5. Discussion

B233241

2025-01-07

Contents

Discussion

Word count: 1064

Aims/objectives/hypotheses

Performance of Methods

Simulation

Null Causal Effect
For a null causal effect, MR-Hevo exhibited comparable accuracy to weighted median estimator (WME), with both methods tending to slightly over-estimate the true value. Considering precision, the mean standard error (SE) for MR-Hevo was almost 2 orders of magnitude smaller than for WME, though mean confidence interval (CI)s were similarly consistent between methods despite this, reflecting differing methodology used in CI construction (see Appendix ref(#appendix-boot)). Despite near identical quantitative effect estimation, MR-Hevo was consistently more accurate in categorical classification as causal/no causal effect, exhibiting a superior false-positive rate in all 24 meta-analyses with null causal effect. This performance lends credence to the claims of MR-Hevo's creators regarding the importance of CI construction method being dissociated from SE estimation[1].

Regarding consistency, there is a trend towards both methods reporting a wider CI with conditions promoting a greater effects of invalid instruments, as might be expected. However, MR-Hevo CIs appear to widen slightly more than for WME, both with increasing proportion of invalid instruments and progressive violation of instrumental variable (IV) assumptions. This does suggest that MR-Hevo deals with pleiotropic effects more conservatively than WME, namely by appropriately reducing the reported precision of the estimate to reflect

additional uncertainty. Greater consistency of effect estimation by MR-Hevo versus WME was particularly marked when moving between different scenarios representing different assumption violations. This may reflect….

Accuracy/Precision Consistency

Positive Causal Effect

When a positive causal effect was present, MR-Hevo exhibited a slightly higher mean sensitivity than WME when all results were pooled, but on per meta-analysis basis it was out-performed in the majority of cases. The accuracy of MR-Hevo versus WME was very slightly lower overall, again tending to over-estimate the true causal effect. Precision was similar to the null causal effect case, with a comparable CI for both methods, but a much smaller SE for MR-Hevo.

Regarding consistency, there are again trends towards both methods reporting wider CIs with a greater proportion of invalid instruments and/or greater violations of IV assumptions; again, this broadening of CIs is more marked for MR-Hevo than for WME. An exception to general trends is the combination of 0% invalid instruments/20,000 participants, where MR-Hevo reports narrower CIs - and therefore correctly reports disproportionately more causal effects - than either WME using that parameter combination, or MR-Hevo at 0% invalid instruments/10,000 participants. This parameter combination appears to drive the somewhat discordant summary of sensitivity results.

It is not clear why this combination is associated with such a high causal report rate for MR-Hevo. If MR-Hevo performed particularly well versus WME in the absence of invalid instruments, this would be expected to hold in the 10,000 participants case also. Similarly, if MR-Hevo were particularly sensitive to the difference in sample size versus WME, larger discrepancies would be expected between the two methods with other parameter/scenario combinations when transitioning between 10,000 to 20,000 participants. Differences in assumption violations between scenarios do not affect this result, as assumption violations are only relevant to invalid instruments, of which there are none in the 0% invalid case. This unexpected result may be an aberrant feature of the particular datasets generated, which could be investigated by re-running the analysis from a different random seed. Alternatively, it may be that, when using MR-Hevo methods, sample size interacts in a non-linear way for and invalid instrument proportions approaching zero with respect to the method's power. If the causal report rate for this parameter/scenario combination remained high after data simulation with a different seed, this possibility could be next investigated using simulated datasets with invalid instrument proportions between 0-10%.

Re-Analysis

Results in Context

False Positive Report Rates

A key concern in the Mendelian randomisation (MR) literature of late has been of suspiciously high numbers of studies reporting causal effects, often in cases where causality does not seem biologically plausible[2]. It is against this backdrop that the creators of MR-Hevo introduce their approach as a potential solution, and it is worth considering this wider context before assessing the relative merits of each method.

Several factors may be driving high positive report rates observed in published MR studies. As with other academic fields, there is likely to be an element of publication bias in favour of studies reporting statistically significant results [?; bowden_modelling_2009]; naturally, no causal effect estimation methods will be able to address this issue. The widespread availability and use of tools such as the `TwoSampleMR` R package[3] facilitate production of MR studies at scale. Genetic studies without a plausible hypothetical basis are at high baseline risk of false positives due to implicit multiple comparisons, given the number of potential genes and/or phenotypes which could be examined[balding_tutorial_2006]. The ability to generate MR studies in an automated way renders all such spurious associations more easily accessible for attempted publication; if these are then preferentially published versus the negative findings, this could contribute to the proliferation

of positive MR studies observed. This was recognised by the creators of MR-Base themselves, prompting the to write a paper which attempted to disincentivise this practice[?]:

> "…we said what we're going to do is do the Mendelian randomization of everything against everything and put it online, and then say no one should be able to publish just the two-sample Mendelian randomization study because we've done them all"[?]

A related concern is that such methods are easily accessible to non-experts, such that the large numbers of studies so produced may also be disproportionately of low quality, without implementing safeguards against such issues. Some authors go so far as to state that MR needs "reclaiming…from the deluge of papers and misleading findings"[2], and recommending evidence "triangulation" - i.e. presenting non-MR data to support each claim of causality detected by MR methods - should be a necessary adjunct to publication of any causal MR finding[?].

By contrast, the group behind MR-Hevo assert that

Citations

The citations search returned 5,417 articles referencing the Bowden et al[4]

Multiple grouped instruments not included.

1. McKeigue PM, Iakovliev A, Spiliopoulou A, Erabadda B, Colhoun HM. Inference of causal and pleiotropic effects with multiple weak genetic instruments: Application to effect of adiponectin on type 2 diabetes [Internet]. medRxiv; 2024 [cited 2024 Oct 23]. Available from: https://www.medrxiv.org/content/10.1101/2023.12.15.23300008v2

2. Stender S, Gellert-Kristensen H, Smith GD. Reclaiming mendelian randomization from the deluge of papers and misleading findings. Lipids in Health and Disease [Internet]. 2024 Sep [cited 2025 Jan 7];23(1):286. Available from: https://doi.org/10.1186/s12944-024-02284-w

3. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Loos R, editor. eLife [Internet]. 2018 May [cited 2025 Jan 7];7:e34408. Available from: https://doi.org/10.7554/eLife.34408

4. Bowden J, Smith GD, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic Epidemiology [Internet]. 2016 Apr [cited 2024 Oct 22];40(4):304. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4849733/