Contribution of incorrect statistical methods to the excess of false-positive results in Mendelian randomization analyses

Paul M McKeigue [1] [2], Tim Old [1], Andrii Iakovliev [2], Buddhiprabha Erabadda [1], Helen M Colhoun [2], Athina Spiliopoulou [1] [2]

1 Usher Institute, College of Medicine and Veterinary Medicine, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland.

2 Institute of Genetics and Cancer, College of Medicine and Veterinary Medicine, University of Edinburgh, Western General Hospital Campus, Crewe Road, Edinburgh EH4 2XUC, Scotland.

Abstract

Recent commentaries have noted a "deluge" of studies that report support for causality using 2-sample Mendelian randomization (2SMR). To investigate reasons for this apparent excess of positive results, we sampled 40 published 2SMR studies. Of the 30 studies that reported support for causality, 27 used the weighted median test, 15 used MR-PRESSO, and 4 used MR-RAPS. In simulations from a null model based on plausible assumptions about the distribution of pleiotropic effects, all three of these methods showed inflation of the variance of the test statistic and the Type 1 error rate, most extreme for the weighted median test. A test based on marginalizing over the direct effects to compute the likelihood of the causal effect parameter had the lowest Type 1 error rate and the lowest Type 2 error rate. We conclude that the proliferation of 2SMR studies reporting evidence of causality is at least partly attributable to widespread use of incorrect statistical methods that are implemented in the MR-Base platform. Used with care, and with measures to control confounding of associations between instrument-exposure and instrument-outcome effects, 2SMR can support systematic causal inference.

Introduction

Instrumental variable analysis with genetic instruments ("Mendelian randomization") has been widely used to infer causal effects of exposures (broadly defined to include behavioural traits, biomarkers and gene expression levels) on diseases. The availability of summary statistics on genotype-exposure and genotype-outcome associations in different studies has made it possible to combine these data in "two-sample MR" (2SMR) studies for "systematic causal inference across the human phenome."[1] Early enthusiasm for this approach has given way to disenchantment, expressed in recent commentaries[2,3] that have noted the rapid growth in published studies using 2SMR that report support for causality:

> *we are unfortunately seeing an ever-increasing number of MR studies that simply use summary GWAS data, and lack negative controls or evidence from other study designs and methodologies to strengthen inference ... there are now relatively few studies applying MR methods that report null results.*

The authors advise reviewers and editors to "simply reject papers that only report 2SMR findings, with no additional supporting evidence,"[2] and recommend that studies should include tests for causal effects on "negative control" outcomes that are not plausibly associated with the exposure under study. This leaves unexplained why current methods for 2SMR analysis are apparently generating false-positive results. Statistical inference given observed data and a model that incorporates prior information is a well-posed problem: it has a unique solution (the posterior) and slight perturbation of the inputs will only slightly perturb this solution.[4] If there is not enough information to infer causality, inference should yield a flat posterior distribution (or flat likelihood) of the causal effect parameter; it should not yield false-positive results supporting causality.

This paper investigates whether defects in currently-used methods for 2SMR analysis underlie what has been described as a "deluge of papers and misleading findings."[2] The paper is organized as follows. First, we briefly describe the statistical model on which 2SMR is based. Second, we examine a sample of papers that report evidence for causality to determine what statistical methods these studies relied on. Third, we use simulation under plausible assumptions about the distribution of nuisance parameters to evaluate the Type 1 and Type 2 error rates of these methods. We comment on why some of these methods yield inflated Type 1 error rates, and conclude with some recommendations for the conduct of 2SMR studies that differ from existing guidelines.

Methods

The main methodological challenge in MR analysis is to infer causality when some of the genetic instruments have direct (pleiotropic) effects on the outcome that are not mediated through the exposure under study. These direct effects are not observed, and their distribution over the instruments is unknown. With modern probabilistic programming languages such as BUGS/JAGS, Stan, PyMC or NumPyro it is straightforward to specify a statistical model with priors on all model parameters including direct effects of the instruments and the causal effect of the exposure on the outcome, and to compute the joint posterior distribution of these parameters, given the data. The marginal likelihood of the causal effect parameter can be obtained by dividing the marginal posterior density by the prior on this parameter. Classical hypothesis tests can be obtained from this likelihood function.

The most widely-used methods for MR analysis, however, are based not on calculating the likelihood function but on constructing "estimators" of the causal effect parameter and demonstrating that these estimators have desirable sampling properties – consistency, unbiasedness, and minimum variance – over hypothetical repetitions of the experiment. As

these frequentist methods are not grounded in likelihood theory, it may not be obvious whether they are likely to generate misleading results.

Statistical model

For a Mendelian randomization study with J unlinked genetic instruments, we specify a model with three parameters:

- $\boldsymbol{\alpha}$ vector of coefficients of effects of the instruments on exposure X.
- $\boldsymbol{\beta}$ vector of coefficients of direct (pleiotropic) effects of the instruments on outcome Y
- θ causal effect of X on Y

The crude effect of the jth instrument on the outcome is the sum of the direct effect and the causal effect.

$$\gamma_j = \beta_j + \theta\alpha_j$$

For instruments with no direct effects, $\beta_j = 0$ and $\theta = \gamma_j/\alpha_j$ and a scatter plot of the true coefficients γ_j against α_j will give points lying on a straight line with gradient θ passing through the origin. For instruments with nonzero direct effects the points will lie off this line, as shown in Fig 1. In a scatter plot of the coefficient estimates $\hat{\gamma}_j$ against $\hat{\alpha}_j$, sampling error will contribute additional dispersion about this line. In testing for causality, we are effectively testing for coupling of the instrument-exposure effects α_j and the instrument-outcome effects γ_j. We discuss later other possible explanations for this coupling.

We note in passing that the statistical power of a 2SMR study depends on the number J of unlinked instruments, because the number of instruments is effectively the number of observations. For type 1 error rate a and type 2 error rate b, the number of observations required to detect an effect of size θ is

$$\frac{\left(\Phi^{-1}\left(1-a/2\right)+\Phi^{-1}\left(1-b\right)\right)^2}{I\theta^2}$$

where $\Phi()$ is the standard normal cumulative distribution function and I is the Fisher information about θ. We can calculate the statistical power for a 2SMR study under the simplifying assumptions that the instrument-exposure and instrument-outcome coefficients are estimated from large GWAS datasets so that their sampling errors are small, and the direct effects β_j have a Gaussian distribution with scale σ. The Fisher information I about θ contributed by the jth instrument is α_j^2/σ^2: the square of the ratio of the instrument-exposure effect to the scale of the direct effects. With exposure and outcome scaled to unit variance, a plausible value for the squared ratio α_j^2/σ^2 is 4, based on our experience of using *trans*-pQTLs as instruments for plasma proteins. This implies that for $\theta = 0.5$, 21 instruments are required for 90% power to detect an effect at $p < 0.001$.

In practice we expect the distribution of the direct effects β_j to be more heavy-tailed than a Gaussian. A convenient prior distribution for β_j is the horseshoe prior, which resembles a spike-and-slab mixture.[5] The observations are the coefficient estimates $\hat{\alpha}_j, \hat{\gamma}_j$, with their standard errors. We model these coefficient estimates as Gaussian variables with means equal to the true values γ_j, α_j and standard deviations equal to the standard errors of the estimates. Where the coefficient estimates are based on large GWAS studies, their standard errors will be small and thus the posterior distribution will not be sensitive to the priors on α_j. Any reasonable prior can be specified for the causal effect parameter θ_j as we will divide the posterior by the prior to obtain the marginal likelihood. Given priors on β_j, α_j and θ, the maximum likelihood estimates $\hat{\alpha}_j, \hat{\gamma}_j$ and their standard errors, we can compute the posterior distribution of θ using a probabilistic programming language.

The method used here (described in the accompanying paper) differs from that used in the MR-HORSE package[6] in that it uses a regularized horseshoe prior rather than the original horseshoe prior, and it obtains a classical hypothesis test from the marginal likelihood of the causal effect parameter θ as described below.

Simulation study

We examined the Type 1 error rates of the most widely-used methods in a simulation study. Without loss of generality, we assume that the instruments and the outcome have been scaled to a standard deviation of 1. For each simulated dataset with $J = 10$ unlinked instruments, we generate data as follows:

- the number of instruments that have nonzero direct effects are drawn from a binomial distribution with probability 0.4 and J trials.
- the direct effects β_j are drawn from a Student t distribution with zero mean, 4 degrees of freedom and scale 0.05
- the instrument-exposure effects α_j are drawn from a Gaussian distribution with zero mean and standard deviation 0.1.
- the instrument-outcome effects γ_j are defined as $\gamma_j = \beta_j + \theta\alpha_j$.
- the coefficient estimates $\hat{\alpha}_j, \hat{\gamma}_j$ are drawn from Gaussian distributions with means α_j, γ_j. The standard deviations (equivalent to the standard errors of the estimates) of these Gaussian distributions are drawn from a Gamma distribution with mean 0.005 and standard deviation 0.0016 (encoded as shape 10 and scale 0.0005). This precision for the coefficient estimates implies an effective sample size of 40,000 for the studies from which they were estimated.

Without loss of generality, we flip the signs of the instruments so that all $\hat{\alpha}_j$ estimates are positive: for each instrument where the sign of $\hat{\alpha}_j$ is flipped, the sign of $\hat{\gamma}_j$ is also flipped.

Calculation of test statistics by different methods.

As a reference method, we used a test based on the marginal likelihood of the causal effect parameter as described above. The computational implementation is described in the accompanying paper. We compared other widely-used methods with this reference method: the fixed-effects and random-effects inverse-variance weighted (IVW) tests, the "raw" and "outlier-corrected" MR-PRESSO tests,[7] the weighted median test,[8] and the MR-RAPS test.[9] For the "inverse-variance weighted" (IVW) and weighted median estimators, we calculated the estimated coefficient ratios $\hat{\gamma}_j/\hat{\alpha}_j$ and their standard errors. Following recommended practice, the delta method (second-order Taylor expansions for the moments of the distribution of the ratio of two independent Gaussian variables) was not used to correct these estimates. The fixed-effect IVW test calculates the sampling variance of the estimator from these standard errors. For the weighted median estimator, we used the procedure described previously to calculate the standard error of the weighted median.[8] For comparison, we also calculated the standard error of the weighted median using the posterior predictive distribution of the direct effects given the instrument-exposure coefficients and their standard errors. The MR-PRESSO tests[7] were calculated from the function `mr_presso()` with argument `SignifThreshold=0.001`. The "outlier-corrected" MR-PRESSO test excludes instruments that are detected as outliers. The MR-RAPS test[9] was calculated with the default settings.

Results

Sampling of published papers

A search of PubMed was performed on 18 October 2024. The query `[("MR-Base" OR "MR Base" OR "MRBase" OR "weighted median") AND "mendelian randomization"]` retrieved 2629 papers. Additional citation searches identified 3174 papers that cited the original exposition of the weighted median estimator,[8] 308 that cited the R package `TwoSampleMR`,[10] 59 that cited the R package `MendelianRandomization`,[11] and 2695 that cited the paper describing the MR-Base platform.[1] After merging and deduplicating these search results, a list of 6311 papers remained.

A random sample of 40 papers from this merged list was examined. These papers were scored for number of unlinked genetic instruments, reported support for causality, and which statistical tests were used; inverse-variance weighted (IVW) mean of ratio estimates, IVW random effects,[12] weighted median, pleiotropy residual sum and outlier test (MR-PRESSO)[7] with or without outlier correction, and Robust Adjusted Profile Likelihood (MR-RAPS).[9]

Of the 40 papers sampled, two were excluded as reviews or methodology papers not reporting original results. Of the 30 papers that reported support for causality, 25 used the fixed effect IVW test, 27 used the weighted median test, 15 used MR-PRESSO, and 4 used MR-RAPS. 8 papers used the random-effect IVW test. Of the 30 papers that reported support for causality, 22 used $p < 0.05$ as a threshold for declaring support.

Simulation under the null

Table 2 compares the variance of the test statistic, and the Type 1 error rate at a threshold of $p < 0.001$ for the various tests. As expected, the variance of the test statistic and the Type 1 error rate were highest for the IVW test, which assumes no direct effects. However the variance of the test statistic was markedly inflated also for tests that were intended to be robust in the presence of pleiotropic effects: 9-fold for the weighted median test, 7-fold for the MR-PRESSO outlier-corrected test, and 4-fold for the MR-RAPS test. The corresponding ratios of observed to nominal Type 1 error rates were 136, 91 and 24. With the MR-PRESSO raw test and the random-effects IVW test, the variance of the test statistic was only moderately inflated (1.5 fold). With the marginal likelihood test, the variance of the test statistic was less than 1 and there were no Type 1 errors in 500 simulations. When the standard error of the weighted median was calculated from the posterior predictive distribution of the direct effects, the variance of the test statistic was less than 1.

Simulation under a non-null value of the causal effect parameter

Table 2 compares the Type 2 error rates at a threshold of $p < 0.001$ in simulations from a model with causal effect size 0.25, for the three methods that achieved reasonable control of the Type 1 error rate in simulations under the null: the random-effects IVW test, the MR-PRESSO raw test, and the marginal likelihood test. The marginal likelihood test had the lowest Type 2 error rate (and thus the highest statistical power) and the smallest mean squared error of the parameter estimate.

Discussion

We have shown that almost all published 2SMR studies that report evidence of causality rely on tests that give highly inflated Type 1 error rates in simulations based on plausible settings for the size and frequency of direct effects of instruments on outcome. The most widely-used test is the weighted median test, with Type 1 error rate inflated more than 100-fold in our simulations. This suggests that the "deluge of papers and misleading findings" may be at least

partly attributable to the use of statistical methods that fail to control the Type 1 error rate. Inflated Type 1 error rates with the weighted median and MR-PRESSO tests have been noted previously,[13] but as these tests are provided on the MR-Base platform researchers have continued to use them. From the sampling of published papers above, we estimate that more than 4000 published papers have relied on incorrect methods to infer support for causality.

We briefly discuss why the weighted median, outlier-corrected MR-PRESSO, and MR-RAPS methods have such high Type 1 error rates. The weighted median test uses a weighted median estimator of the coefficient ratios. The sampling distribution of this estimator is calculated by what is described as a "parametric bootstrap" method.[8] Parametric bootstrapping models the observed data as generated by a known distribution or family of distributions, estimates the parameters of that distribution from the data, and uses this predictive distribution to draw new observations under the null. The method described in the original publication does not sample new observations (instruments) from the posterior predictive distribution but simulates new estimates of the instrument-exposure coefficients for the same instruments. Not surprisingly this procedure yields a standard error for the weighted median that is too small. This incorrect procedure is replicated in the R package `TwoSampleMR`,[10] the R package `MendelianRandomization`,[11,14] and in the `MR-Base` platform.[1] It is possible to calculate the sampling distribution of the weighted median by generating new instruments and new observations of instrument-exposure coefficients from the posterior predictive distribution, but there is no advantage to this if we have computed the posterior distribution of all model parameters as the marginal likelihood of the parameter of interest can be calculated directly.

The random-effects IVW test estimates the factor by which the direct effects inflate the residual variance, and uses this to rescale the sampling variance of the IVW estimator.[12] The "raw" MR-PRESSO test is similar to the random-effects IVW test, using a weighted regression model without intercept term to estimate the variance of the direct effects. The "outlier-corrected" MR-PRESSO test however re-estimates the variance of the direct effects after excluding outliers. Not surprisingly, dropping outliers gives a standard error for the causal effect parameter that is too small. The defect in the MR-RAPS method is less obvious. MR-RAPS fits a model under the null hypothesis of no causal effect, with direct effects held at their maximum likelihood values.[9] A score test is constructed based on the gradient and second derivative of the profile likelihood (the likelihood function with nuisance parameters held at their maximum likelihood values) of the causal effect parameter at the null. Under some regularity conditions profile likelihoods behave similarly to marginal likelihoods, but these conditions may not hold when the number of nuisance parameters equals the number of observations.[15] Again there is no advantage to using the profile likelihood when the marginal likelihood can easily be calculated.

Why causal inference should be based on the likelihood

Bayesian hypothesis testing is based on the Likelihood Principle: all information conveyed by observations that supports one model or one parameter value over another is contained in the ratio of the likelihoods of these models or parameter values, given the data. The likelihood-based approach to causal inference described in this paper is aligned with the not yet universally-accepted principle that causal inference is just a special case of statistical inference, requiring attention to assumptions about exchangeability between observations and predictions of the effect of perturbing the exposure.[16] Where the log-likelihood is asymptotically quadratic, statistical theory guarantees that the maximum-likelihood estimate has the sampling properties that are desirable for an estimator: consistency, minimum variance and unbiasedness. Where the log-likelihood is not approximately quadratic, it is preferable to base inference directly on the likelihood (or the posterior) rather than on the sampling properties of an estimator.

Other explanations for coupling of instrument-exposure and instrument-outcome effects

The standard statistical model for 2SMR assumes that the effects of the instruments on the exposure and the direct effects of the instruments on the outcome are independent in magnitude and direction. The causal effect parameter can then be inferred from the relation between the instrument-exposure coefficients and the instrument-outcome coefficients. The assumption that the instrument-exposure effects and the direct effects are uncoupled in direction is termed "balanced pleiotropy". It is possible in principle to infer a causal effect if the instrument-exposure effects and the direct effects are assumed to be uncoupled in magnitude even if they are coupled in direction[17] but it is not obvious why this would be a biologically realistic assumption. The assumption that the magnitudes of the instrument-exposure effects and the direct instrument-outcome effects are independent has been denoted the InSIDE (Instrument Strength Independent of Direct Effect) assumption.[18]

If the assumption of no coupling between instrument-exposure effects and direct instrument-outcome effects is relaxed to allow coupling in magnitude and direction by an unmeasured confounder, it is not possible to distinguish between causality and confounding without a strong prior on the strength of coupling, because models with and without a causal effect can fit the data equally well with the same number of adjustable parameters. The MR-HORSE model allows the instrument-exposure effects and the instrument-outcome direct effects to be correlated, but can infer causality in this situation only by specifying a strong prior on these correlations such that most of the prior mass is close to zero.[9]

One proposed approach to this problem is to infer the genetic correlations of the exposure and the outcome with a latent causal factor that gives rise to genetic covariance between the exposure and the outcome.[19] This however does not establish causality; an exposure that is strongly correlated with a latent variable may simply be a good biomarker for that variable. More fundamentally, if coupling of instrument-exposure coefficients and instrument-outcome coefficients is attributable to confounding by a shared pathway that affects the outcome, we would usually want to investigate this pathway further. For instance the effects of genetic variants on expression of interferon-stimulated genes are strongly coupled with their effects on systemic lupus erythematosus.[20] This upregulation of interferon-stimulated genes is recognizable as an "interferon signature" pointing to the interferon signalling pathway itself, rather than the genes regulated by this pathway, as a target for therapeutic intervention.[21]

Other mechanisms that give rise to coupling of instrument-exposure coefficients and instrument-outcome coefficients is attributable to confounding may be less interesting but relatively easy to control by standard epidemiological methods, provided that individual-level data are available. Reverse causation can be excluded by imposing temporal sequence between instrument-exposure effects and instrument-outcome effects: thus where the outcome is a common disease such as type 2 diabetes that has onset in older adults, we can estimate the instrument-exposure coefficients in younger adults before the disease has begun to develop. Where the outcome is a common condition such as depression, we can estimate the instrument-exposure coefficients in those who are free of the condition. Where the exposure is the transcript level of a gene in whole blood, the relation between the instrument-exposure coefficients and the instrument-outcome coefficients may be confounded by cell type proportions in whole blood. Given transcript levels measured in a single-cell reference panel, we can impute the cell type proportions from bulk transcriptomic measurements, and recalculate the instrument-exposure coefficients with adjustment for cell type proportions.

Revised guidelines for MR analysis

This is not the first time that mis-application of methods in complex trait genetics has led to an excess of positive findings;[22] with the advent of GWASs, these problems were resolved by

quality control procedures to detect and reduce inflation of the variance of the test statistic, and stringent p-value thresholds for declaring association. Guidelines for the conduct of Mendelian randomization studies have recently been updated.[23,24] Based on the work described here, some suggestions for revisions to these guidelines are listed below.

- Investigators should not "pick a sensible range of methods", but should use a correct statistical method based on computing the likelihood or the posterior distribution of the causal effect parameter given the model and the data, and should assess the sensitivity of their results to prior assumptions encoded in the model.
- Although current guidelines recommend the random-effects IVW test as the primary analysis method for 2SMR studies using summary data,[24] this recommendation has not been widely taken up. The random-effects IVW test achieves reasonable control of the Type 1 error rate, but has lower statistical power than the likelihood-based method when the direct effects have a heavy-tailed distribution.
- There is no need to exclude weak instruments, as likelihood-based methods do not rely on assumptions about the sampling properties of coefficient ratios. To maximize the strength of the instruments given available summary statistics on SNP-exposure coefficients, scalar instruments should be constructed from multiple SNPs as described in the accompanying paper.
- For adequate statistical power to detect a causal effect in the presence of direct effects that are not much smaller than the causal effect, the number of unlinked genetic instruments should be at least 20.
- For declaring evidence of causality in a discovery study (rather than confirming a hypothesis supported by other evidence), p-value thresholds should be considerably more stringent than $p < 0.05$.
- Where possible, multiple exposures should be studied so that pleiotropic effects of genetic instruments can be observed directly. Statistical methods to model multiple exposures and to exploit the information that this provides are described in the accompanying paper.
- If correct statistical methods are used it should not be necessary to include as a sanity check "negative control" outcomes that are not plausibly associated with the exposures under study. It is of course valuable to model the causal effects of the exposure on multiple outcomes, as this may help to identify shared pathways that couple instrument-exposure and instrument-outcome coefficients.
- Where evidence of a causal effect is detected, other explanations for coupling between instrument-exposure and instrument-outcome coefficients – reverse causation or confounding by an unmeasured factor – should be considered. Where such explanations are plausible, they can be tested by recomputing the instrument-exposure coefficients using restriction by age to control temporal sequence, or using adjustment to control measured confounders. A shared pathway that couples instrument-exposure and instrument-outcome coefficients may itself be relevant as a possible therapeutic target.
- Individual-level data will usually be required to construct scalar instruments from multiple SNPs, and to exclude confounding or reverse causation. This limits the usefulness of platforms such as MR-Base that provide only summary-level data.
- Used correctly, 2-sample Mendelian randomization can allow "systematic causal inference", even without other supporting evidence.

Declarations

Data and code availability

Code is available at https://github.com/molepi-precmed/mrhevo.

Acknowledgements

No specific funding was received for this work. AI was supported by the Medical Research Council Cross Disciplinary Fellowship (XDF) Programme (MC_FE_00035). HC is supported by an endowed chair from the AXA Research Foundation.

Declaration of interests

The authors declare no competing interests.

Web resources

References

1. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., 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. https://doi.org/10.7554/eLife.34408.

2. Stender, S., Gellert-Kristensen, H., and Smith, G.D. (2024). Reclaiming mendelian randomization from the deluge of papers and misleading findings. Lipids in Health and Disease *23*, 286. https://doi.org/10.1186/s12944-024-02284-w.

3. Munafò, M.R., Brown, J., Hefler, M., and Davey Smith, G. (2024). Managing the exponential growth of Mendelian randomization studies. Addiction (Abingdon, England). https://doi.org/10.1111/add.16627.

4. Jaynes, E.T. (1973). The well-posed problem. Foundations of Physics *3*, 477–492. https://doi.org/10.1007/BF00709116.

5. Carvalho, C.M., Polson, N.G., and Scott, J.G. (2010). The horseshoe estimator for sparse signals. Biometrika *97*, 465–480.

6. Grant, A.J., and Burgess, S. (2024). A Bayesian approach to Mendelian randomization using summary statistics in the univariable and multivariable settings with correlated pleiotropy. American Journal of Human Genetics *111*, 165–180. https://doi.org/10.1016/j.ajhg.2023.12.002.

7. Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature Genetics *50*, 693–698. https://doi.org/10.1038/s41588-018-0099-7.

8. Bowden, J., Davey Smith, G., Haycock, P.C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic Epidemiology *40*, 304–314. https://doi.org/10.1002/gepi.21965.

9. Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D.S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. The Annals of Statistics *48*, 1742–1769. https://doi.org/10.1214/19-AOS1866.

10. Hartwig, F.P., Davies, N.M., Hemani, G., and Davey Smith, G. (2016). Two-sample Mendelian randomization: Avoiding the downsides of a powerful, widely applicable but

potentially fallible technique. International Journal of Epidemiology *45*, 1717–1726. https://doi.org/10.1093/ije/dyx028.

11. Yavorska, O.O., and Burgess, S. (2017). MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. International Journal of Epidemiology *46*, 1734–1739. https://doi.org/10.1093/ije/dyx034.

12. Bowden, J., Del Greco M, F., Minelli, C., Smith, G.D., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in twosample summary data Mendelian randomization. Statistics in Medicine *36*, 1783. https://doi.org/10.1002/sim.7221.

13. Burgess, S., Foley, C.N., Allara, E., Staley, J.R., and Howson, J.M.M. (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nature Communications *11*, 376. https://doi.org/10.1038/s41467-019-14156-4.

14. Patel, A., Ye, T., Xue, H., Lin, Z., Xu, S., Woolf, B., Mason, A.M., and Burgess, S. (2023). MendelianRandomization v0.9.0: Updates to an R package for performing Mendelian randomization analyses using summarized data. Wellcome Open Research *8*, 449. https://doi.org/10.12688/wellcomeopenres.19995.1.

15. McCullagh, P., and Nelder, J.A. (1989). Generalized Linear Models 2nd ed. (Chapman and Hall) https://doi.org/10.1201/9780203753736.

16. Rohde, D. (2022). Causal Inference, is just Inference: A beautifully simple idea that not everyone accepts. In I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021 (PMLR), pp. 75–79.

17. Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. International Journal of Epidemiology *44*, 512–525. https://doi.org/10.1093/ije/dyv080.

18. Bowden, J., Burgess, S., and Smith, G.D. (2017). Difficulties in Testing the Instrument Strength Independent of Direct Effect Assumption in Mendelian Randomization. JAMA Cardiology *2*, 929–930. https://doi.org/10.1001/jamacardio.2017.1572.

19. O'Connor, L.J., and Price, A.L. (2018). Distinguishing genetic correlation from causation across 52 diseases and complex traits. Nature Genetics *50*, 1728–1734. https://doi.org/10.1038/s41588-018-0255-0.

20. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nature Genetics *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.

21. Baechler, E.C., Batliwalla, F.M., Karypis, G., Gaffney, P.M., Ortmann, W.A., Espe, K.J., Shark, K.B., Grande, W.J., Hughes, K.M., Kapur, V., et al. (2003). Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus. Proceedings of the National Academy of Sciences of the United States of America *100*, 2610–2615. https://doi.org/10.1073/pnas.0337679100.

22. Colhoun, H.M., McKeigue, P.M., and Smith, G.D. (2003). Problems of reporting genetic associations with complex outcomes. Lancet *361*, 865—872.

23. Skrivankova, V.W., Richmond, R.C., Woolf, B.A.R., Davies, N.M., Swanson, S.A., VanderWeele, T.J., Timpson, N.J., Higgins, J.P.T., Dimou, N., Langenberg, C., et al. (2021). Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): Explanation and elaboration. BMJ (Clinical research ed.) *375*, n2233. https://doi.org/10.1136/bmj.n2233.

24. Burgess, S., Davey Smith, G., Davies, N.M., Dudbridge, F., Gill, D., Glymour, M.M., Hartwig, F.P., Kutalik, Z., Holmes, M.V., Minelli, C., et al. (2023). Guidelines for performing Mendelian randomization investigations: Update for summer 2023. Wellcome Open Research *4*, 186. https://doi.org/10.12688/wellcomeopenres.15555.3.

Figures

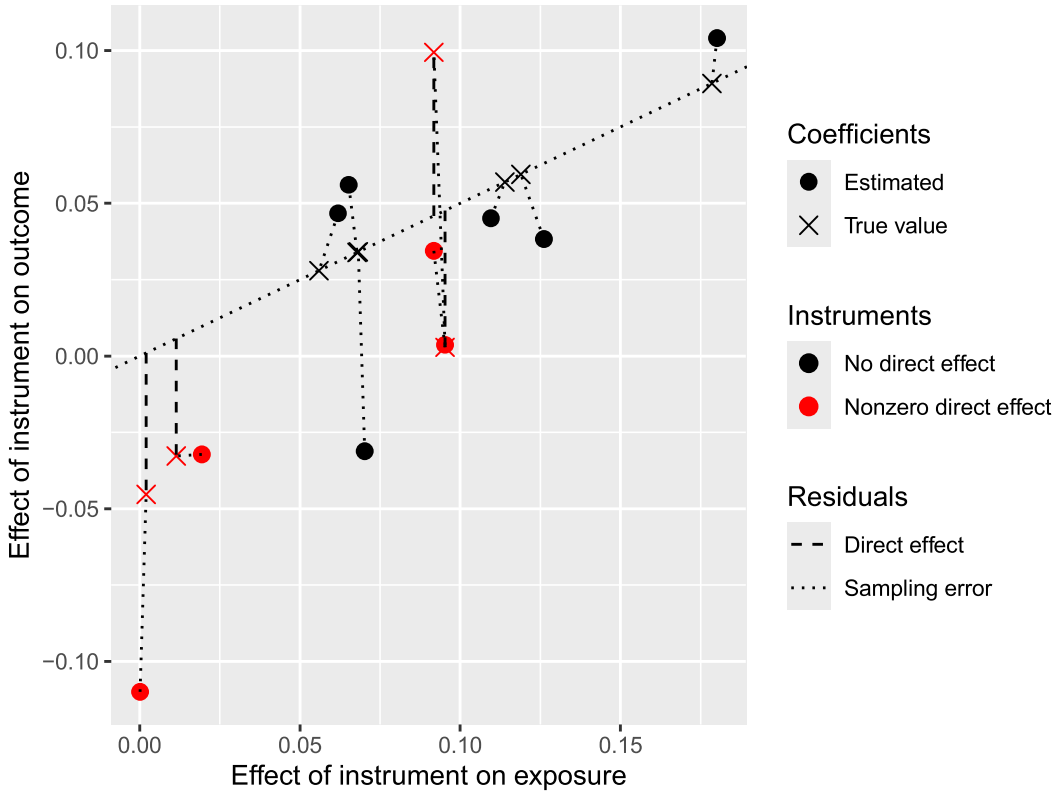

Fig 1. Plot of instrument-outcome coefficients against instrument-exposure coefficients, based on simulated data. Of the 10 instruments, 6 have no direct effect; the coordinates of the true values of the coefficients of these instruments (black crosses) lie on a straight line passing through the origin. The slope of this line is the causal effect parameter. For the other 4 instruments, the coordinates of the true values of the coefficients (red crosses) are shifted vertically by the direct effects. The coordinates of the estimated values of the coefficients (filled circles) are shifted vertically and horizontally from the true values by sampling errors

Tables

Table 1. Tests of causality used in a sample of 38 publications reporting original results of 2-sample Mendelian randomization analysis

PubMed ID	Year	Number of unlinked instruments	Support for causality reported	Methods used to test for causality					p-value threshold
				IVW	Random effects IVW	Weighted median	MR-PRESSO	MR-RAPS	
29974225	2019	60	+	+	.	+	.	.	<0.05
30723964	2019	37 to 78	+	+	.	+	.	.	<0.05
31253830	2019	3	+	+	.	.	.	.	<0.001
31801993	2020	4 to 62	-	-	.	-	.	.	<0.05
32162118	2020	16	+	+	.	+	.	.	<0.05
32467347	2020	12	-	-	.	-	.	.	<0.05
32996171	2021	10 to 704	+	+	.	+	+	.	<0.05
32997995	2020	7 to 96	+	.	+	.	.	.	<0.05
33594380	2021	32 to 124	+	+	.	+	+	.	<0.01
34187478	2021	56	+	+	.	.	.	.	<0.001
34279564	2021	1 to 7	-	-	.	-	-	.	<0.05
35856088	2022	103 to 139	+	+	+	+	.	.	<0.001
35995490	2022	22 to 524	+	+	.	+	.	.	<0.05
36118886	2022	11	+	+	.	+	.	.	<0.05
36722420	2023	43	-	-	.	-	-	.	<0.05
36972688	2023	294 to 361	+	+	.	+	.	.	<0.05
37275551	2023	16 to 22	+	+	.	+	+	+	<0.05
37480221	2023	7	+	+	.	+	+	.	<0.05
37547307	2023	7 to 81	-	-	.	-	-	.	<0.01
37805541	2023	3 to 71	+	.	+	+	+	.	<0.05
37900136	2023	132 to 580	+	+	+	+	+	.	<0.05
38169560	2024	35	+	+	.	+	+	.	<0.05
38316767	2024	45	+	-	.	+	.	.	<0.05
38427644	2024	32 to 91	-	-	-	-	-	.	<0.05
38455666	2024	16 to 25	+	+	+	+	+	.	<0.05
38469141	2024	26	+	+	.	+	+	+	<0.05
38605947	2024	16	+	+	.	+	+	.	<0.05
38659993	2024	5 to 19	+	+	.	+	+	.	<0.05
38681765	2024	14 to 37	-	-	.	-	+	.	<0.05
38784581	2024	11 to 142	+	+	+	+	+	.	<0.05
38903105	2024	Not stated	+	+	.	+	+	+	<0.001
38909247	2024	8 to 17	+	.	+	+	.	.	<0.001
38990854	2024	6 to 64	+	+	.	+	+	+	<0.001
39052079	2024	87	-	-	.	-	.	.	<0.05
39148061	2024	12 to 49	+	.	+	+	.	.	<0.05
39173282	2024	11	+	+	.	+	+	.	<0.05
39193020	2024	11 to 29	+	+	.	+	+	.	<0.001
39262796	2024	26 to 130	+	+	.	+	.	.	<0.05

"+" indicates positive result, "-" indicates negative result, "." indicates method not used

Table 2. Comparison of Type 1 error rates, based on 2000 simulations from a null model with 10 unlinked instruments

Test	N	Variance of test statistic	Number of tests with $p < 0.001$, out of 2000	Ratio of observed to nominal Type 1 error rate	Mean outliers detected by MR-PRESSO (4 expected)
Weighted median with SE from posterior predictive distribution	2000	0.4	0	0.0	.
Marginal likelihood	2000	0.5	1	0.5	.
MR-RAPS	1398	1.2	10	7.2	.
Random effects IVW	2000	1.6	42	21.0	.
MRPRESSO Raw	2000	1.6	42	21.0	4
Weighted median Bowden2016	2000	11.7	299	149.5	.
MRPRESSO Outlier-corrected	1830	83.9	164	89.6	4
IVW	2000	134.6	1256	628.0	.

Number of instruments with nonzero direct effects distributed as binomial with probability 0.4. Nonzero direct effects distributed as Student t with 4 df, mean zero, scale 0.05. Instrument-exposure coefficients α_j distributed as Gaussian with mean zero, standard deviation 0.1. Standard errors for the instrument-exposure and instrument-outcome coefficients distributed as Gamma with shape 10, mean 0.005 (encoded by specifying scale as 0.0005)

Table 3. Comparison of Type 2 error rates, based on 200 simulations from a model with causal effect size 0.25

Test	N	Type 1 error rate at $p < 0.001$	Type 2 error rate at $p < 0.001$	Power
Marginal likelihood	200	0	0.47	0.53
MRPRESSO Raw	200	0	0.70	0.30
Random effects IVW	200	0	0.78	0.22

Number of instruments with nonzero direct effects distributed as binomial with probability 0.4.
Nonzero direct effects distributed as Student t with 4 df, mean zero, scale 0.05.
Instrument-exposure coefficients α_j distributed as Gaussian with mean zero, standard deviation 0.1.
Standard errors for the instrument-exposure and instrument-outcome coefficients distributed as Gamma with mean 0.005, shape 10)