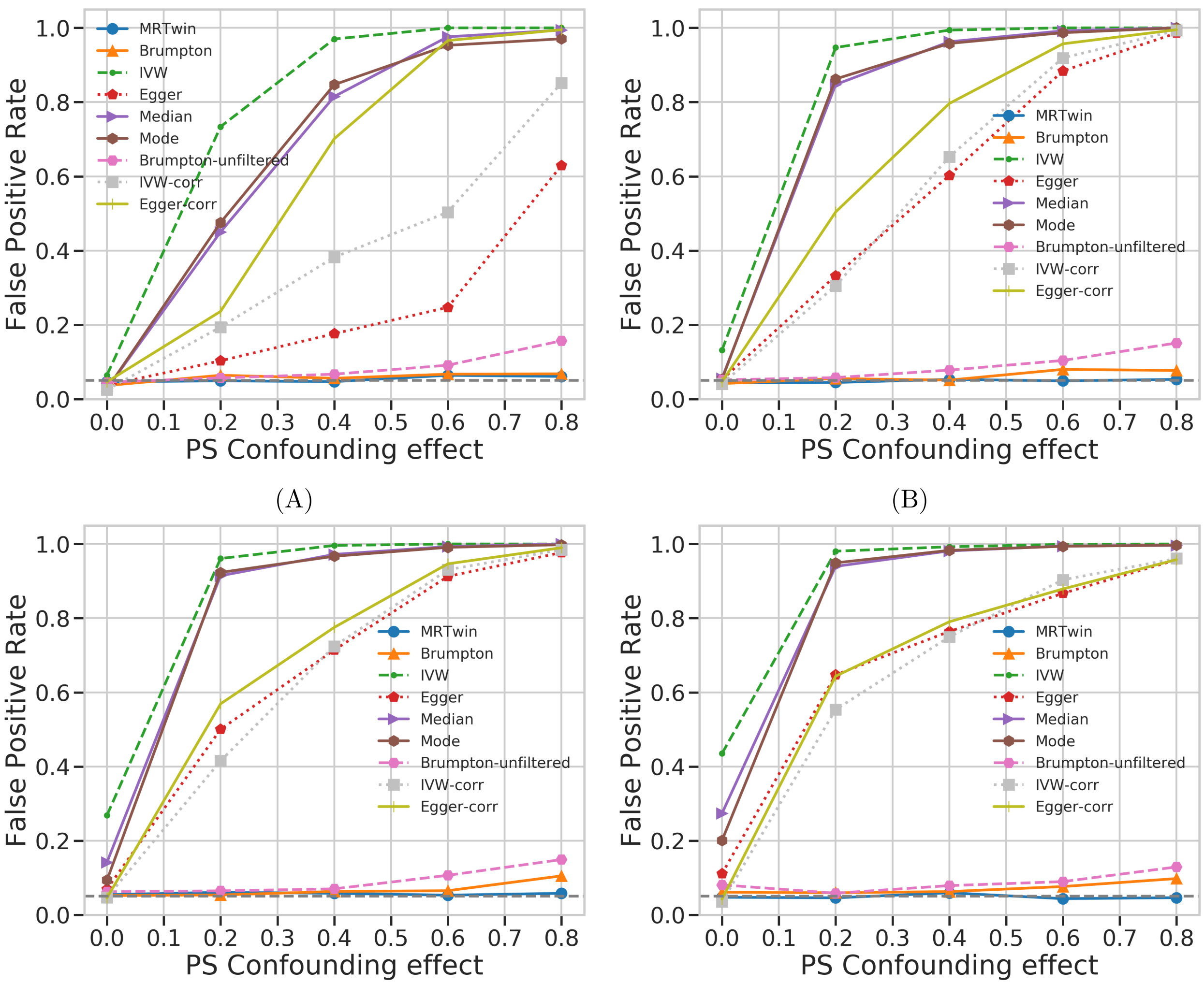
List of Figures



(A)

(B)

Figure S1: Power comparison between various methods with different numbers of simulated families. The axes are power (y-axis) and number of simulated families (x-axis), and the methods compared are (A) standard MR methods, Brumpton, and the trio mode of MR-Twin; (B) the trio, duo, and sibling modes of MR-Twin. "MR-sib (likelihood)" is a sibling-based approach where digital siblings are drawn based on a weighted average over the possible parental genotypes, while "MR-sib (shuffling)" generates the digital siblings by randomly shuffling the haplotypes of the true offspring. Results are averaged over 1000 simulation replicates.



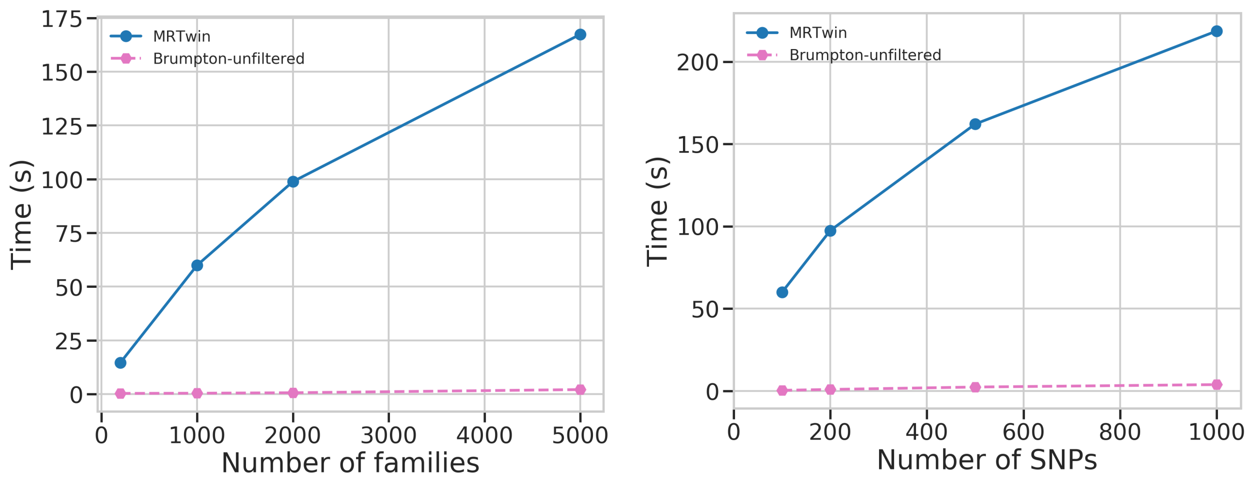
(A)

(B)

(C)

(D)

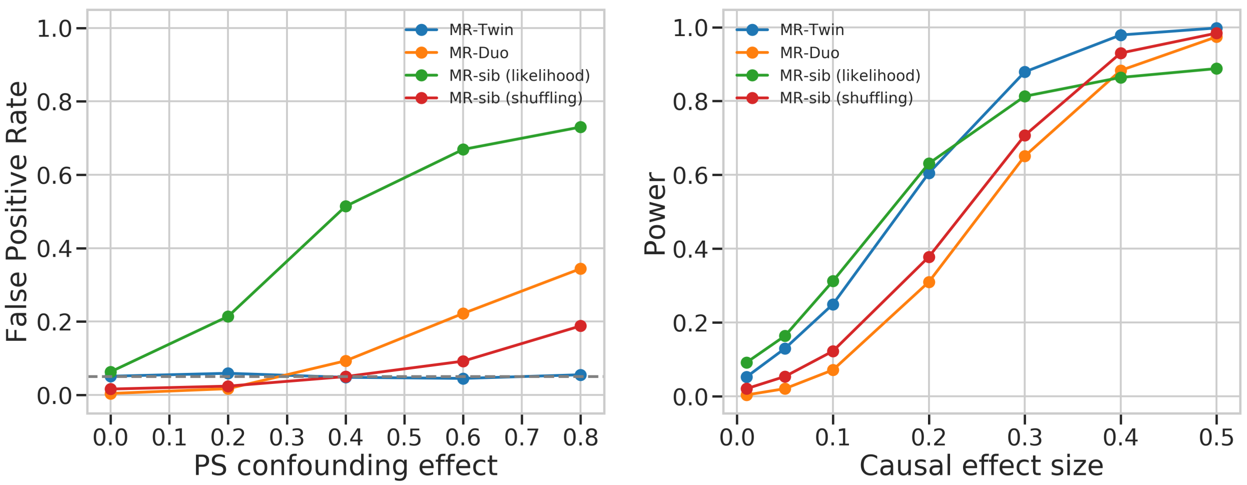
Figure S2: False Positive Rate (FPR) comparison on simulated data with different amounts of population structure. False positive rate (y-axis) under varying levels of confounding due to population stratification (PS), with the x-axis describing the magnitude of the effect of the effect of the population labels on the exposure and outcome trait. The subplots show results with the F_{ST} set to (A) 0.01; (B) 0.05; (C) 0.1; (D) 0.2. Results are averaged over 1000 simulation replicates.



(A)

(B)

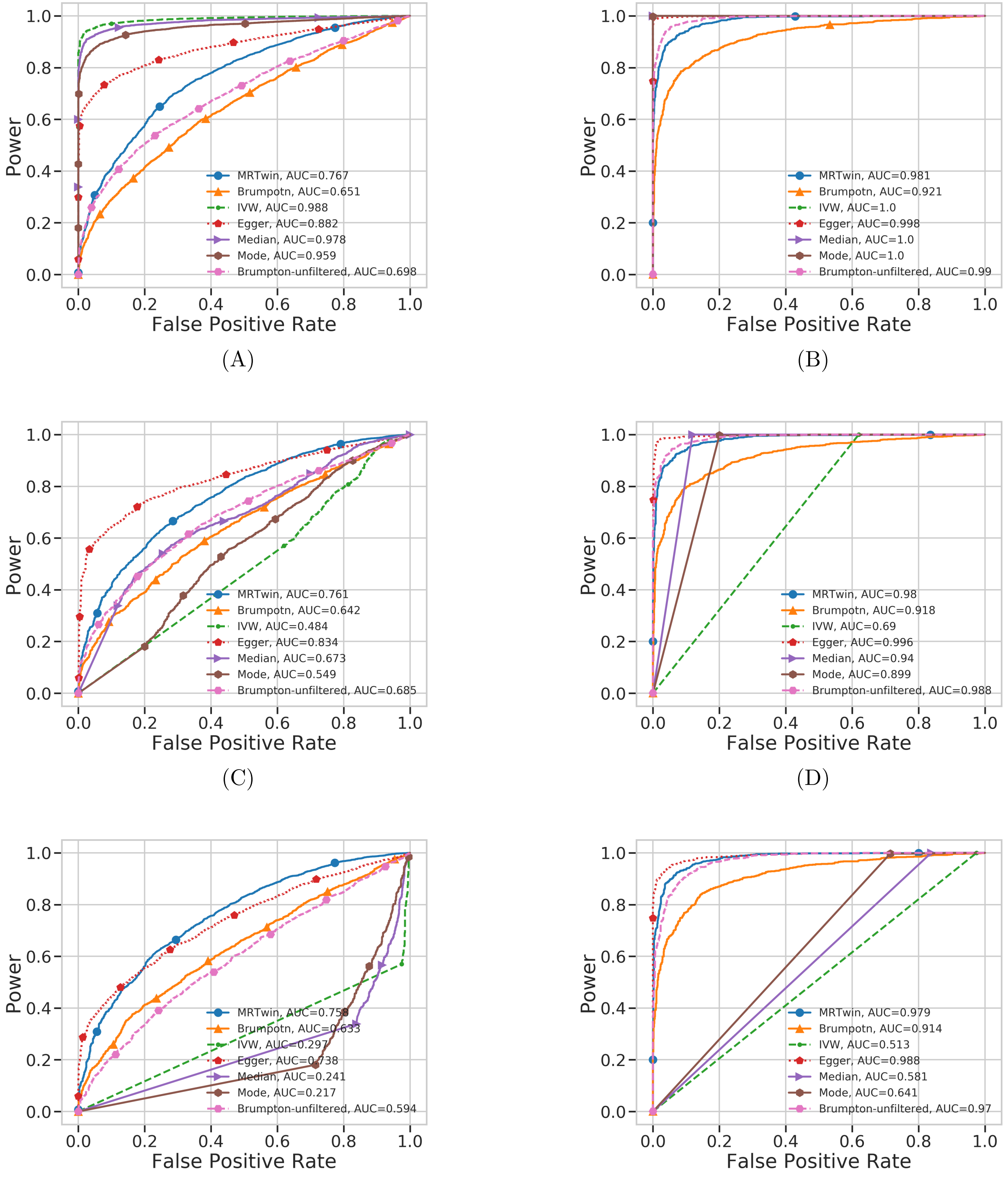
Figure S3: Time complexity analysis. Run time (y-axis) comparison between MR-Twin and Brumpton for (A) different numbers of families; (B) different numbers of SNPs (x-axis). Other MR methods had similar running time to Brumpton and are excluded for simplicity. Results are averaged over 10 simulation replicates. MR-Twin results use 100 digital twins. MR-Twin digital twins were simulated serially, not in parallel.



(A)

(B)

Figure S4: False Positive Rate (FPR) and Power comparison between the trio, duo, and sibling modes of MR-Twin run on simulated data. (A) False positive rate (y-axis) under varying levels of confounding due to population stratification, with the x-axis describing the magnitude of the effect between the population labels and the exposure and outcome trait. (B) Power (y-axis) with various causal effect sizes (x-axis). Results are averaged over 1000 simulation replicates.

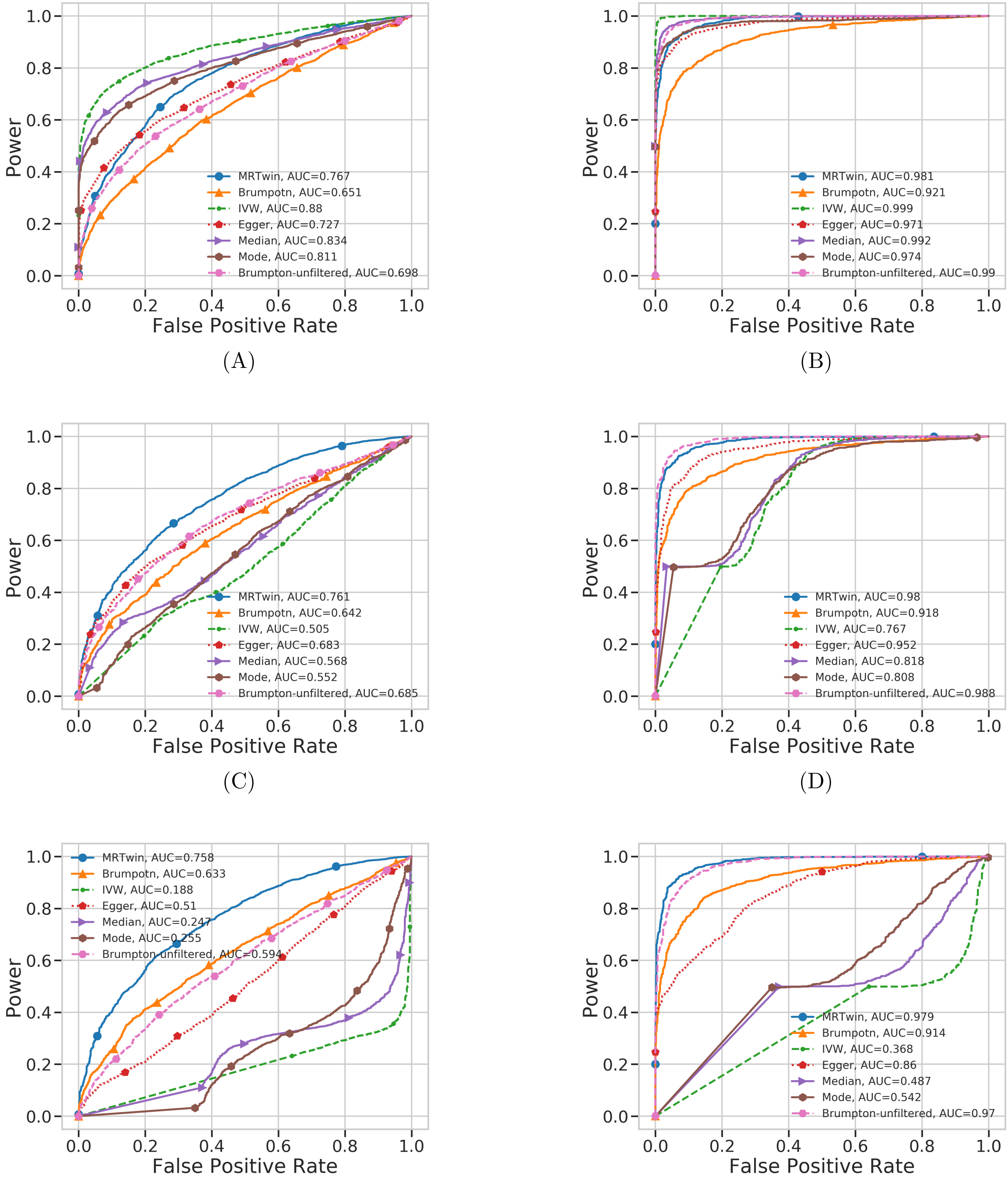


(A) (B)

(C) (D)

(E) (F)

Figure S5: Receiver Operating Characteristic (ROC) curve comparing various methods run on simulated trio data. From left to right, the causal effects of the power simulation are 0.1 and 0.3, respectively. From top to bottom, the confounding effects imposed in null settings are 0, 0.4, and 0.8, respectively. Both power and calibration settings have 1000 replicates in each setting, and there is no confounding effect in any of the power simulations.



(A)

(B)

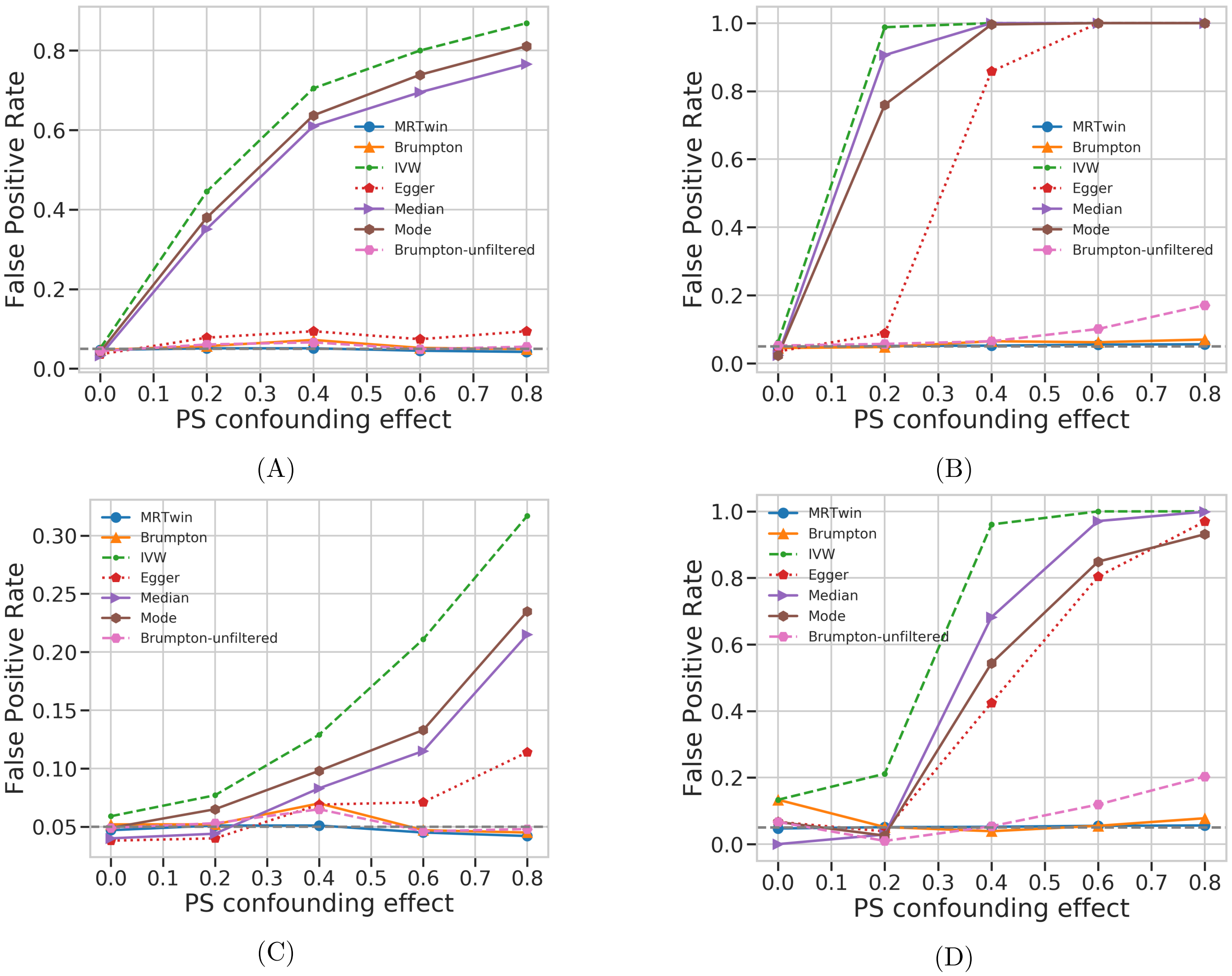
(C)

(D)

(E)

(F)

Figure S6: Receiver Operating Characteristic (ROC) curve comparing various methods run on simulated trio data. Similar to Fig S5, except that IVW, Egger, Median, and Mode are run on the offspring of the trio dataset instead of the large "external" group of unrelated individuals, such that all methods have the same sample size.



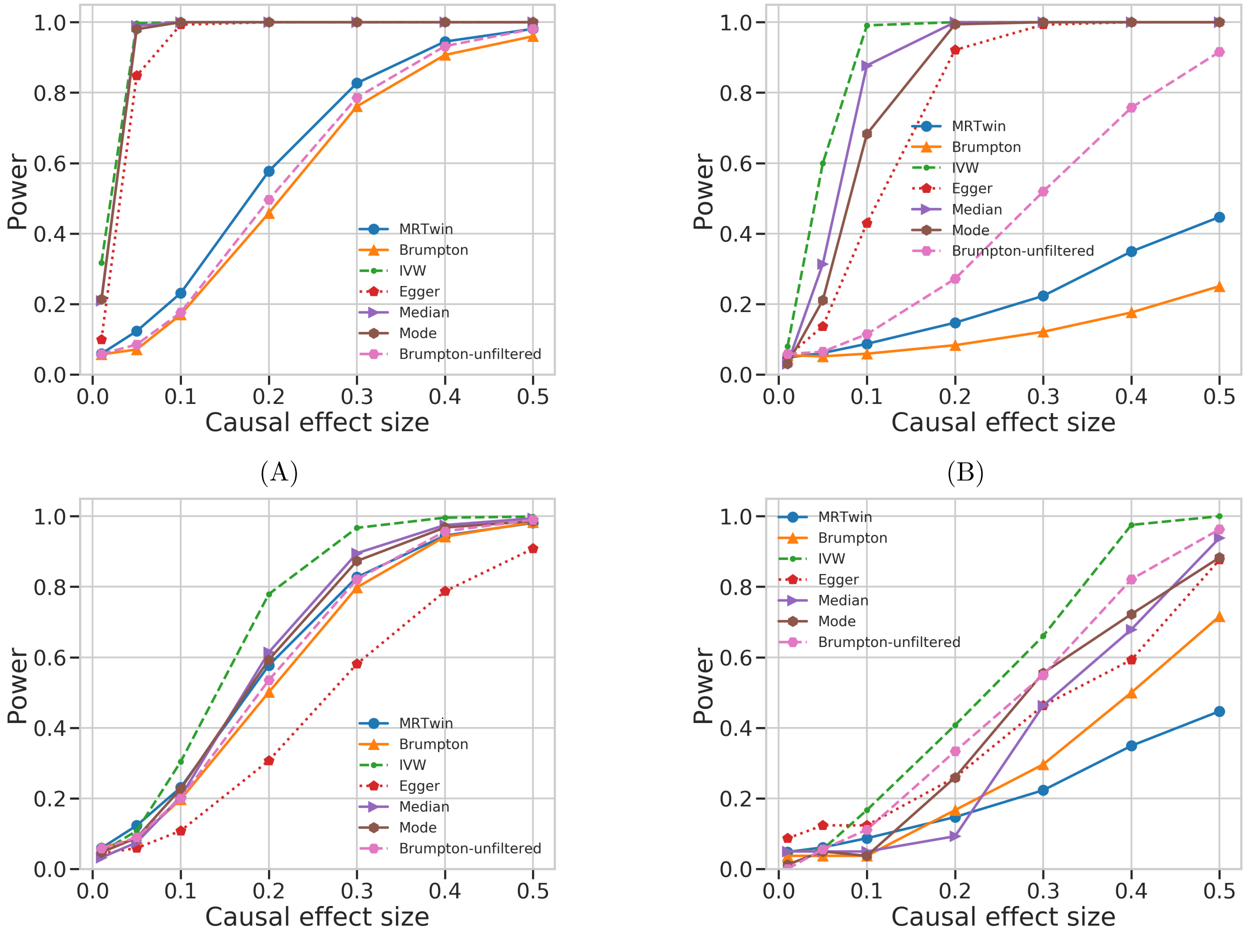
(A)

(B)

(C)

(D)

Figure S7: 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, run on simulated data in settings with very few SNPs or very low heritability. (A) and (B) are similar to Figure 2a, except that for (A), the total number of SNPs is 10 (8 of them are causal); for (B), the heritability $h^2 = 0.02$. (C) and (D) are similar to Figure 3a, except that for (C), the total number of SNPs is 10 (8 of them are causal); for (D), the heritability $h^2 = 0.02$.



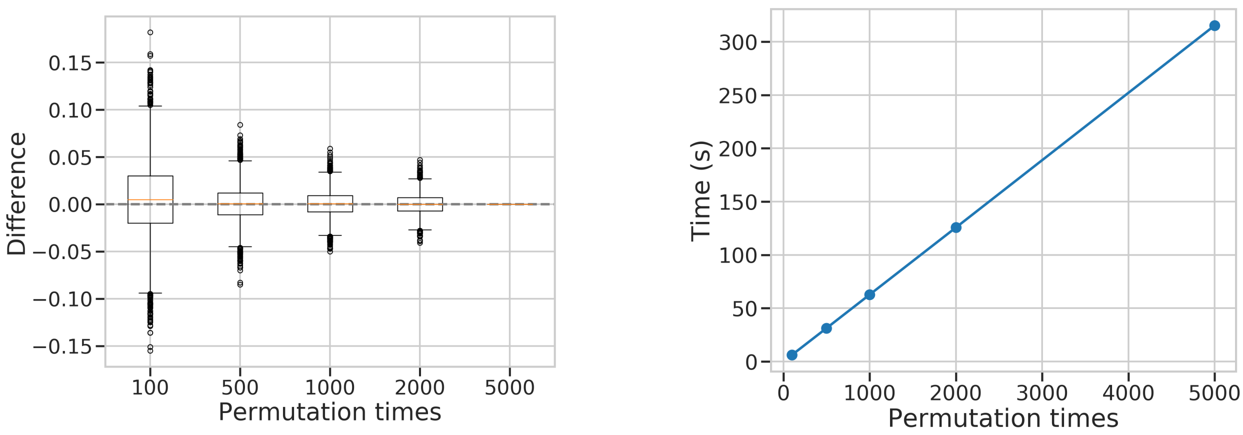
(A)

(B)

(C)

(D)

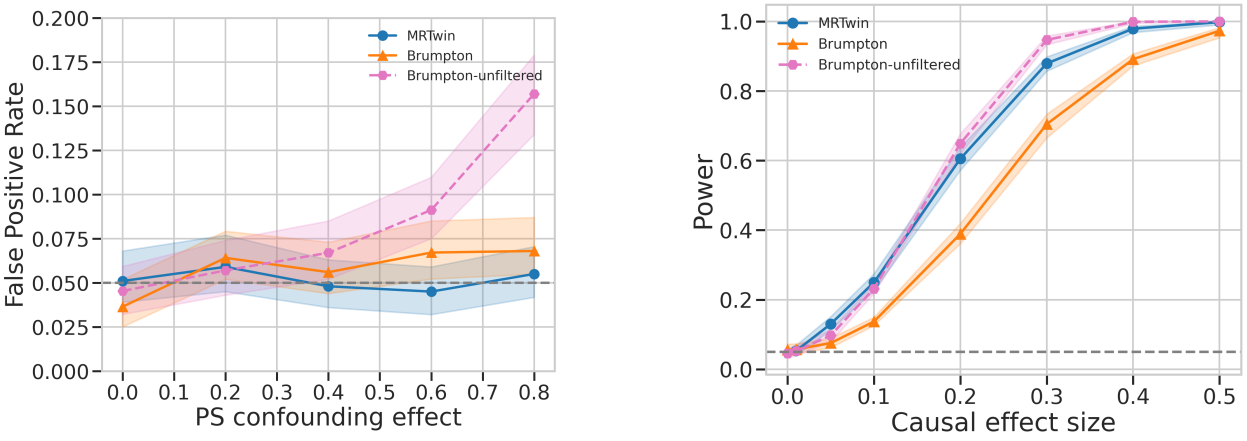
Figure S8: Power (y-axis) as a function of the magnitude of the causal effect of the exposure on the outcome trait (x-axis) in a setting with no confounding, run on simulated data in settings with very few SNPs or very low heritability. (A) and (B) are similar to Figure 2b, except that for (A), the total number of SNPs is 10 (8 of them are causal); for (B), the heritability $h^2 = 0.02$. (C) and (D) are similar to Figure 3b, except that for (C), the total number of SNPs is 10 (8 of them are causal); for (D), the heritability $h^2 = 0.02$.



(A)

(B)

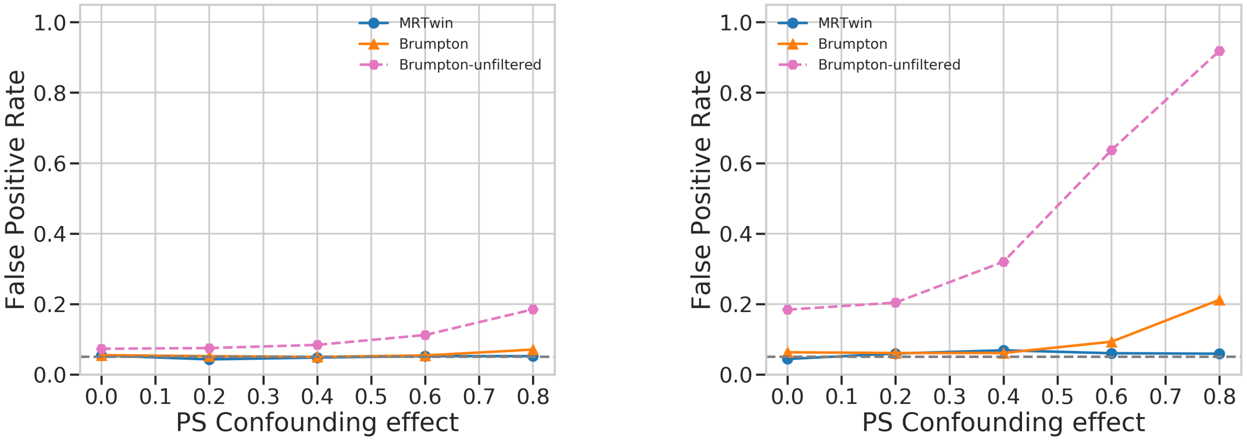
Figure S9: MR-Twin p-value stability and time complexity for different numbers of simulated digital twins. (A) A stable "baseline" MR-Twin p-value was computed on simulated data using a very large number of digital twins (5000). MR-Twin was then run 1000 times for several different numbers of digital twins (x-axis), and the difference between each run's p-value and the baseline was computed (y-axis). No confounding effect was simulated; parameter settings were otherwise the same as the False Positive Rate experiments. (B) Running time (y-axis) averaged over 10 runs of MR-Twin for different numbers of digital twins (x-axis).



(A)

(B)

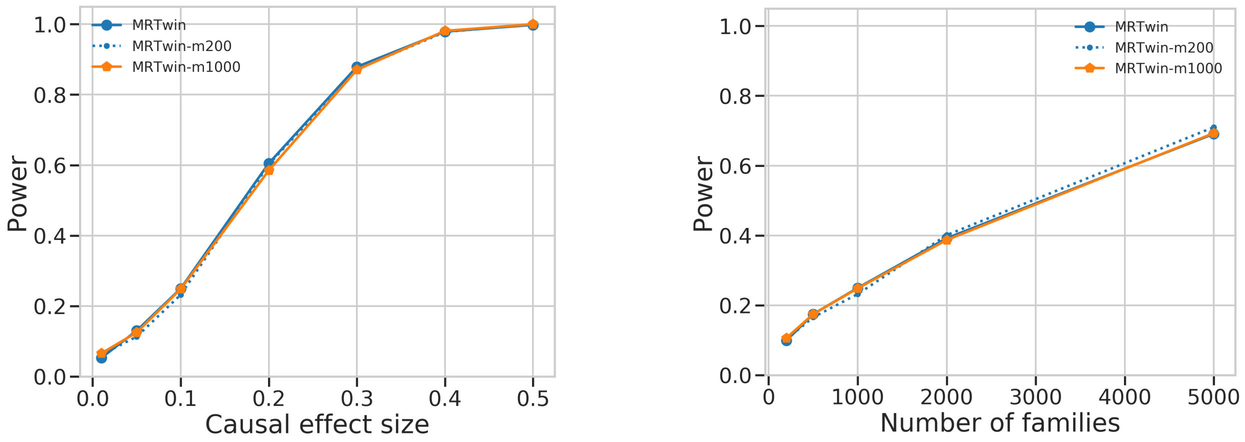
Figure S10: Calibration and power analysis of MR-Twin and Brumpton with confidence intervals. (A) False positive rate (y-axis) under varying levels of confounding due to population stratification (PS), with the x-axis describing the magnitude of the confounding effect of population labels on the exposure 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. The 95% confidence intervals are computed for MR-Twin and Brumpton by bootstrapping the original test statistics 1000 times with replacement.



(A)

(B)

Figure S11: False positive rate (FPR) comparison with no SNP filtering. 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 effect of the population labels on the exposure and outcome trait. Unlike the main text figures, no SNPs were filtered out via the external dataset, so many instruments are expected to be weak. The total number of SNPs are (A) 100; and (B) 1000, respectively.



(A)

(B)

Figure S12: Power analysis of MR-Twin with and without SNP filtering. Here *MRTwin* represents the standard usage of MR-Twin with the same p-value filtering procedure via the external data as described in the main text. *MRTwin-m200* and *MRTwin-m1000* represent running MR-Twin without any SNP filtering via the external data with 200 or 1000 SNPs, respectively, so that many instruments are expected to be weak.

List of Tables

Trait Name	UK Biobank Field ID	Num. of Genetic Instruments
(Heel) Bone Mineral Density	78	321
Body Mass Index	21001	260
Body Fat (Percentage)	23099	203
(Total) Cholesterol	30690	290
Diastolic Blood Pressure	4079	96
Glucose	30740	46
(Standing) Height	50	1502
LDL Cholesterol	30780	257
Systolic Blood Pressure	4080	110
Townsend Deprivation Index	189	0
Triglycerides	30870	307
Weight	21002	339

Table S1: List of traits used in the UK Biobank analysis, along with the number of genetic instruments (SNPs) identified according to the filtering procedures described in the text (genome-wide significance and linkage disequilibrium pruning).