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

Paul McKeigue

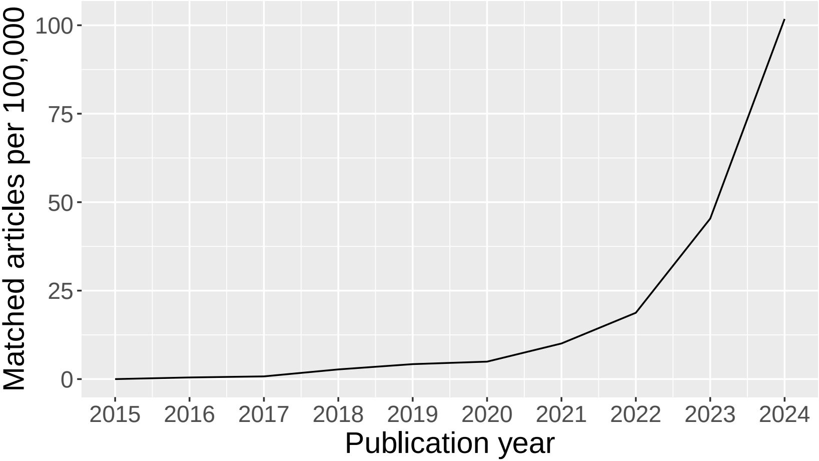
College of Medicine & Veterinary Medicine, University of Edinburgh, 07 March 2025

Mendelian randomization: early hopes

- Taubes (1995): Epidemiology faces its limits
- Keavney (2000, 2005): Fibrinogen and coronary heart disease: test of causality by 'Mendelian randomization'
- MRC Integrative Epidemiology Unit (2013): established as "a leading centre for research into methods for causal inference".
    - Hartwig (2016): Two-sample Mendelian randomization: uses genotype-exposure coefficients and genotype-outcome coefficients, estimated from different datasets
    - Bowden (2016): Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator.
    - Hemani (2018): The MR-Base platform supports systematic causal inference across the human phenome.
- 2-sample MR, using MR-Base to compute tests from summary GWAS statistics, has become the most widely used method.

Growth of articles published 2015-2024

```
(("MR-Base" OR "MR Base" OR "MRBase" OR "weighted
median") AND "mendelian randomization")  OR
TwoSampleMR OR MendelianRandomization
```



Sounding the alarm

- Munafò, Brown, Hefler, Davey Smith (2024). Managing the exponential growth of Mendelian randomization studies

  we are unfortunately seeing an ever-increasing number of MR studies that simply use summary GWAS data . . . down to current incentive structures that reward publication over knowledge . . . there are now relatively few studies applying MR methods that report null results.

- Stender, Gellert-Kristensen, Davey Smith (2024). Reclaiming Mendelian randomization from the deluge of papers and misleading findings

  Sadly, MR has run off the rails . . . a powerful and elegant scientific method for assessing causality in epidemiology is now being exploited for mass production of low-quality research, and is also reporting misleading findings . . .

  We advise editors to simply reject papers that only report 2SMR findings, with no additional supporting evidence.

Why is Mendelian randomization analysis generating false-positive results?

- Statistical inference given observed data and a model that incorporates prior information is a **well-posed problem** (Jaynes 1973):
  - unique solution given the inputs: posterior $\propto$ prior $\times$ likelihood
  - slight perturbation of the inputs will only slightly perturb the solution
  - if inputs are uninformative, solution will be uninformative (rather than false-positive)

With correct methods, mass production of research should not lead to low-quality output.

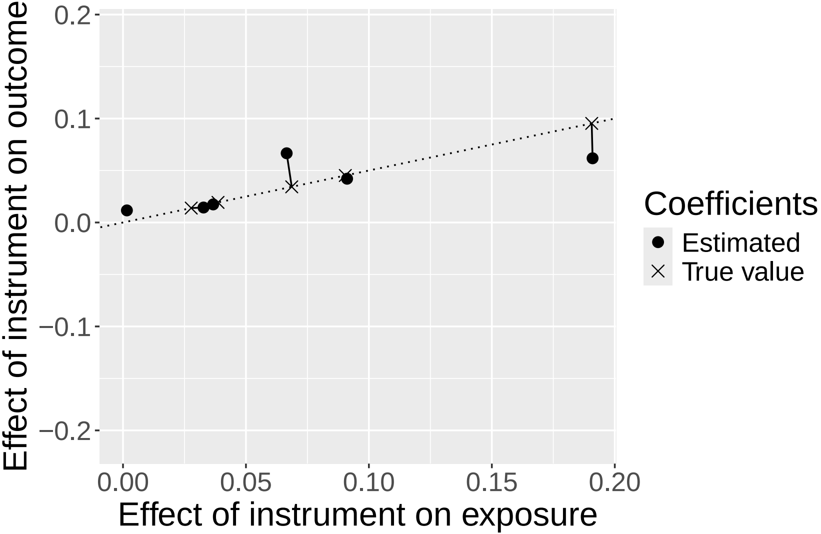
Statistical model for 2-sample Mendelian randomization

- $\boldsymbol{\alpha}$ vector of coefficients of effects of J unlinked **genetic instruments** on exposure X.
- $\boldsymbol{\beta}$ vector of coefficients of direct (pleiotropic) effects of the instruments on outcome Y, assumed to be independent of $\boldsymbol{\alpha}$
- θ causal effect of X on Y
- Crude effect γ_j of jth instrument on the outcome is the sum of the direct effect and the causal effect:

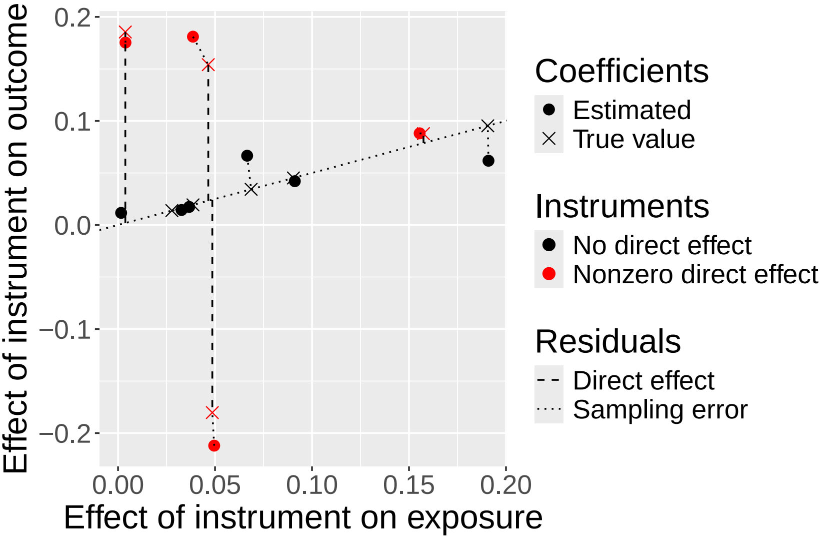
$$\gamma_j = \beta_j + \theta\alpha_j$$

For instruments with no direct effects, plot of the true values for the coefficients γ_j against those for α_j will give points lying on a straight line passing through the origin, with gradient θ.

Plot of simulated data, excluding instruments with direct effects



Plot including instruments with direct effects



Inference of the causal effect parameter

- Specify a horseshoe prior (equivalent to a spike-and-slab) on the direct effects β_j
- Observed coefficient estimates $\hat{\alpha}_j, \hat{\gamma}_j$ are modelled as Gaussian variables with means α_j, γ_j and standard deviations equal to their standard errors.
- Specify weak priors on $\alpha_j, \gamma_j, \theta$.
- Compute the posterior distribution of all parameters using a probabilistic programming language: `JAGS` (Grant 2024), `Stan`, `PyMC`, or `NumPyro` (McKeigue 2024).
    - Divide the posterior density of the causal effect parameter θ by the prior on θ to obtain the **marginal likelihood** of θ.
    - Fit a quadratic function to the log-likelihood and construct a classical hypothesis test for $\theta = 0$.

How is causal inference possible without “valid instruments”?

- Pearl 2000 - Structural Causal Model defines graphical conditions for causal effects to be identifiable.
  - instrumental variable analysis requires “valid instruments” that influence the outcome only through the exposure
- Rohde *Proc Machine Learning Res* (2022) - Causal inference is just inference
- With multiple unlinked instruments, information about the causal effect parameter accumulates as the number of instruments increases, if the direct instrument-outcome effects are independent of the instrument-exposure effects,.
  - Statistical power to detect a causal effect depends upon the number of (observations) instruments.

Pubmed query for articles published up to October 2024

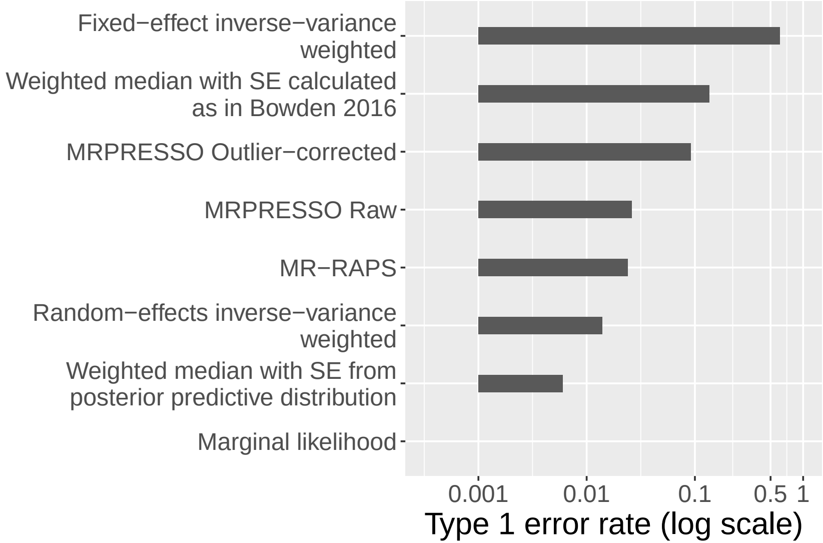
- Query `[("MR-Base" OR "MR Base" OR "MRBase" OR "weighted median") AND "mendelian randomization"]` retrieved 2629 papers
- Citation searches identified:
  - 3174 papers that cited the derivation of the weighted median estimator (Bowden 2016)
  - 308 that cited the R package `TwoSampleMR` (Hartwig 2016)
  - 59 that cited the R package `MendelianRandomization` (Yavorska 2017)
  - 2695 that cited the paper describing the MR-Base platform (Hemani 2018).

6311 unique papers remained after merging and deduplicating.

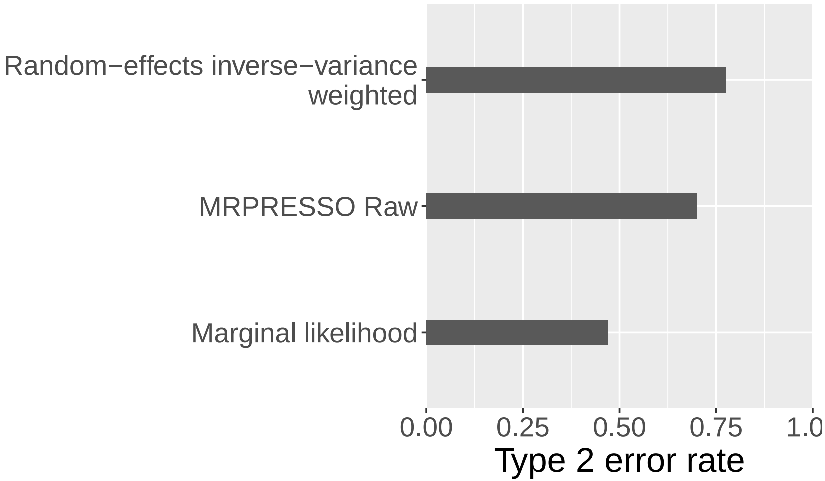
Commonly used statistical tests: sample of 40 articles

- 38 papers reported original results
- Of 30 papers that reported support for causality, 25 used the fixed-effect inverse variance weighted test (assumes no direct effects). For inference in the presence of direct effects:
  - 27 used weighted median test: calculates standard error of estimator by a "parametric bootstrap".
  - 15 used "outlier-corrected" Pleiotropy RESidual Sum and Outlier (MR-PRESSO) test.
  - 4 used Robust Adjusted Profile Score (MR-RAPS) test: profile likelihood of causal effect parameter is calculated by holding nuisance parameters (direct effects) at their maximum likelihood values
  - 22 used $p < 0.05$ as a threshold for declaring support for causality

Simulations from null model: Type 1 error rates



Simulations from non-null model: Type 2 error rates



Why do some tests yield inflated Type 1 error rates?

- Weighted median estimator:
  - a “parametric bootstrap” is a method for obtaining the sampling distribution of a test statistic by simulating new observations from the predictive distribution given the model parameters.
  - published code simulates not new observations but new coefficient estimates from the same observations.
  - incorrect procedure is replicated in R package `TwoSampleMR`, R package `MendelianRandomization`, and the `MR-Base` platform.
- MR-PRESSO: “outlier-corrected” procedure drops outliers, so standard error for the causal effect parameter is too small
- MR-RAPS: where number of nuisance parameters (direct effects) equals the number of observations, the profile likelihood does not behave as a likelihood.

Why causal inference should be based on the marginal likelihood

- Frequentist inference relies on constructing “estimators” that have desirable sampling properties: consistency, minimum variance and unbiasedness
  - even a genius can get this wrong: R A Fisher’s “fiducial inference”
- Bayesian inference requires us to specify a model and to compute the likelihood of the parameter given the model and the data, marginalizing over nuisance parameters.
  - all information favouring one value of the parameter over another is conveyed by the difference in log-likelihoods
  - in large samples the maximum-likelihood estimate is guaranteed to have desirable sampling properties.

What if effects of instruments on outcome and direct effects of instruments on exposure are coupled?

- If direct instrument-outcome effects are coupled with instrument-exposure effects, we cannot infer causality without controlling this confounding.
- This is achievable, but usually requires access to individual-level data on genotypes and exposures
  - for instance where exposure is gene transcript levels in whole blood, a likely confounder is cell type proportions.
  - can impute cell type proportions, and re-estimate the genotype-transcript coefficients with adjustment for cell type proportions
- Confounders that couple genetic effects on exposure and outcome may be of interest in their own right:
  - in systemic lupus erythematosus and psoriasis, coupling of effects on expression with effects on disease is recognizable as an "interferon signature".

Excluding reverse causation:

- Excluding reverse causation also requires individual-level data on genotypes and exposure
- If individual-level data from the dataset used to estimate genotype-exposure coefficients are available, we can exclude reverse causation.
  - for instance to study the effect of obesity on depression we can construct instruments for obesity in people who are not depressed, and vice versa.
  - can establish temporal sequence by estimating genotype-exposure coefficients before typical age of onset of disease
- Reverse causation is unlikely to explain an apparent causal effect of exposure on disease if the disease is rare.

Suggested revisions to existing guidelines for 2-sample MR analysis

- At least 20 unlinked genetic instruments are required for adequate statistical power.
- Inference should be based on the likelihood – no need to "pick a sensible range of methods"
- p-value thresholds for declaring evidence of causality should be more stringent than $p < 0.05$.
- Individual-level data will usually be required to construct scalar instruments from multiple SNPs, and to exclude confounding or reverse causation.
- Where possible, multiple exposures should be studied so that pleiotropic effects of genetic instruments can be observed directly.

Conclusions

- Used correctly, 2-sample Mendelian randomization can allow “systematic causal inference”, even without other supporting evidence
- About 4000 papers since 2015 that reported causality based on Mendelian randomization have relied on statistical methods that are likely to generate false-positive results.
- Flaws in widely-used scientific methods can be resistant to correction, especially when resources are concentrated in centres of research excellence:
    - Wood et al. Some statistical aspects of the Covid-19 response, *J R Stat Soc Series A*, meeting 10 April 2025.