Causal Effect Estimation in Mendelian Randomisation Studies - Evaluating a Novel Bayesian Approach To Genetic Pleiotropy Versus Established Weighted Median Methodology

B233241

September 2024 - June 2025

Contents

Acknowledgements

I would like to acknowledge

Contributions

Mine others

Statement of Originality

I confirm that all work is my own except where indicated, that all sources are clearly referenced….

Word Count

Word count: 4053

1 Introduction and Background

1.1 Introduction to Mendelian Randomisation (MR)

Epidemiology is the study of determinants and distribution of disease across populations; a common epidemiological study aim is therefore to seek evidence as to whether a given exposure (e.g. cigarette smoking) may cause a given outcome (e.g. lung cancer)[1]. Logistics limit experimental interventions across large groups, so insights into associations between exposures and outcomes are gleaned from observational data of people in the population of interest. Comparing health outcomes between individuals with different levels of a particular exposure may highlight potential links, e.g. higher cancer incidence in those who smoke more is consistent with a causal role for cigarettes in carcinogenesis[1].

However, correlation does not prove causation. A key epidemiological challenge is accounting for so-called "confounding" factors; these are other variables, associated with both the exposure and the outcome of interest, which represent an alternative causal explanation for any exposure-outcome links observed[2]. If smokers also drink more alcohol than non-smokers, then an observed link between smoking and increased cancer risk could plausibly be caused by increased alcohol exposure, either partially or entirely. Another potential issue with observational data is "reverse causation", where the presumed outcome is in fact a cause of the exposure; this might be the case if a cancer diagnosis drove individuals to drink and smoke more, and data were collected without respect to exposure timings.

Mendelian randomisation (MR) is a methodology intended to support causal inference from observational data. It applies the principles of instrumental variable (IV) analysis to genetic data, performing a type of natural experiment often likened to a randomised-controlled trial (RCT)[3].

In a properly conducted RCT, causality can be inferred due to a randomisation process being used as an "instrument" to allocate different levels of exposures to different experimental groups. If groups are randomly allocated, any confounding variables which might otherwise influence exposure-outcome relationships should be evenly distributed between groups, whether these confounders are known or not. As such, there should be no systematic differences between individuals from different groups in the exposure of interest - that is, there should be no bias[4]. Statistical methods can quantify the probability that any observed outcome differences could have occurred by chance, and thereafter any outcome differences can be interpreted as caused by exposure differences. As allocation and receipt of exposures is known to precede outcome measurements, reverse causality is impossible.

In MR, naturally occurring genetic variants - "genetic instruments" – are chosen based on their known association to an exposure of interest. Provided that assumptions of IV analysis are met, random assignment of alleles (i.e. variants of a given gene) from parents to offspring during meiosis creates randomisation analagous to that performed for an RCT – both measured and unmeasured confounders should be distributed evenly between the groups created, allowing valid causal inference after other sources of bias and random variation are accounted for[5].

1.2 Causal Effect Estimation in MR

At its simplest, the relationship between two continuous variables - an exposure X and outcome Y - can be represented as a linear model:

$$Y = \alpha + \beta X + \epsilon \tag{1}$$

where α represents all non-X determinants of Y, β is the causal effect of X on Y and ϵ is an error term. The β term is a numerical measure of strength of causal exposure-outcome association, where:

- $\beta = 0$ implies no causal link between exposure and outcome
- $\beta > 0$ implies X causes Y

- $\beta < 0$ implies X prevents Y

To estimate a causal effect using a genetic variant in an IV analysis, three key assumptions must be met[6]:

1. Relevance – the genetic variant must be associated with the exposure of interest
2. Independence – the genetic variant is independent of confounders of the relationship between exposure and outcome
3. Exclusion restriction – the genetic variant must not be associated with the outcome except via the exposure

These assumptions are represented graphically in Figure 1.

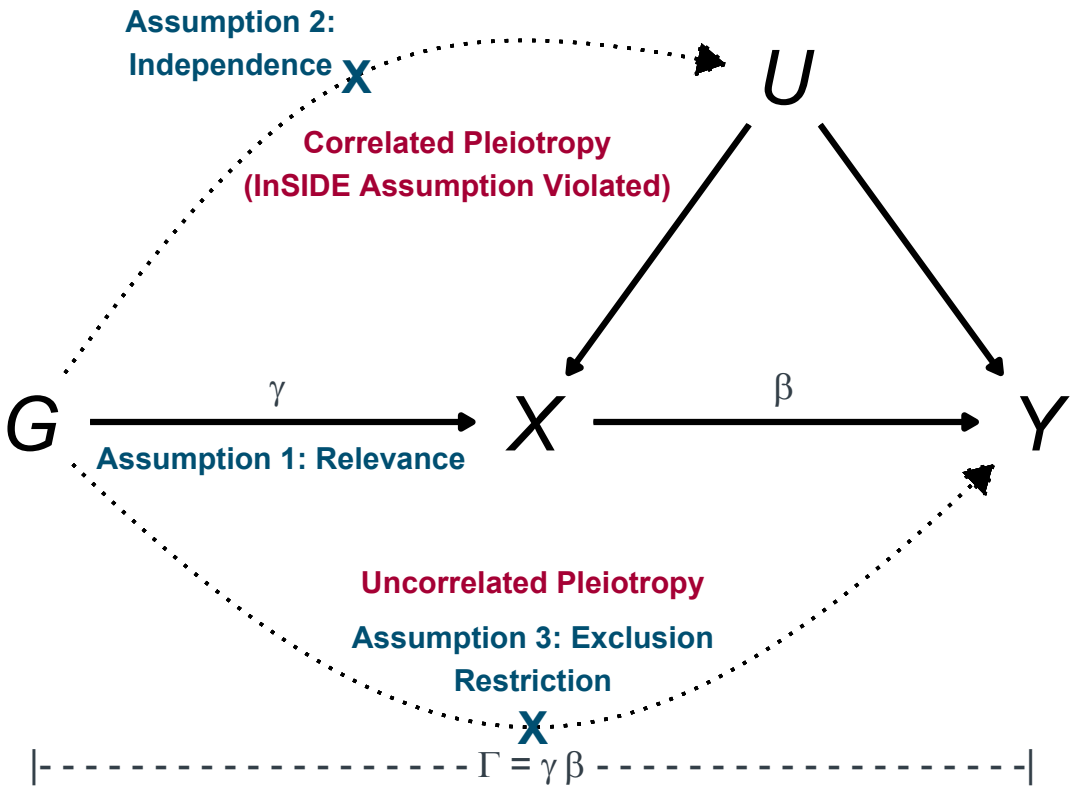


Figure 1: Causal diagram illustrating the relationships between genetic instrument G, exposure X, outcome Y and confounders of the exposure-outcome relationship U in Mendelian randomisation studies. Blue text & crosses represent key assumptions to ensure valid inference of causal effect of X on Y using G as an instrumental variable. Red text represents violations of these assumptions that may lead to invalid inference through opening of alternate causal pathways. Greek characters represent the key parameters/association coefficients to be estimated. Adapted from Burgess et al 2016[7]

Typically, MR studies estimate causal effect using a set of several genetic instruments; the causal effect estimate derived from the jth instrument is denoted $\hat{\beta}_j$. Each estimate $\hat{\beta}_j$ acknowledges there will be specific effects on the observed values of exposure and outcome given the presence of that specific genetic variable G_j under study, i.e. $\hat{\beta}_j$ is based on the instrument-conditioned exposure $X|G_j$ and outcome $Y|G_j$. These observed values of exposure and outcome can be described by their own linear models:

$$X|G_j = \gamma_0 + \gamma_j G_j + \epsilon_{X_j} \tag{2}$$

$$Y|G_j = \Gamma_0 + \Gamma_j G_j + \epsilon_{Y_j} \tag{3}$$

where, for exposure and outcome respectively:

- γ_0 and Γ_0 reflect base values without influence of the genetic variant
- γ_j and Γ_j are coefficients of association with the genetic variant, representing the extent to which an effect allele of G_j will perturb the value of X or Y versus the non-effect allele
- ϵ_{X_j} and ϵ_{Y_j} are error terms, containing contributions from confounders of the exposure-outcome relationship (U in the causal diagram), and all genetic variants except G_j.

It can be shown that a simple causal effect estimate for the exposure on the outcome can be obtained from a single genetic instrument by the Wald method, dividing the coefficient of gene-outcome association by the coefficient of gene-exposure association, i.e.:

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j} \tag{4}$$

Each instrument may be valid or invalid, depending on it meeting the above assumptions. The overall causal effect estimate $\hat{\beta}$ from any given MR method will typically seek to pool effect estimates from several instruments so as to minimise effects of any invalid instruments included, e.g. by removing/down-weighting contributions of genetic instruments which violate one or more assumptions. This is equivalent to plotting all estimated coefficients of gene-outcome association ($\bar{\Gamma}$) versus all estimated coefficients of gene-exposure association ($\bar{\gamma}$) for the set of instruments, then using the gradient of a regression line through the points as the causal effect estimate $\hat{\beta}$; picking an MR methodology is analogous to choosing the method to draw the line of best fit (Figure 2). For binary outcomes, the causal effect estimate can be converted to an odds ratio (OR) through exponentiation, i.e.:

$$OR = e^{\hat{\beta}} \tag{5}$$

1.3 Violations to Assumptions

In practice, only the relevance assumption can be directly tested and proven. Typically, genetic variants for MR studies are selected as instruments based on Genome Wide Association Studies (GWAS), which quantify associations between small genetic variations - known as single nucleotide polymorphism (SNP)s - and various phenotypes. Association between genetic variants and a phenotypes representing exposures of interest can be partly assured by selection using an appropriate genome-wide significance level (e.g. $p < 10^{-8}$). Statistical testing can also quantify the gene-exposure relationship; commonly used measures include the r^2 statistic, representing the proportion of variance in the exposure explained by the genotype, and the related F-statistic, which additionally accounts for the the sample size under investigation[9]. An F-statistic of ≥ 10 is generally considered to represent a strong enough gene-exposure association to consider a genetic instrument for use[2].

The assumptions of independence and exclusion restriction depend on all possible confounders of the exposure-outcome association, both measured and unmeasured; as such, these can never be proven absolutely. Various methods have been proposed to quantify and account for violations of these two additional assumptions, including the weighted median estimator, described below[8].

The main methods to avoid violations of the independence assumption relate to appropriate selection of populations studied to avoid confounding due to ancestry or population stratification. For example, in two-sample MR studies, where gene-exposure and gene-outcome coefficients are estimated from two separate GWAS studies, it is recommended to select GWAS studies performed in similar population groups (e.g. both

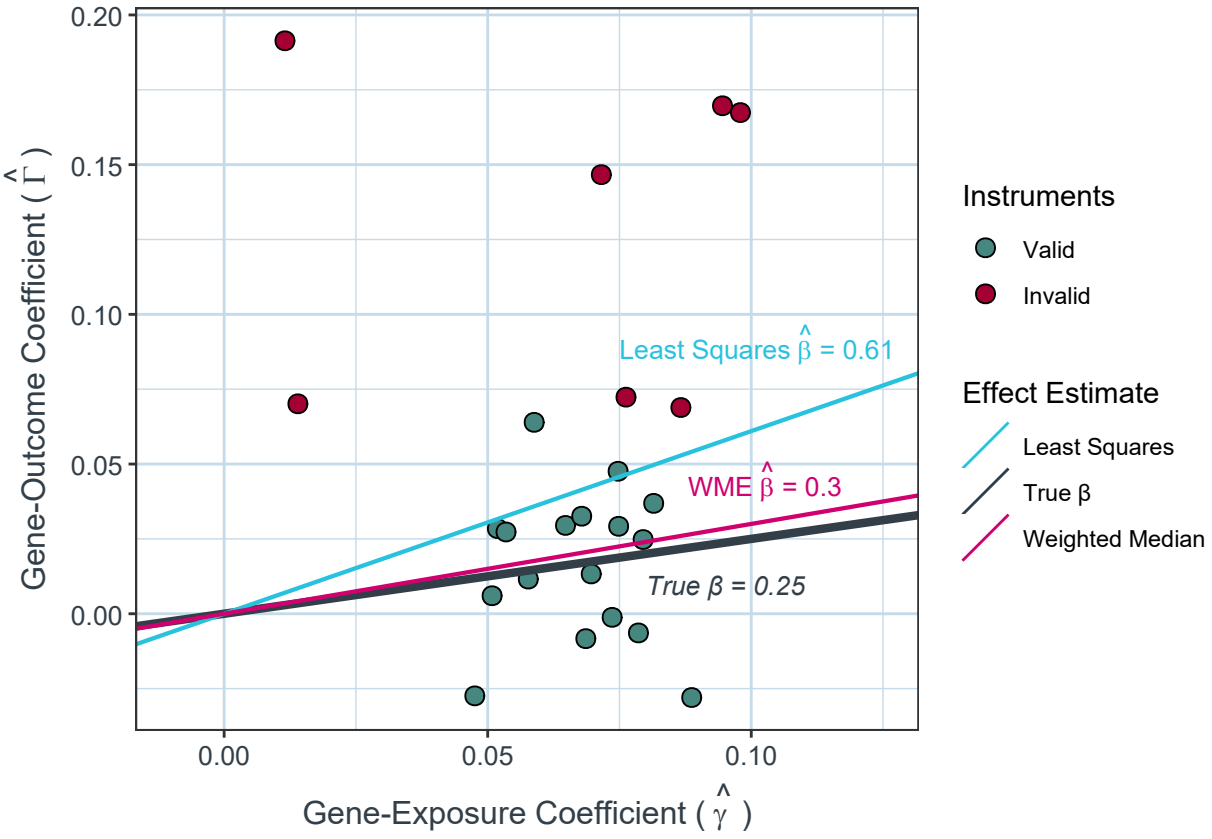


Figure 2: Simulated MR Study on 10,000 individuals using 25 genetic instruments, of which 30% are invalid (red points) and introduce directional pleiotropic effects. The true value of the exposure-outcome causal effect is 0.25 (grey line, causal effect represented by gradient). Regression using an unajusted least-squares linear model (light blue line) results in a biased estimate in the positive direction due to the influence of the invalid instruments. Using the Weighted Median Estimator method (pink line) attenuates the effects of the invalid instruments, resulting in an estimate closer to the true value. Adapted from Bowden et al 2016[8]

in Western Europeans). This practice helps avoid spurious exposure-outcome associations being generated by confounding due to underlying differences in allele frequency, baseline disease risks etc between ancestrally different populations[9].

Exclusion restriction is a particularly universal issue in MR, due to so-called (horizontal) genetic pleiotropy, where a single genetic variant may have multiple "pleiotropic" effects – i.e. it may influence several traits simultaneously. Such pleiotropic effects may be unknown and open unmeasured causal pathways between a genetic instrument and the outcome (Figure 1), thus potentially biasing MR estimates of the association between exposure and outcome. As pleiotropy influences outcome separate to the path involving the exposure of interest, the term "direct effects" is also used[10]. Where pleiotropic effects are in both positive and negative directions with a mean of zero - "balanced pleiotropy" - then they only add noise to causal effect estimation[11]. By contrast, "directional pleiotropy", where the mean of pleiotropic effects is non-zero, may introduce bias[8].

If such an additional causal pathway acts between gene G and outcome Y via a confounding factor U, then the magnitude of direct/overall effects of G on Y will correlate with the effects of G on X (i.e. $\Gamma \propto \gamma$), and "correlated pleiotropy" is present. If an additional causal pathway acts directly between gene G and outcome Y independent of both exposure X and confounders U, this results in "uncorrelated pleiotropy" (Figure 1). Both correlated and uncorrelated pleiotropy can introduce bias which distorts the estimate of the true causal effect. In general, correlated pleiotropy is more challenging to account for; several MR methods explicitly require an additional assumption of Instrument Strength Independent of Direct Effect (InSIDE), i.e no correlated pleiotropy to be present[12].

1.4 Weighted Median Estimator (WME)

A common approach to produce exposure-outcome causal effect estimates robust to violations of the exclusion restriction assumption is the weighted median estimator (WME) method, proposed by Bowden et al[8].

In WME analysis, several genetic instruments are used to estimate the exposure-outcome causal effect $\hat{\beta}$. Each instrument is known to be associated with the exposure of interest, but an unknown proportion of these instruments may be invalid due to pleiotropic genetic effects. Any instrument linked to an outcome via multiple pleiotropic causal pathways will exhibit a less consistent gene-outcome association than a relationship mediated by a single pathway; this results in larger variance in causal estimates derived from invalid/pleiotropic genetic instruments versus estimates from valid instruments.

WME therefore assigns a weight to each genetic instrument's estimate of the causal effect according to the inverse of the variance of the estimate; these weighted effect estimates are used to construct a cumulative distribution function for probability of true causal effect size across the range of estimated values. The 50th percentile of this distribution can then be taken as a "weighted median estimate" of the true causal effect, theoretically producing consistent causal estimates even if up to 50% of the included information comes from invalid instruments[8]. An example of WME attenuating the effects of invalid instruments is shown in Figure 2.

1.5 Issues With WME CIs

WME calculation methods are available via several prolific MR tools: the R packages "MendelianRandomization"[13] and "TwoSampleMR", and the MR-Base web platform[14]. However, these implement the original authors' suggested process of generating 95% confidence intervals for WME, which deviates from accepted re-sampling methodology:

> "We found the bootstrap confidence interval…too conservative. However, the bootstrap standard error… gave more reasonable coverage using a normal approximation (estimate ±1.96 x standard error) to form a 95% confidence interval"[8]

This modification, explicitly aiming to boost estimate precision artificially, would be expected to lead to a high Type 1 error rate, which has been a growing concern in the field of late[15]. The theoretical issues with this approach, and the fundamentals of bootstrapping in general, are covered in Appendix @(ref:appendix-boot).

1.6 MR-Hevo

MR-Hevo is an R package which uses more typical Bayesian methodology to estimate MR causal effects and corresponding 95% confidence intervals. It uses the probabilistic programming language, Stan, to directly sample the posterior probability distribution of pleiotropic effects on the outcome, rather than making untested assumptions about the shape of this distribution as current WME implementations do[16].

MR-Hevo incorporates several additional features which its creators claim further aid valid causal inference. Most MR methods can only account for one genetic variant per genetic locus (i.e. per location in the genome). If multiple variants exist at a given locus, generally only one can be selected as an instrument for further MR analysis. HR-Hevo handles multiple instruments per genetic locus via scalar construction, essentially assigning a "score" to each locus based on the variant(s) present, thus incorporating more information than if closely grouped variants had been discarded[16]. As MR-Hevo is based on a Bayesian approach, it generates estimates which incorporate relevant existing information generated by prior studies, increasing the amount of data informing each estimate. In this case, MR-Hevo specifies a prior probability distribution reflecting prior knowledge that most individual genetic instruments will have only small effects on complex traits[17,18], further aiding biologically plausible inference regarding distribution of pleiotropic effects.

1.7 Aims and Objectives

The main aim of this study will be to demonstrate if the WME approach gives over-confident causal estimates in the presence of pleiotropy, and whether this issue is more correctly handled by the MR-Hevo Bayesian approach. This will be achieved through addressing the research questions and objectives as outlined below:

1.7.1 Research Questions:

1. How does MR-Hevo perform versus the weighted median estimator when estimating causal effects in MR studies?
2. Do conclusions of existing MR studies using weighted median causal effect estimation change if MR-Hevo methods are used?

1.7.2 Objectives:

1. Quantify the precision of MR-Hevo causal estimates for simulated data under differing sets of common assumptions, with reference to the weighted median estimator
2. Evaluate the consistency of MR-Hevo causal estimates for simulated data under differing sets of common assumptions, with reference to the weighted median estimator
3. Compare the conclusions drawn from MR-Hevo causal effect estimation versus the weighted median estimator on real-world data

2 Methods

2.1 Simulation Study

To evaluate the performance of MR-Hevo causal estimation relative to WME, the precision and consistency of both methods were quantified using simulated datasets with known parameter values.

2.1.1 Data Simulation

To aid comparability with existing methods and literature, the simulation methodology of the original WME exposition was reproduced based on published models and parameters in Appendix 3 of its supplementary materials[8]. Full details of simulation reproduction, including code and validation of outputs, is presented in C.

In brief, simulations were created based on three different scenarios, each representing a common set of assumptions about underlying data used for MR, and each increasingly challenging to the performance of any given MR causal estimation methodology:

1. Balanced pleiotropy, InSIDE assumption satisfied - A proportion of invalid genetic instruments are present and introduce pleiotropic effects uncorrelated with the instrument strength; these pleiotropic effects are equally likely to be positive as negative with a mean value = 0, thus introducing noise into the estimation of causal effect.
2. Directional pleiotropy, InSIDE assumption satisfied - A proportion of invalid genetic instruments are present and introduce pleiotropic effects uncorrelated with the instrument strength; these pleiotropic effects are positive only, with a mean value > 0, thus biasing the causal effect estimate in a positive direction.
3. Directional pleiotropy, InSIDE assumption not satisfied - A proportion of invalid genetic instruments are present and introduce pleiotropic effects correlated with the instrument strength through action via a confounder; these pleiotropic effects are positive only, with a mean value > 0, thus potentially biasing the causal effect estimate in a positive direction to an even greater extent than Scenario 2.

1,000 simulated datasets of participant-level data were generated for every combination of each scenario and each the following simulation parameters:

- Proportion of invalid instruments: 0%, 10%, 20% or 30%
- Number of participants: $n = 10,000$ or $n = 20,000$
- Causal effect: null ($\beta = 0$) or positive ($\beta = 0.1$)

The same set of 25 simulated genetic instruments were used across all datasets, with the status of each as valid/invalid determined by random draw per instrument at the start of each simulation run of 1,000 datasets.

Genotypes were simulated as for a two-sample setting: where number of particpants was $n = 10,000$, 20,000 genotypes were simulated - 10,000 for the cohort used to estimate gene-exposure association ($\hat{\gamma}$), and a separate cohort of 10,000 used to estimate gene-outcome association ($\hat{\Gamma}$). Parameter values for effect allele frequency were not specified by Bowden et al, though initial testing showed values around 0.5 produced WME causal effect estimates closest to published values when other parameters were matched[8]. As such, effect allele frequencies were assigned per instrument from a uniform distribution between 0.4 to 0.6. Each effect allele freqency thus generated per instrument was then used as a probability to assign each simulated participant effect alleles for each instrument via two draws from a binomial distribution.

2.1.2 Analysis of Simulated Data

Each dataset generated was analysed using both WME and MR-Hevo methods, via functions from the `TwoSampleMR` and `mrhevo` packages, respectively[14,16]. Results were aggregated per group of 1,000 simulated datasets corresponding to a particular combination of scenario and parameter values. This resulted in one meta-analysis reported per combination of scenario/parameter values, each including 1,000 simulated MR studies using the same 25 genetic instruments in the same population. Aggregated measures for both WME and MR-Hevo per meta-analysis were mean causal effect estimate; mean standard error of the causal effect estimate; and causality report rate, i.e. percentage of simulated studies reported as showing a non-null causal effect, either by p-value <0.05 (WME), or by a 95% credible interval for causal effect estimate not including 0 (MR-Hevo).

Results of the above aggregations were tabulated as per Tables 2 and 3 of Bowden et al[8] to allow direct comparisons of both methods versus each other and versus the published characteristics of existing MR causal estimation methods.

2.2 Re-Analysis of Published Data

To investigate the potential implications of any differences in performance between WME and MR-Hevo methods, a selection of published studies resporting causal effect estimates using the WME method was re-analysed. A sample size of 10 published studies was decided as a pragmatic compromise between the scope of this study and the need to check consistency of any observed differences. In the original Bowden et al simulation studies, the WME causal estimation method was shown to generate a false-positive report rate of $\geq 30\%$ with relatively minor violations of relevant assumptions[8]; therefore, even this relatively small sample of 10 studies might be expected to demonstrate differences between methods if the MR-Hevo approach is as appropriately conservative as its creators propose.

To estimate the upper bound of the potential impact of MR-Hevo versus existing WME methodology, studies were chosen for re-analysis based on their number of citations in the wider MR literature. Compared to studies with few or no citations, highly-cited studies would be expected to have a larger impact on their respective fields if their conclusions were to change. In addition, highly-cited works will typically have been submitted to more scrutiny than less-cited works - both during peer review whilst under consideration by journals likely to produce highly-cited works, and from the wider scientific community following the widespread dissemination evidenced by a high citation count. As such, it would be expected that highly-cited works are likely to be free of significant methodological flaws which may impede interpretation of any re-analysis.

2.2.1 Citation Search

The Scopus search platform [@] was used on 15/04/2025 to retrieve all articles citing the original weighted median estimator exposition paper[8]. The articles returned were sorted by the number of times each article itself had been cited, and the resulting list was saved to RIS format in blocks of ten articles for upload into the Covidence evidence synthesis platform. Abstracts were screened by a single reviewer (B233241), starting with the most cited article and proceeding in descending order of citation count, against the following inclusion and exclusion criteria:

Inclusion criteria:

- Original two-sample MR study
- Able to determine samples' ancestry sufficient to establish presence/potential degree of participant overlap between groups
- Reporting ≥ 20 human genetic instruments relating to exposure

- Reports details of effect/non-effect alleles
- Regression coefficients and standard errors and/or confidence intervals available for each genetic instrument used
- Uses Weighted Median Estimator

Exclusion criteria:

- Methodology paper, review article, editorial or letter
- English full-text not accessible

Where eligibility could not be determined from abstract screening alone, full texts were retrieved and screened against the same criteria. Screening of abstracts and full texts was undertaken in blocks of ten articles, until the target of ten included studies for reanalysis had been reached.

Where an article reported multiple exposure-outcome associations, data were only extracted for the association with the highest number of genetic instruments available, or else for the first reported association where several were based on the same number of instruments. Data were extracted from full texts of included studies using a standardised data collection template, which included publication details, citation count, primary study question, degree of participant overlap between groups, number/details of genetic instruments used, effect estimates/standard errors calculated, and conclusion regarding causality as determined by the weighted median estimator method.

2.3 Data Manipulation and Analysis

All simulations, data manipulations and data analyses were performed in R version 4.4.3 (2025-02-28 ucrt)[19].

For the simulation study, full details of computation are available in Appendix C.

For citation search data, a standardised data collection form was Microsoft Excel[20] to create .csv files for subsequent analysis in R; Excel's "Get Data" function was also used to extract tables of genetic instruments where these were presented in non-csv format (e.g. pdf).

Data cleaning for citation search data was primarily undertaken using the Tidyverse suite of R packages[21]. A full list of packages used can be found in Appendix @(ref:appendix-pkg).

Data were manually screened at summary level and relevant features were extracted. Data were checked for completeness, consistency, duplicate values and plausibility. Data were transformed to an appropriate data type, and encoding of genetic variables was standardised into a single format. Missing values for association coefficients and standard error (SE)s were imputed as the mean value calculated per dataset. It was noted during early testing that MR-Hevo functions do not operate correctly when zero values are present in coefficients of genetic association or their standard errors; such zero values were therefore re-coded as an arbitrarily low value of 10^{-100}.

2.4 Ethical Approval

The protocol for this work has been reviewed and approved by the Usher Masters Research Ethics Group (UMREG) at the University of Edinburgh, Ethics ID: UM241126. Due to the nature of the project, using simulated and publically available data only, no significant ethical issues were foreseen, and sponsorship was deemed unnecessary by the UMREG reviewing panel.

3 References

1. Coggon D, Rose G, Barker D. Chapter 1. What is epidemiology? | The BMJ. In: The BMJ | The BMJ: Leading general medical journal Research Education Comment [Internet]. 2003 [cited 2025 Apr 29]. Available from: https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/1-what-epidemiology

2. Martens EP, Pestman WR, Boer A de, Belitser SV, Klungel OH. Instrumental Variables: Application and Limitations. Epidemiology [Internet]. 2006 May [cited 2025 Apr 29];17(3):260. Available from: https://journals.lww.com/epidem/fulltext/2006/05000/instrumental_variables__application_and.10.aspx

3. Hernán MA, Robins JM. Instruments for Causal Inference: An Epidemiologist's Dream? Epidemiology [Internet]. 2006 Jul [cited 2025 Apr 29];17(4):360. Available from: https://journals.lww.com/epidem/fulltext/2006/07000/instruments_for_causal_inference___an.4.aspx#JCL1-2

4. Stel VS, Dekker FW, Zoccali C, Jager KJ. Instrumental variable analysis. Nephrology Dialysis Transplantation [Internet]. 2013 Jul [cited 2025 Apr 29];28(7):1694–9. Available from: https://doi.org/10.1093/ndt/gfs310

5. Davies NM, Holmes MV, Smith GD. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. BMJ [Internet]. 2018 Jul [cited 2025 Jan 7];362:k601. Available from: https://www.bmj.com/content/362/bmj.k601

6. Lousdal ML. An introduction to instrumental variable assumptions, validation and estimation. Emerging Themes in Epidemiology [Internet]. 2018 Jan [cited 2025 Apr 29];15:1. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5776781/

7. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. Epidemiology (Cambridge, Mass) [Internet]. 2016 Nov [cited 2024 Oct 22];28(1):30. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC5133381/

8. 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/

9. Richmond RC, Smith GD. Mendelian Randomization: Concepts and Scope. Cold Spring Harbor Perspectives in Medicine [Internet]. 2022 Jan [cited 2024 Oct 22];12(1):a040501. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8725623/

10. Hemani G, Bowden J, Smith GD. Evaluating the potential role of pleiotropy in Mendelian randomization studies. Human Molecular Genetics [Internet]. 2018 May [cited 2024 Oct 23];27(R2):R195. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6061876/

11. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nature Genetics [Internet]. 2020 Jul [cited 2025 May 22];52(7):740–7. Available from: https://www.nature.com/articles/s41588-020-0631-4

12. Grant AJ, Burgess S. A Bayesian approach to Mendelian randomization using summary statistics in the univariable and multivariable settings with correlated pleiotropy. The American Journal of Human Genetics [Internet]. 2024 Jan [cited 2025 May 20];111(1):165–80. Available from: https://www.cell.com/ajhg/abstract/S0002-9297(23)00433-0

13. Yavorska OO, Burgess S. MendelianRandomization: An R package for performing Mendelian randomization analyses using summarized data. International Journal of Epidemiology [Internet]. 2017 Dec [cited 2024 Oct 23];46(6):1734–9. Available from: https://doi.org/10.1093/ije/dyx034

14. 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

15. 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

16. 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

17. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature genetics [Internet]. 2010 Jun [cited 2024 Oct 23];42(7):570. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC4615599/

18. Piironen J, Vehtari A. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics [Internet]. 2017 Jan [cited 2025 Jan 7];11(2):5018–51. Available from: https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-11/issue-2/Sparsity-information-and-regularization-in-the-horseshoe-and-other-shrinkage/10.1214/17-EJS1337SI.full

19. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2025. Available from: https://www.R-project.org/

20. Microsoft Corporation. Microsoft Excel [Internet]. 2018. Available from: https://office.microsoft.com/excel

21. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. Journal of Open Source Software. 2019;4(43):1686.

22. Buscaglia DLS&R. Chapter 3 Confidence Intervals via Bootstrapping | Introduction to Statistical Methodology, Second Edition [Internet]. 2020 [cited 2025 May 25]. Available from: https://bookdown.org/dereksonderegger/570/3-confidence-intervals-via-bootstrapping.html

23. Ross SM. Chapter 6 - Distributions of Sampling Statistics. In: Ross SM, editor. Introduction to Probability and Statistics for Engineers and Scientists (Fifth Edition) [Internet]. Boston: Academic Press; 2014 [cited 2025 May 25]. p. 207–33. Available from: https://www.sciencedirect.com/science/article/pii/B978012394811350006X

24. Cata JP, Klein EA, Hoeltge GA, Dalton JE, Mascha E, O'Hara J, et al. Blood Storage Duration and Biochemical Recurrence of Cancer After Radical Prostatectomy. Mayo Clinic Proceedings [Internet]. 2011 Feb [cited 2025 May 25];86(2):120–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3031436/

25. Higgins P. medicaldata: Data package for medical datasets [Internet]. 2021. Available from: https://CRAN.R-project.org/package=medicaldata

26. Chaput R. acronymsdown: Acronyms and glossaries support for RMarkdown [Internet]. 2025. Available from: https://github.com/rchaput/acronymsdown

27. Xie Y. bookdown: Authoring books and technical documents with R markdown [Internet]. Boca Raton, Florida: Chapman; Hall/CRC; 2016. Available from: https://bookdown.org/yihui/bookdown

28. Xie Y. bookdown: Authoring books and technical documents with r markdown [Internet]. 2025. Available from: https://github.com/rstudio/bookdown

29. Fox J, Weisberg S. An R companion to applied regression [Internet]. Third. Thousand Oaks CA: Sage; 2019. Available from: https://www.john-fox.ca/Companion/

30. Wilke CO. cowplot: Streamlined plot theme and plot annotations for "ggplot2" [Internet]. 2024. Available from: https://CRAN.R-project.org/package=cowplot

31. Csárdi G. crayon: Colored terminal output [Internet]. 2024. Available from: https://CRAN.R-project.org/package=crayon

32. Wickham H, Hester J, Chang W, Bryan J. devtools: Tools to make developing r packages easier [Internet]. 2022. Available from: https://CRAN.R-project.org/package=devtools

33. Barrett M. ggdag: Analyze and create elegant directed acyclic graphs [Internet]. 2024. Available from: https://CRAN.R-project.org/package=ggdag

34. Yutani H. gghighlight: Highlight lines and points in "ggplot2" [Internet]. 2023. Available from: https://CRAN.R-project.org/package=gghighlight

35. Rodriguez-Sanchez F, Jackson CP. grateful: Facilitate citation of R packages [Internet]. 2024. Available from: https://pakillo.github.io/grateful/

36. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2025. Available from: https://www.R-project.org/

37. Müller K. here: A simpler way to find your files [Internet]. 2020. Available from: https://CRAN.R-project.org/package=here

38. Couch SP, Bray AP, Ismay C, Chasnovski E, Baumer BS, Çetinkaya-Rundel M. infer: An R package for tidyverse-friendly statistical inference. Journal of Open Source Software. 2021;6(65):3661.

39. Zhu H. kableExtra: Construct complex table with "kable" and pipe syntax [Internet]. 2024. Available from: https://CRAN.R-project.org/package=kableExtra

40. Xie Y. knitr: A comprehensive tool for reproducible research in R. In: Stodden V, Leisch F, Peng RD, editors. Implementing reproducible computational research. Chapman; Hall/CRC; 2014.

41. Xie Y. Dynamic documents with R and knitr [Internet]. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC; 2015. Available from: https://yihui.org/knitr/

42. Xie Y. knitr: A general-purpose package for dynamic report generation in R [Internet]. 2025. Available from: https://yihui.org/knitr/

43. Bengtsson H. matrixStats: Functions that apply to rows and columns of matrices (and to vectors) [Internet]. 2025. Available from: https://CRAN.R-project.org/package=matrixStats

44. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2025. Available from: https://www.R-project.org/

45. Xie Y, Allaire JJ, Grolemund G. R markdown: The definitive guide [Internet]. Boca Raton, Florida: Chapman; Hall/CRC; 2018. Available from: https://bookdown.org/yihui/rmarkdown

46. Xie Y, Dervieux C, Riederer E. R markdown cookbook [Internet]. Boca Raton, Florida: Chapman; Hall/CRC; 2020. Available from: https://bookdown.org/yihui/rmarkdown-cookbook

47. Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, et al. rmarkdown: Dynamic documents for r [Internet]. 2024. Available from: https://github.com/rstudio/rmarkdown

48. Stan Development Team. RStan: The R interface to Stan [Internet]. 2025. Available from: https://mc-stan.org/

49. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genetics [Internet]. 2017;13(11):e1007081. Available from: https://doi.org/10.1371/journal.pgen.1007081

50. Hemani G, Zheng J, Elsworth B, Wade K, Baird D, Haberland V, et al. The MR-base platform supports systematic causal inference across the human phenome. eLife [Internet]. 2018;7:e34408. Available from: https://elifesciences.org/articles/34408

51. Marwick B. wordcountaddin: Word counts and readability statistics in r markdown documents [Internet]. 2025. Available from: https://github.com/benmarwick/wordcountaddin

A Appendix: List of Abbreviations

CI confidence interval
CLT central limit theorem
IV instrumental variable
InSIDE Instrument Strength Independent of Direct Effect
MR Mendelian randomisation
RCT randomised-controlled trial
SD standard deviation
SE standard error
SNP single nucleotide polymorphism
UMREG Usher Masters Research Ethics Group
WME weighted median estimator

B Appendix: Bootstrapping

B.1 Bootstrapping - General Method

The typical process for "bootstrap" generating an estimate, SE and confidence interval (CI)s of a population parameter (e.g. population mean μ) from a sample x is as follows[22]:

1. A sample, x, of n individuals is selected from a total population, X, of N individuals
2. This sample x is then treated as the "bootstrap population"; the empirical distribution of values in the n individuals in the bootstrap population is taken to be broadly representative of the distribution of values in the underlying population X of N individuals
3. A "bootstrap sample", x^*, is then obtained by re-sampling individuals from the bootstrap population with replacement n times per bootstrap sample, i.e. the new bootstrap sample comprises n sampled individuals, $x_1^*, x_2^*, ...x_n^*$. As such, individuals from the original bootstrap population x may contribute once, more than once or not at all to each bootstrap sample x^*.
4. A total of k bootstrap samples are generated, $x^{*1}, x^{*2}, ...x^{*k}$, and the statistic of interest (e.g. sample mean $\bar{x}$) is estimated in each individual sample, $\bar{x}^{*i}$, giving the complete set of $\bar{x}^{*1}, \bar{x}^{*2}, ...\bar{x}^{*i}...\bar{x}^{*k}$.
5. The set of k statistics are combined to form a "bootstrap distribution"; as expected from central limit theorem (CLT)[23], this is typically closer to a normal distribution than the underlying distribution of values in either the bootstrap population x or the total population X. (See Figure 3 for an example of this)
6. The final values are derived as follows:

    - the parameter estimate (e.g. estimate of the true population mean, $\hat{\mu}$) is taken as the mean of the bootstrap distribution of k estimates, $(\sum_{i=1}^{k} \bar{x}) \div n$
    - the CIs are taken as the values at the appropriate centiles at the edges of the sampling distribution, e.g. a 95% CI would be generated using values at the 2.5th and 97.5th centiles
    - the SE of the estimate is taken as the standard deviation (SD) of the sampling distribution, given by $\sqrt{\frac{1}{k-1}\sum_{i=1}^{k}(\bar{x_i} - \hat{\mu})^2}$

B.2 Bootstrapping - Example: Prostate Volume

The above process is illustrated in 3. Data on prostate volume in 307 prostate cancer patients demonstrates a right-skewed distribution (A). An empirical distribution from a sample of 100 of these patients mirrors this right skew, and is used as the "bootstrap population" (B) for further re-sampling. As the bootstrap population is re-sampled more and more times, the "bootstrap distribution" of the sample means generated (C and D) gradually tends towards a normal distribution. The 95% CI is given by the bounds defining the middle 95% of the bootstrap distribution of estimated means, as shown.

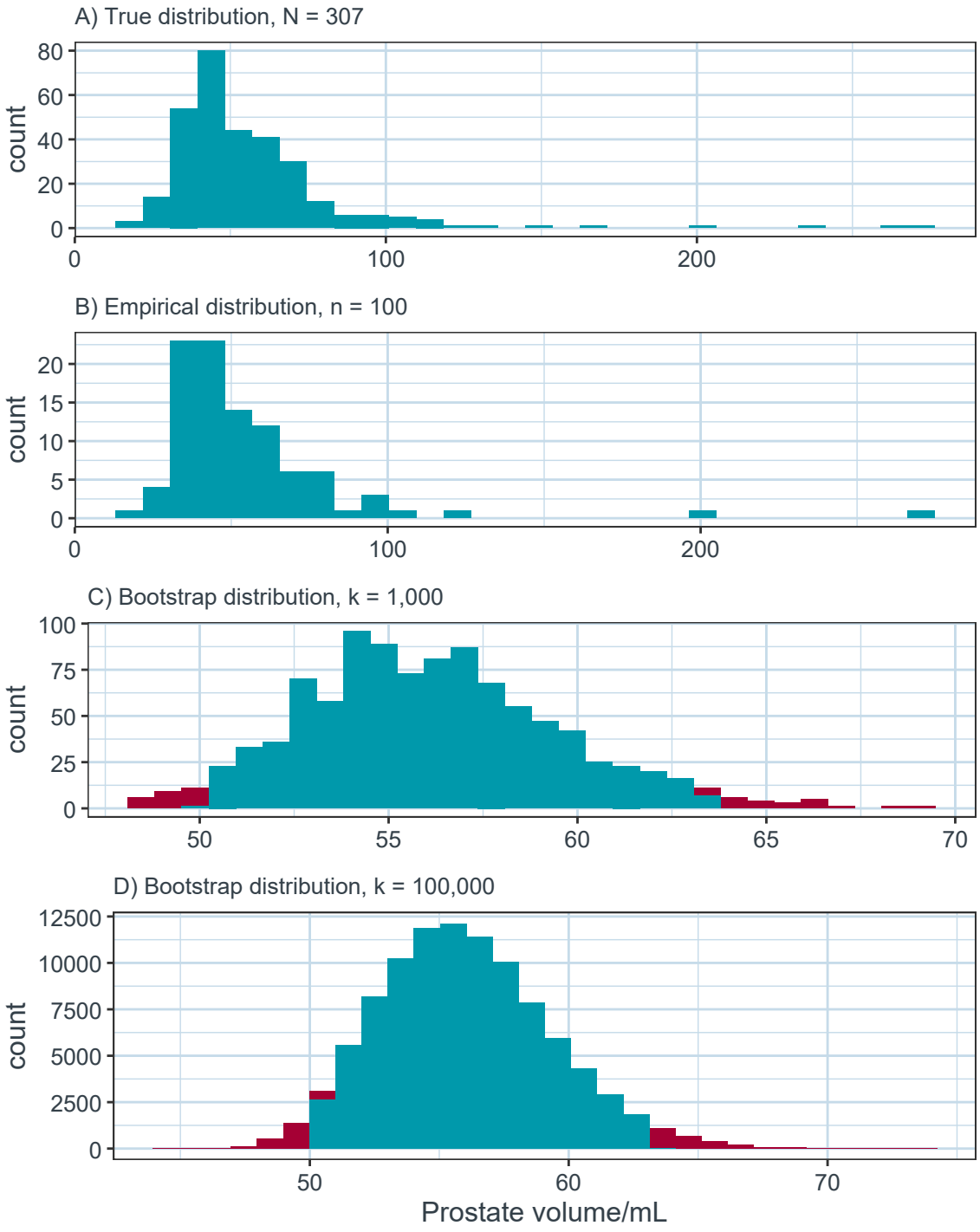


Figure 3: Histograms demonstrating distribution of prostate volumes in patients with prostatic cancer, taken from Cata et al 2011[24] via the R package `medicaldata`[25]. A) Distribution from whole study population of 307 patients with non-missing data, exhibiting right-skew. B) Distribution from random sample of 100 patients, still exhibiting right-skew. C) Bootstrap distribution generated by re-sampling 1,000 bootstrap samples from the original sample of 100 patients, right-skew less apparent. D) Bootstrap distribution generated by re-sampling 100,000 bootstrap samples from the original sample of 100 patients, approaching normality. 95% confidence intervals are demonstrated in plots C and D by marking the 2.5th and 97.5th centiles.

B.3 Bootstrapping - Relevance to WME

In current implementations of WME, the WME estimate of the causal effect ($\hat{\beta}_{WME}$) is calculated as described in Bowden et al[8], and the 95% CI is generated separately using bootstrapping, though notably not using the method described above.

The bootstrapping process begins similarly, with re-sampling undertaken (a default of $k = 1000$ times) to generate k bootstrap samples $x^{*1}, x^{*2}, ...x^{*k}$. Each individual bootstrap sample x^{*i} is used to estimate the causal effect using the WME method $\hat{\beta}^{*i}_{WME}$, and thus a bootstrap distribution of k values of WME is created, $\hat{\beta}^{*1}_{WME}, \hat{\beta}^{*2}_{WME}....\hat{\beta}^{*k}_{WME}$.

At this stage, however, the bootstrap distribution is then assumed to be approximately normally distributed without verifying this assumption. The 95% CI of the bootstrap estimate is then calculated as 1.96 SDs of the bootstrap distribution either side of the mean estimate, i.e. $\hat{\beta}_{WME} \pm 1.96 \times SE$. This approach may be problematic for several reasons.

Although CLT leads us to expect that the bootstrap distribution will approach normality as the number of bootstrap iterations k increases, the extent to which this occurs for a given k may depend on the inital distribution of values in the population X, and so also on the distribution in the sample/bootstrap population x. If the true distribution of values is very non-normal, as may be the case for traits determined by complex genetic and environmental influences, it may take relatively more bootstrap iterations for the bootstrap distribution to become sufficiently normal to assume mean and SD accurately describe it.

Additionally, the bootstrap SE is inversely proportional to the number of bootstrap iterations k, as opposed to the usual standard error (given by $SE = \frac{SD}{\sqrt{n}}$), which is inversely proportional to the square root of the sample size n. It is therefore possible to generate smaller SEs by increasing the number of bootstrap samples obtained. This may lead to false confidence in estimates generated despite potential issues with initial sample x, e.g. if it too small, or sampled in such a way that it is not representative of the underlying population X. Although such issues are inherent to any bootstrapping approaches, the usual method of generating bootstrapped CIs detailed above uses more information (i.e. using the entire bootstrap distribution) to generate these values than the parameter-based $estimate \pm 1.96 \times SE$ method (i.e. using approximate summary statistics to represent the distribution). The usual method of bootstrap CI generation may therefore be expected to highlight any variation or uncertainty present more readily than the parameter-based approach; this would be represented as wider CIs.

C Appendix: Simulation Code

C.1 Generating Data and Models

The data generating model used was from Appendix 3 of Bowden et al[8]; the relevant section describing their model is reproduced below:

> "...
>
> $$U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \epsilon_i^U \tag{6}$$
>
> $$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_i^X \tag{7}$$
>
> $$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \beta X_i + U_i + \epsilon_i^Y \tag{8}$$
>
> *for participants indexed by* $i = 1, ..., N$*, and genetic instruments indexed by* $j = 1, ..., J$*.*
>
> *The error terms* $\epsilon_i^U, \epsilon_i^X$ *and* ϵ_i^Y *were each drawn independently from standard normal distributions. The genetic effects on the exposure* j *are drawn from a uniform distribution between 0.03 and 0.1. Pleiotropic effects* α_j *and* ϕ_j *were set to zero if the genetic instrument was a valid instrumental variable. Otherwise (with probability 0.1, 0.2, or 0.3):*
>
> *1. In Scenario 1 (balanced pleiotropy, InSIDE satisfied), the* α_j *parameter was drawn from a uniform distribution between* -0.2 *and 0.2.*
>
> *2. In Scenario 2 (directional pleiotropy, InSIDE satisfied), the* α_j *parameter was drawn from a uniform distribution between 0 and 0.2.*
>
> *3. In Scenario 3 (directional pleiotropy, InSIDE not satisfied), the* ϕ_j *parameter was drawn from a uniform distribution between* -0.2 *and 0.2.*
>
> *The causal effect of the exposure on the outcome was either* $\beta X = 0$ *(null causal effect) or* $\beta X = 0.1$ *(positive causal effect). A total of 10 000 simulated datasets were generated for sample sizes of N = 10 000 and 20 [sic] participants. Only the summary data, that is genetic associations with the exposure and with the outcome and their standard errors as estimated by univariate regression on the genetic instruments in turn, were used by the analysis methods. In the two-sample setting, data were generated on 2N participants, and genetic associations with the exposure were estimated in the first N participants, and genetic associations with the outcome in the second N participants."*[8]

To reproduce this model, code was written in R to generate the relevant participant level data. First, a function (`get_simulated_MR_data`) was written which included parameters specified by Bowden et al, and also to allow testing of data simulation:

```r
# Define function to create data generating model
# Arguments/default values based on Bowden et al
get_simulated_MR_data <- function(n_participants = as.integer(),
                                  n_instruments = as.integer(),
                                  n_datasets = as.integer(),
                                  prop_invalid = 0.1,
                                  causal_effect = TRUE,
                                  balanced_pleio = TRUE,
                                  InSIDE_satisfied = TRUE,
                                  rand_error = TRUE,      # remove random errors, for testing
                                  two_sample = TRUE,      # 1- or 2-sample MR toggle, for testing
                                  beta_val = 0.1,         # size of causal effect
                                  allele_freq_min = 0.4,  # frequency of effect allele 0.01/0.99
                                  allele_freq_max = 0.6,  #?0.4/0.6
                                  gamma_min = 0.03,       # size of pleiotropic effects on exposure
                                  gamma_max = 0.1,
                                  alpha_min = -0.2,       # size of pleiotropic effects on outcome
                                  alpha_max = 0.2,
                                  phi_min = -0.2,         # size of additional pleiotropic effects
                                  phi_max = 0.2,          # when InSIDE not satisfied
                                  seed = 14101583){       # Set seed for reproducibility

  # Set seed to ensure comparability across scenarios
  set.seed(seed)

  # Initialise blank lists to receive datasets for
  # each of:
  #     U (vector: unmeasured confounding exposures per participant),
  #     X (vector: exposure:outcome associations estimated per participant)
  #     Y (vector: gene:outcome association estimated per participant),
  #     G (Matrices: Genotype data)
  #
  #     gamma (vector: pleiotropic effects of each instrument on exposure)
  #     alpha (vector: pleiotropic effects of each instrument on outcome)
  #     phi (vector: additional pleiotropic effects of each instrument when InSIDE
  #     assumption not satisfied)
  U_list <- list()
  X_list <- list()
  Y_list <- list()
  G_X_list <- list()
  G_Y_list <- list()

  gamma_list <- list()
  alpha_list <- list()
  phi_list <- list()


  n_participants_list <- list()
  n_instruments_list <- list()
  prop_invalid_list <- list()
  beta_val_list <- list()
```

```r
# --- Assign features common to all datasets --- #

# size of causal effect
beta <- if_else(causal_effect == TRUE,
                beta_val,
                0)

# create vector of participant indices for 1st n participants
# i.e. participants used for estimating gene:exposure coefficient
sample_1_ref <- 1:n_participants


# Default is to estimate gene:outcome coefficient from different sample
# to gene:exposure coefficient (i.e. simulating 2-sample MR)
# two_sample == FALSE toggles to single sample for testing simulation
ifelse(two_sample == FALSE,
       sample_2_ref <- sample_1_ref, # 1 sample MR
       sample_2_ref <- (n_participants+1):(2*n_participants)) # 2 sample MR

# --- Set characteristics for each genetic instrument --- #

# Set genetic effects of each instrument on the exposure,
# drawn from uniform distribution, min/max as per Bowden
# et al
gamma_vect <- runif(n = n_instruments,
                    min = gamma_min,
                    max = gamma_max)



# Set which instruments invalid, 0 = valid, 1 = invalid
invalid_instrument_vect <- rbinom(n = n_instruments,
                                  size = 1,
                                  prob = prop_invalid)


# Probability of effect allele set per dataset
# for each instrument, default value set at
# random between 0.01-0.99 (i.e. both effect +
# reference are common alleles)
allele_freq_vect <- runif(n = n_instruments,
                          min = allele_freq_min,
                          max = allele_freq_max)

# Set pleiotropic effects on outcome, Scenarios and
# min/max from Bowden et al
alpha_vect <- double() # Pleiotropic effects of instruments on outcome
phi_vect <- double() # Pleiotropic effects of confounders on outcome


for(j in 1:n_instruments){
  ifelse(invalid_instrument_vect[j] == 0, # alpha = 0 if valid
         alpha_vect[j] <- 0,
```

```
          ifelse(balanced_pleio == TRUE,
                 alpha_vect[j] <- runif(n = 1, # balanced
                                        min = alpha_min,
                                        max = alpha_max),
                 alpha_vect[j] <- runif(n = 1, # directional
                                        min = 0,
                                        max = alpha_max)
          )
  )


  # Assign default phi = 0 unless directional pleiotropy &
  # InSIDE assumption not satisfied & genetic instrument invalid
  if(balanced_pleio == FALSE & InSIDE_satisfied == FALSE){
    ifelse(invalid_instrument_vect[j] == 0,
           phi_vect[j] <- 0,
           phi_vect[j] <- runif(n = 1,
                                min = phi_min,
                                max = phi_max)
    )

  }
  else{
    phi_vect[j] <- 0
  }
}

# Re-set seed to ensure consistency across datasets
# N.B. above two if/ifelse statements cause de-sync
# of number of randomised functions between valid/invalid
set.seed(seed)

# --- Create separate datasets --- #

# Create N datasets by simulating genotype matrices with
# 1 row per participant, 1 column per genetic instrument
# Use these to estimate U, X + Y

for(n in 1:n_datasets){

  # --- Create matrix of genotypes --- #

  # Assign genotypes by sampling from binomial distribution
  # twice (as two alleles) per participant with probability
  # equal to frequency of effect allele
  # Create twice as many genotypes as participants in sample
  # to simulate 2 sample MR, i.e. first half used to estimate
  # Gene:Exposure, second half used to estimate Gene:Outcome

  # Matrix where columns are instruments, rows are participants
  # Values 0, 1 or 2
  # 0 = reference, i.e. zero effect alleles,
  # 1 = 1 effect allele, 2 = 2 effect alleles
```

```r
G_mat <- matrix(rbinom(n = 2 * n_participants * n_instruments,
                       size = 2,
                       prob = rep(allele_freq_vect, 2 * n_participants)),
                nrow = 2 * n_participants,
                ncol = n_instruments,
                byrow = TRUE)


# Create error terms for U, X + Y per participant,
# each drawn from standard normal distribution
# unless random error turned off (for testing)

ifelse(rand_error == TRUE,
       U_epsilon_vect <- rnorm(n = 2 * n_participants),
       U_epsilon_vect <- rep(0, 2 * n_participants))

ifelse(rand_error == TRUE,
       X_epsilon_vect <- rnorm(n = n_participants),
       X_epsilon_vect <- rep(0, n_participants))

ifelse(rand_error == TRUE,
       Y_epsilon_vect <- rnorm(n = n_participants),
       Y_epsilon_vect <- rep(0, n_participants))


# --- Combine Gene matrix/parameters to recreate model --- #

# Create vectors of estimates for U, X and Y per individual,
# i.e. Ui, Xi and Yi. Uses matrix inner product operator " %*%"
# https://stackoverflow.com/questions/22060515/the-r-operator
# http://matrixmultiplication.xyz/

#     U (vector: unmeasured confounding exposures per participant),
#     X (vector: exposure:outcome associations estimated per participant)
#     Y (vector: gene:outcome association estimated per participant)

Ui_vect <-  G_mat %*% phi_vect + U_epsilon_vect

Xi_vect <-  G_mat[sample_1_ref, ] %*% gamma_vect +
  Ui_vect[sample_1_ref, ] +
  X_epsilon_vect

Yi_vect <-  G_mat[sample_2_ref, ] %*% alpha_vect +
  beta * Xi_vect +
  Ui_vect[sample_2_ref, ] +
  Y_epsilon_vect


# Add vectors of estimates from this dataset to lists of
# estimates from all datasets
U_list[[n]] <- Ui_vect

X_list[[n]] <- Xi_vect
```

```
    Y_list[[n]] <- Yi_vect

    G_X_list[[n]] <- G_mat[sample_1_ref, ]

    G_Y_list[[n]] <- G_mat[sample_2_ref, ]


    # Include actual parameter values generated for simulation
    alpha_list[[n]] <- alpha_vect

    gamma_list[[n]] <- gamma_vect

    phi_list[[n]] <- phi_vect

    # Include inputs for reference/testing
    n_participants_list[[n]] <- n_participants
    n_instruments_list[[n]] <- n_instruments
    prop_invalid_list[[n]] <- prop_invalid
    beta_val_list[[n]] <- beta_val


  }



  # --- Combine all outputs to return --- #

  combined_list <- list(U = U_list,         # Estimates
                        X = X_list,
                        Y = Y_list,

                        G_X = G_X_list,     # Genotypes of 1st sample
                        G_Y = G_Y_list,     # Genotypes of 2nd sample

                        alpha = alpha_list, # Actual values for validating simulation
                        gamma = gamma_list,
                        phi = phi_list,

                        n_participants = n_participants_list, # Inputs
                        n_instruments = n_instruments_list,
                        prop_invalid = prop_invalid_list,
                        beta_val = beta_val_list
  )

  return(combined_list)
}
```

This initial simulation function generated data in the following format:

```
# Check data produced in expected format
#set.seed(1701)
test_data_sim <- get_simulated_MR_data(n_participants = 1000,
                                       n_instruments = 25,
```

```
                                        n_datasets = 2,
                                        prop_invalid = 0.3,
                                        rand_error = FALSE,
                                        causal_effect = TRUE,
                                        balanced_pleio = TRUE,
                                        InSIDE_satisfied = TRUE)

str(test_data_sim)
```

```
## List of 12
##  $ U             :List of 2
##   ..$ : num [1:2000, 1] 0 0 0 0 0 0 0 0 0 0 ...
##   ..$ : num [1:2000, 1] 0 0 0 0 0 0 0 0 0 0 ...
##  $ X             :List of 2
##   ..$ : num [1:1000, 1] 1.59 1.29 1.02 1.89 1.49 ...
##   ..$ : num [1:1000, 1] 1.85 1.84 2.18 1.16 1.44 ...
##  $ Y             :List of 2
##   ..$ : num [1:1000, 1] 0.0704 0.1351 0.1589 0.0944 0.0161 ...
##   ..$ : num [1:1000, 1] 0.59017 0.10039 0.00743 0.27896 0.27746 ...
##  $ G_X           :List of 2
##   ..$ : int [1:1000, 1:25] 2 0 1 2 1 0 1 0 0 2 ...
##   ..$ : int [1:1000, 1:25] 0 1 2 1 1 1 0 2 0 2 ...
##  $ G_Y           :List of 2
##   ..$ : int [1:1000, 1:25] 1 1 0 1 2 0 1 0 1 2 ...
##   ..$ : int [1:1000, 1:25] 1 1 1 1 0 1 2 0 2 0 ...
##  $ alpha         :List of 2
##   ..$ : num [1:25] 0 0 0.1157 0 -0.0634 ...
##   ..$ : num [1:25] 0 0 0.1157 0 -0.0634 ...
##  $ gamma         :List of 2
##   ..$ : num [1:25] 0.0938 0.0808 0.0755 0.0342 0.0443 ...
##   ..$ : num [1:25] 0.0938 0.0808 0.0755 0.0342 0.0443 ...
##  $ phi           :List of 2
##   ..$ : num [1:25] 0 0 0 0 0 0 0 0 0 0 ...
##   ..$ : num [1:25] 0 0 0 0 0 0 0 0 0 0 ...
##  $ n_participants:List of 2
##   ..$ : num 1000
##   ..$ : num 1000
##  $ n_instruments :List of 2
##   ..$ : num 25
##   ..$ : num 25
##  $ prop_invalid  :List of 2
##   ..$ : num 0.3
##   ..$ : num 0.3
##  $ beta_val      :List of 2
##   ..$ : num 0.1
##   ..$ : num 0.1
```

A function (`get_models`) was then written to create linear models from each dataset generated as per Bowden et al:

```
# Create plotting tibble with Mean/SD X + Y grouped by
# Dataset + instrument
get_models <- function(sim){
```

```r
output_list <- list()

# Create linear models per dataset to get coefficients
# for gene:exposure association (coeff_G_X) and gene:outcome
# association (coeff_G_Y)
for(dataset in 1:length(sim$X)){

  X <- sim$X[[dataset]]
  Y <- sim$Y[[dataset]]
  Instruments_X <- sim$G_X[[dataset]]
  Instruments_Y <- sim$G_Y[[dataset]]

  alpha <- sim$alpha[[dataset]]
  gamma <- sim$gamma[[dataset]]
  phi <- sim$phi[[dataset]]
  beta <- sim$beta_val[[dataset]]
  prop_invalid <- sim$prop_invalid[[dataset]]
  n_instruments <- sim$n_instruments[[dataset]]
  n_participants<- sim$n_participants[[dataset]]


  # Model for gene:exposure ***0 + deleted from X_lm and Y_lm***
  X_lm <- lm(X ~ 0 + Instruments_X)
  coeff_G_X_vect <- coef(summary(X_lm))[1:(ncol(Instruments_X)), 1]
  SE_coeff_G_X_vect <- coef(summary(X_lm))[1:(ncol(Instruments_X)), 2]

  R2_stat <- summary(lm(X ~ Instruments_X))$r.squared
  F_stat <- summary(lm(X ~ Instruments_X))$fstatistic[[1]]


  # Model for gene:outcome
  Y_lm <- lm(Y ~ 0 + Instruments_Y)
  coeff_G_Y_vect <- coef(summary(Y_lm))[1:(ncol(Instruments_Y)), 1]
  SE_coeff_G_Y_vect <- coef(summary(Y_lm))[1:(ncol(Instruments_Y)), 2]

  output_list[[dataset]] <- as_tibble(list(dataset = dataset,
                                           Instrument = c(1:ncol(Instruments_X)),
                                           coeff_G_X = coeff_G_X_vect,
                                           coeff_G_X_SE = SE_coeff_G_X_vect,
                                           gamma = gamma,
                                           F_stat = F_stat,
                                           R2_stat = R2_stat,
                                           coeff_G_Y = coeff_G_Y_vect,
                                           coeff_G_Y_SE = SE_coeff_G_Y_vect,
                                           alpha = alpha,
                                           phi = phi,
                                           beta = beta,
                                           prop_invalid = prop_invalid,
                                           n_instruments = n_instruments,
                                           n_participants = n_participants),
                                      .name_repair = "unique")
}
```

```
  return(output_list)

}
```

These models generated estimates of the coefficient of gene-exposure association (`coeff_G_X`), coefficient of gene-outcome association (`coeff_G_Y`), and the relevant standard errors of these estimates. The values of parameters inputted were also returned to aid in further testing of data/model generation, i.e. actual gene-exposure associations (`gamma`), pleiotropic effects of invalid instruments (`alpha`), additional pleiotropic effects when InSIDE assumption not satified (`phi`), causal effect of exposure on outcome (`beta`) and the proportion of invalid genetic instruments with pleiotropic effects on the outcome (`prop_invalid`).

```
test_extract_model <- get_models(test_data_sim)

summary(test_extract_model[[1]])
```

```
##     dataset    Instrument   coeff_G_X        coeff_G_X_SE
##  Min.   :1   Min.   : 1   Min.   :0.03419   Min.   :6.659e-17
##  1st Qu.:1   1st Qu.: 7   1st Qu.:0.05645   1st Qu.:6.807e-17
##  Median :1   Median :13   Median :0.06823   Median :6.873e-17
##  Mean   :1   Mean   :13   Mean   :0.06791   Mean   :6.889e-17
##  3rd Qu.:1   3rd Qu.:19   3rd Qu.:0.08594   3rd Qu.:6.935e-17
##  Max.   :1   Max.   :25   Max.   :0.09379   Max.   :7.313e-17
##      gamma             F_stat             R2_stat    coeff_G_Y
##  Min.   :0.03419   Min.   :6.224e+27   Min.   :1   Min.   :-0.109067
##  1st Qu.:0.05645   1st Qu.:6.224e+27   1st Qu.:1   1st Qu.: 0.004951
##  Median :0.06823   Median :6.224e+27   Median :1   Median : 0.006555
##  Mean   :0.06791   Mean   :6.224e+27   Mean   :1   Mean   : 0.013297
##  3rd Qu.:0.08594   3rd Qu.:6.224e+27   3rd Qu.:1   3rd Qu.: 0.008890
##  Max.   :0.09379   Max.   :6.224e+27   Max.   :1   Max.   : 0.162629
##   coeff_G_Y_SE          alpha                phi         beta      prop_invalid
##  Min.   :0.001550   Min.   :-0.115363   Min.   :0   Min.   :0.1   Min.   :0.3
##  1st Qu.:0.001579   1st Qu.: 0.000000   1st Qu.:0   1st Qu.:0.1   1st Qu.:0.3
##  Median :0.001591   Median : 0.000000   Median :0   Median :0.1   Median :0.3
##  Mean   :0.001598   Mean   : 0.006717   Mean   :0   Mean   :0.1   Mean   :0.3
##  3rd Qu.:0.001624   3rd Qu.: 0.000000   3rd Qu.:0   3rd Qu.:0.1   3rd Qu.:0.3
##  Max.   :0.001653   Max.   : 0.156224   Max.   :0   Max.   :0.1   Max.   :0.3
##  n_instruments n_participants
##  Min.   :25    Min.   :1000
##  1st Qu.:25    1st Qu.:1000
##  Median :25    Median :1000
##  Mean   :25    Mean   :1000
##  3rd Qu.:25    3rd Qu.:1000
##  Max.   :25    Max.   :1000
```

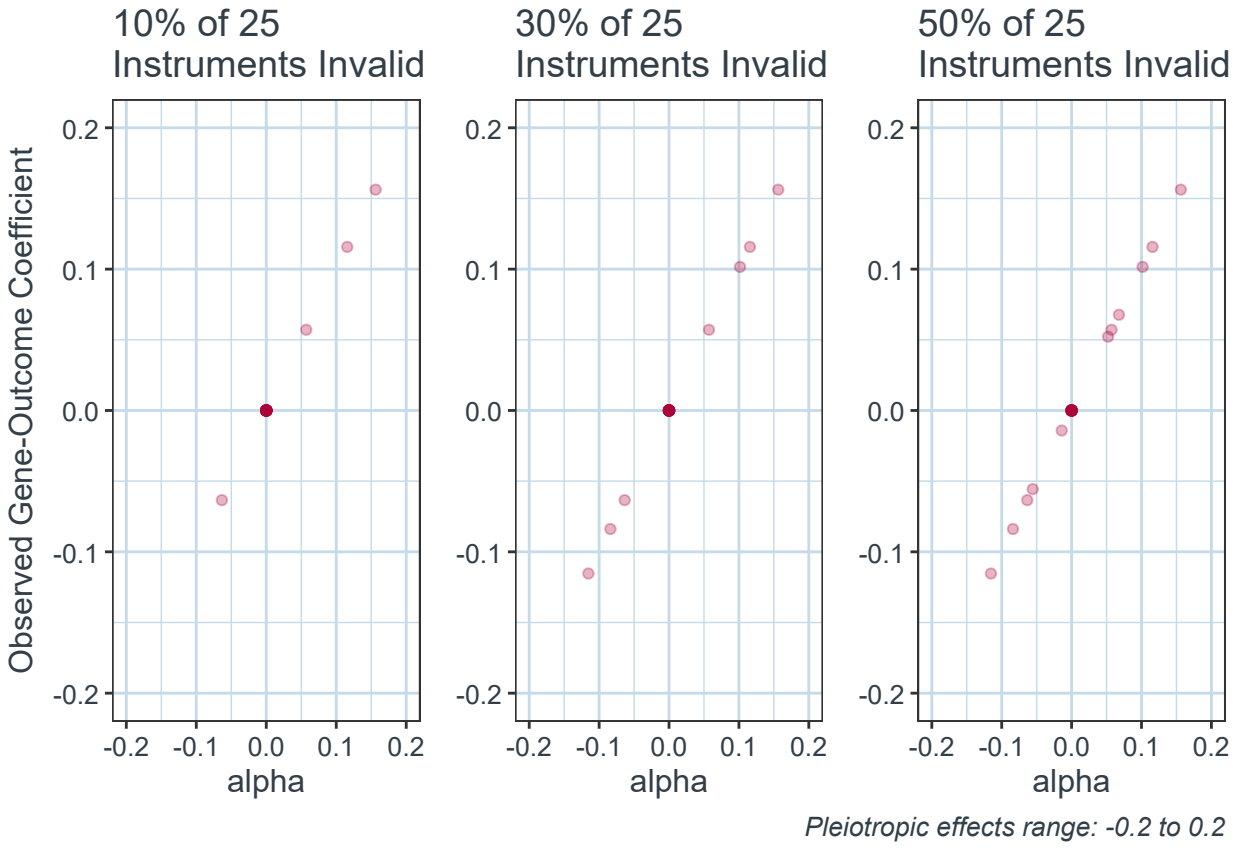
C.2 Testing Generation of Data and Models

A series of test plots were used to verify that data were simulated as intended under the various conditions specified by input parameters. Test plots were not created for the parameters `n_participants`, `n_instruments` or `n_datasets`, as the functioning of these parameters could be readily inferred from the structure of the datasets outputted, as above.

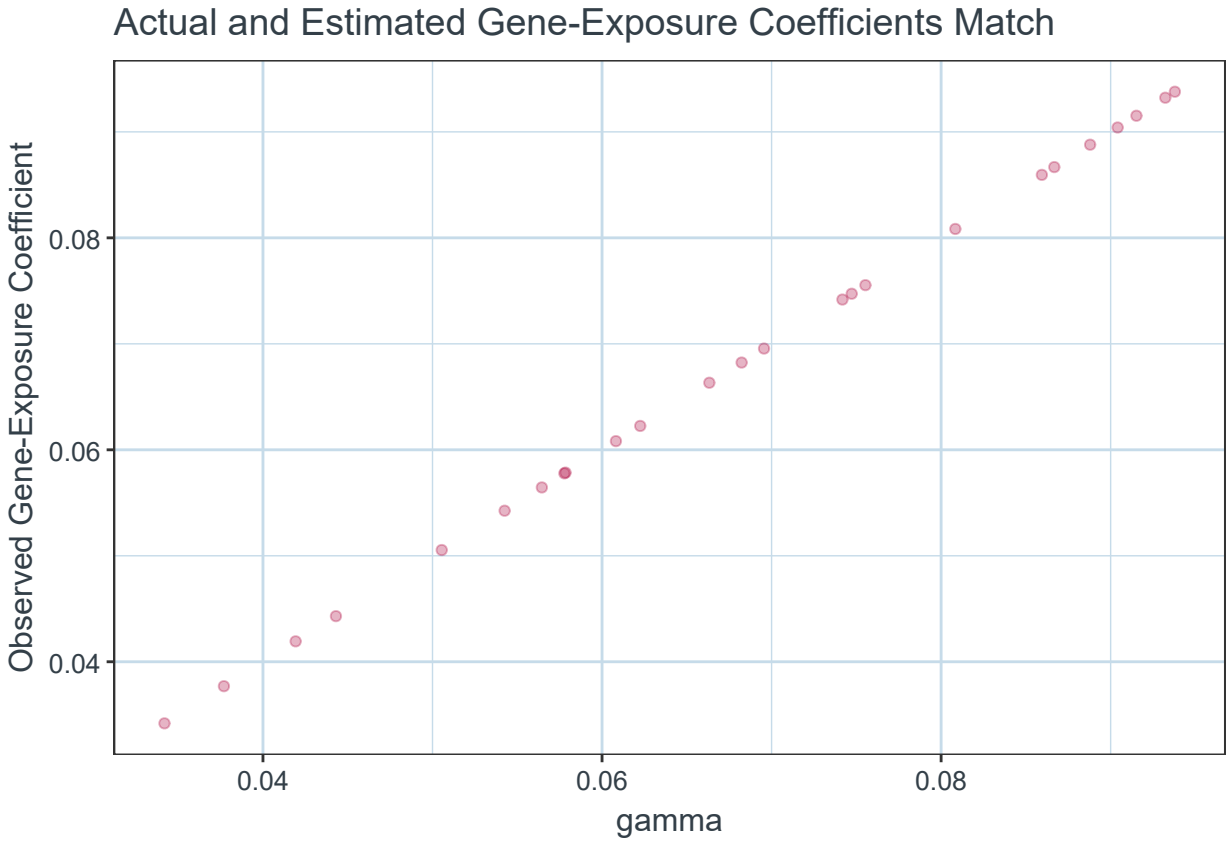
C.2.1 Proportion of Invalid Instruments

The `prop_invalid` parameter specifies the proportion of invalid genetic instruments simulated, i.e. the proportion of genetic instruments affecting the outcome via direct/pleiotropic effects, and thus not solely via the exposure of interest. If simulated correctly, increasing the value of `prop_invalid` should increase the number of instruments with pleiotropic effects, i.e. instruments with `alpha` $\neq 0$. With random error terms set to 0 and no causal effect present (i.e. `rand_error = FALSE` and `causal_effect = FALSE`), the estimated gene-outcome coefficient estimated using any given instrument will equal the pleiotropic effects of that instrument (i.e. `coeff_G_Y = alpha`), and therefore will only be non-zero for invalid instruments with non-zero pleiotropic effects on the outcome . Plotting `coeff_G_Y` against `alpha` for simulated data with no causal effect or random error should therefore yield a graph where

- For valid instruments: gene-outcome coefficient = alpha = 0
- For invalid instruments: gene-outcome coefficient = alpha $\neq 0$, with values spread uniformly between `alpha_min` and `alpha_max`



Similarly, with random error terms set to 0 (`rand_error = FALSE`) and no causal effect present (`causal_effect = FALSE`), gene-exposure coefficients estimated for each instrument should exactly match the actual values simulated, i.e. `coeff_G_X = gamma` for all instruments:

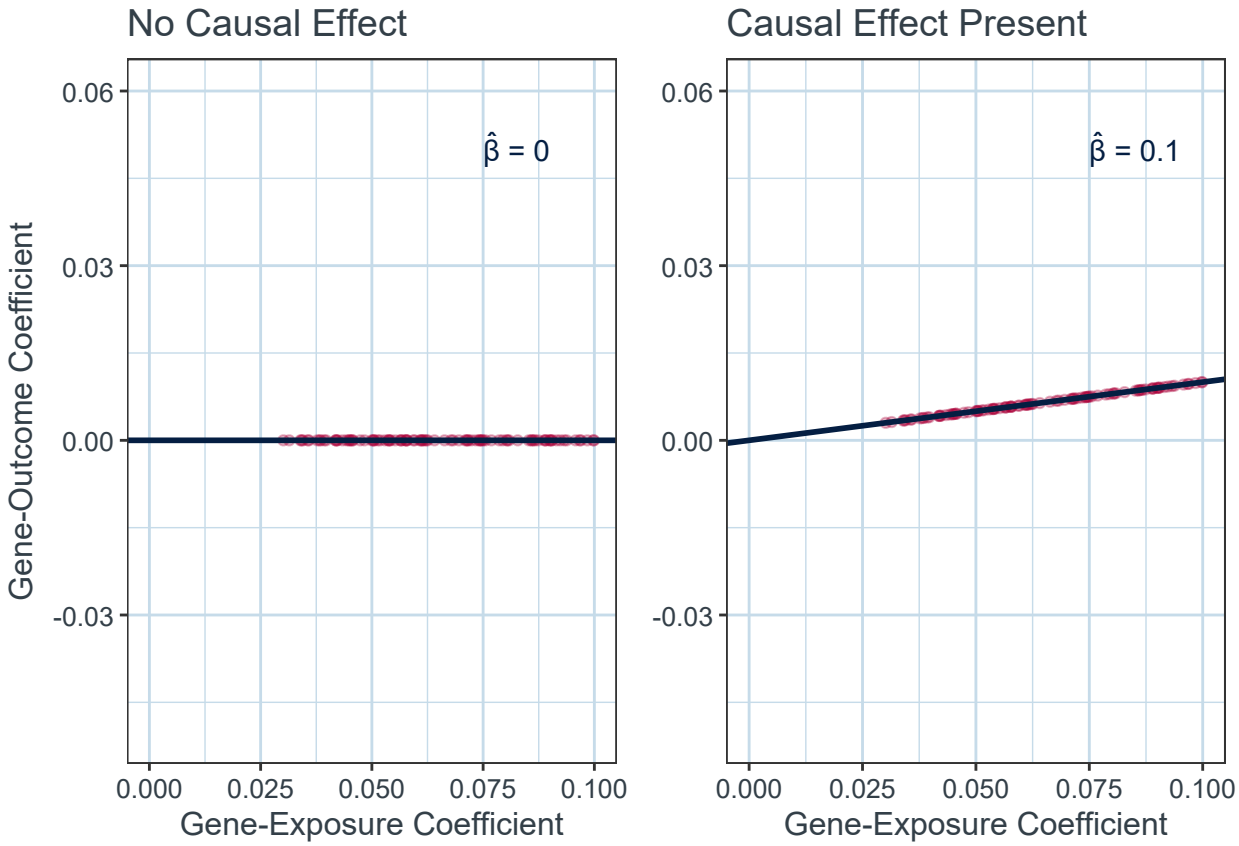


C.2.2 Gene-Exposure Coefficient Versus Gene-Outcome Coefficient Plots

For the next phase of testing, a function (`plot_GY_GX`) was written to plot the coefficients for gene-exposure versus gene-outcome as estimated using the previously created linear models:

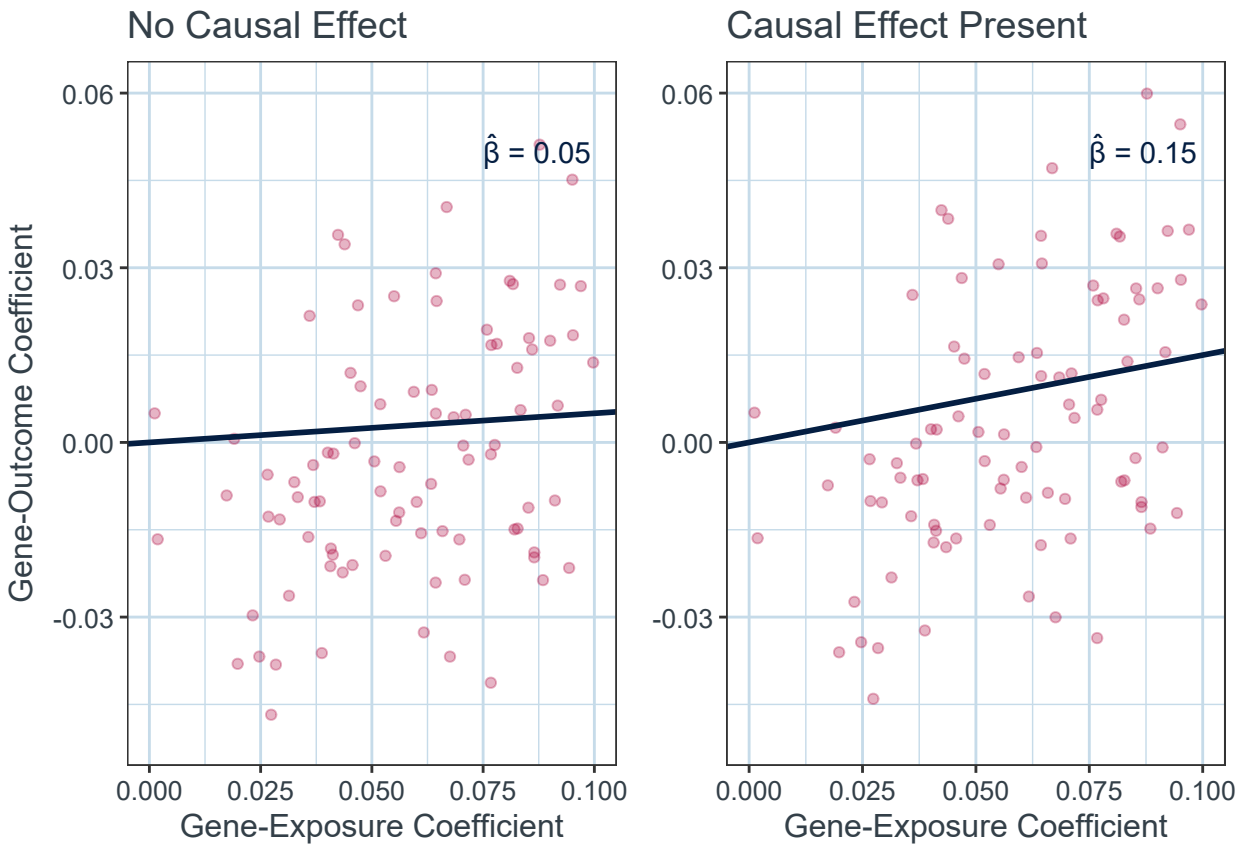
```
plot_GY_GX <- function(model_tib,
                       plot_title = as.character(NA),
                       x_min = 0,                      # set x-axis limits
                       x_max = 0.1,
                       y_min = -0.05,                  # set x-axis limits
                       y_max = 0.06,
                       beta_x = 0.075,                 # set beta-hat position
                       beta_y = 0.05,
                       hat_offset = 0.003
)
{

  model_tib %>%
    mutate(Gradient = round(coefficients(lm(coeff_G_Y ~ 0 + coeff_G_X)[1], 5),
                            digits = 2)) %>%
    plot_template() + # pre-formatted plot template - call to ggplot with UoE colours
    aes(x = coeff_G_X, y = coeff_G_Y) +
    geom_point(colour = edin_bright_red_hex, alpha = 0.3) +
    geom_abline(aes(intercept = 0,
                    slope = Gradient),
                size = 1,
                colour = edin_uni_blue_hex) +
    geom_text(aes(label = paste0("\U03B2 = ", as.character(Gradient))), #beta
              x = beta_x, # labels with gradient (causal effect estimate)
              y = beta_y,
              colour = edin_uni_blue_hex,
              hjust = 0,
              data = . %>% slice_head()# prevent over-printing
    ) +
    #label = expression("True" ~ hat(beta)~ "= 0.25"),
    annotate("text",
             x = beta_x,      # add hat to beta
             y = beta_y + hat_offset,
             label = paste("\U02C6"),
             colour = edin_uni_blue_hex,
             hjust = -0.4,
             vjust = 0.9
    ) +
    labs(title = plot_title,
         x = "Gene-Exposure Coefficient",
         y = "Gene-Outcome Coefficient") +
    xlim(x_min, x_max) +
    ylim(y_min, y_max)

}
```

With random error terms set to 0 (`rand_error = FALSE`) and no causal effect present, a graph of gene-exposure coefficients versus gene-outcome coefficients should be a straight line through the origin with gradient = 0; causal effect of $\beta = 0.1$ present (`beta_val = 0.1`, `causal_effect = TRUE`), the slope of a graph of gene-exposure coefficients versus gene-outcome coefficients from the same sample should be a straight line through the origin with gradient = 0.1:



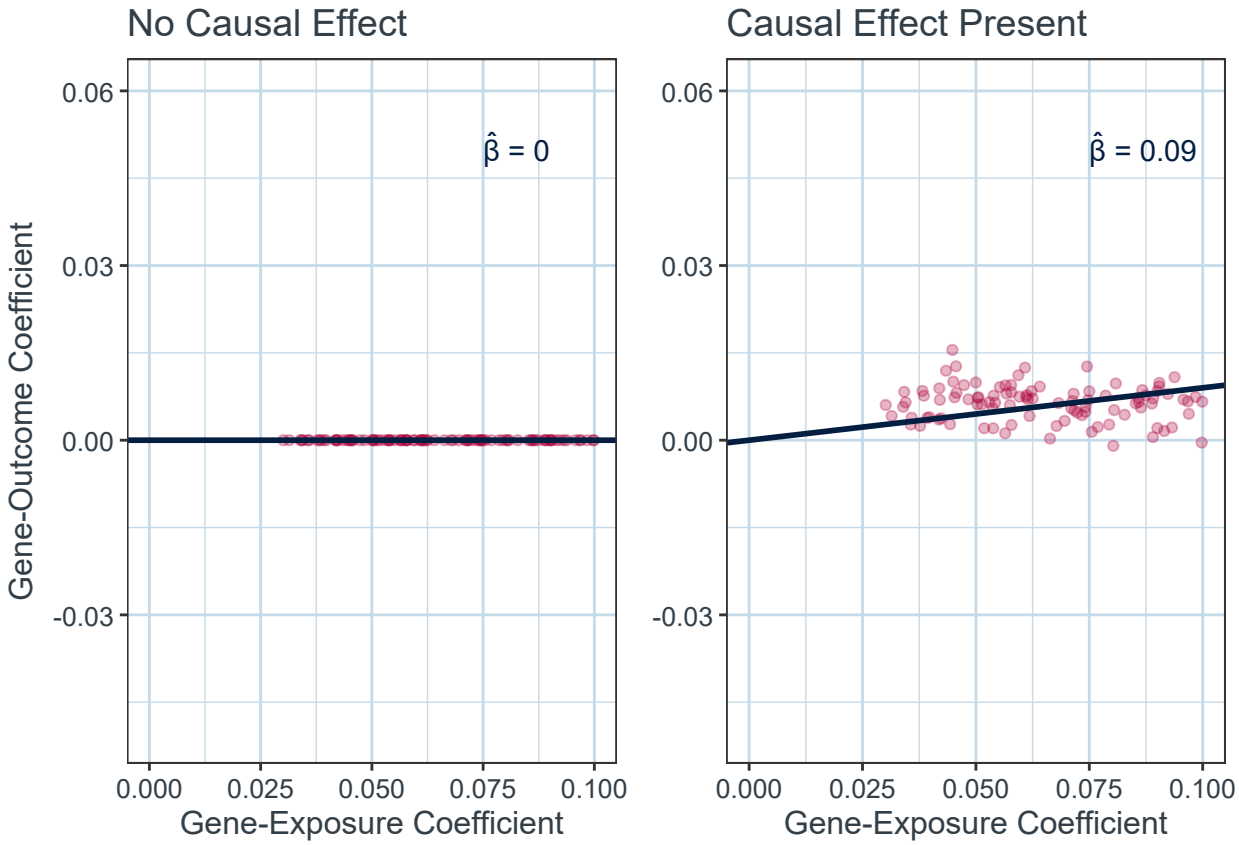
C.2.3 Random Errors

Re-plotting the same graphs with non-zero random error terms (`rand_error = TRUE`) should produce similar graphs with Gaussian spread around lines passing through the origin with gradients of 0 and 0.1 for no causal effect and causal effect, respectively:



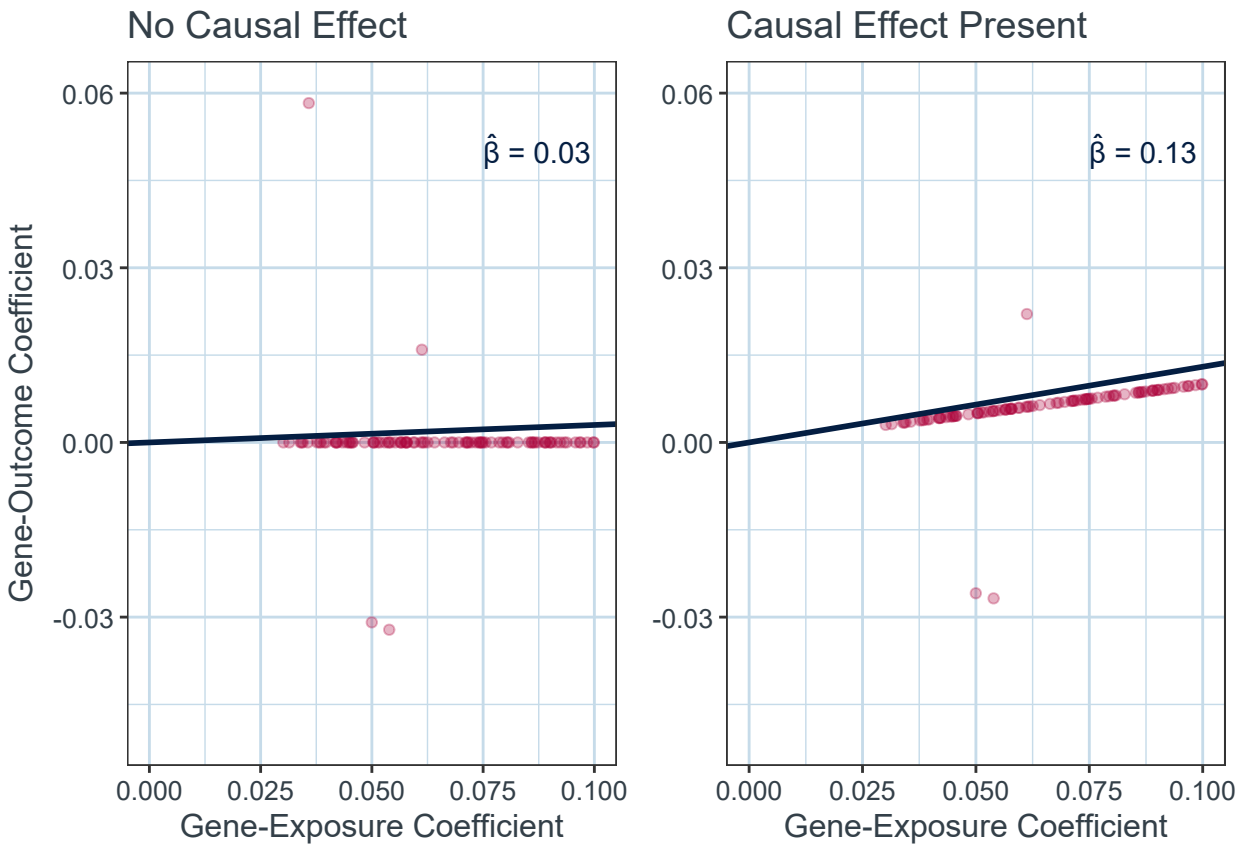
C.2.4 One versus Two Sample MR

Where gene-exposure coefficients and gene-outcome coefficients are estimated from two separate samples rather than one (i.e. `two_sample = TRUE`, simulating 2 sample MR), even with random error terms set to zero, error will be introduced into causal effect estimation through random sampling of different combinations of effect alleles. However, where a causal effect is not present, the effect estimated will consistently be zero regardless of the combinations of alleles sampled, so random error should not be introduced:



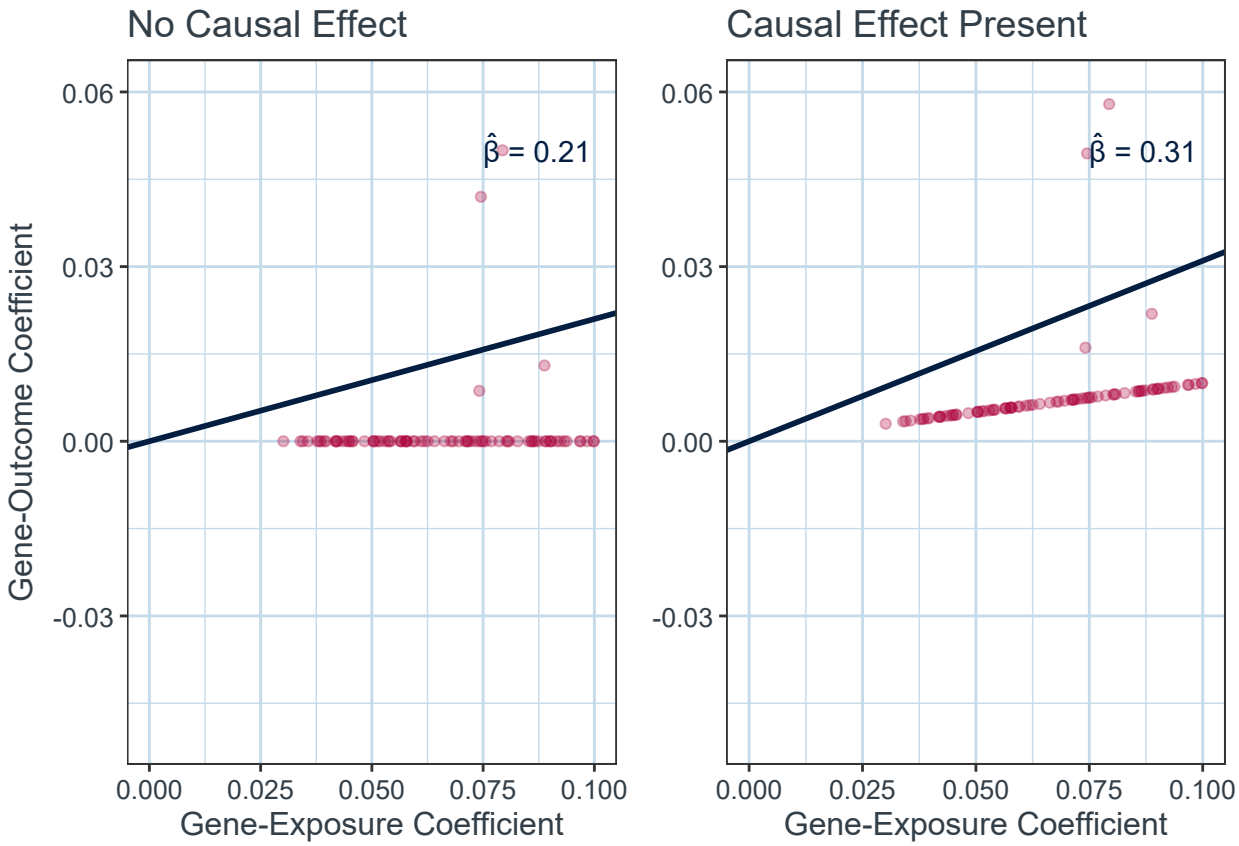
C.2.5 Invalid Instruments

Where invalid instruments are present (i.e. `prop_invalid` $\neq$ `0`) and random error terms are set to 0, graphs of gene-exposure coefficients versus gene-outcome coefficients should be straight lines through the origin and all points representing valid instruments; the invalid instruments should appear as outliers to this line:



C.2.6 Balanced Versus Directional Pleiotropy

Replotting the above with unbalanced pleiotropy present (`balanced_pleio = FALSE`), the invalid instruments should all appear as outliers in the positive direction, i.e. steepening the line of best fit and leading to overestimation of the causal effect:



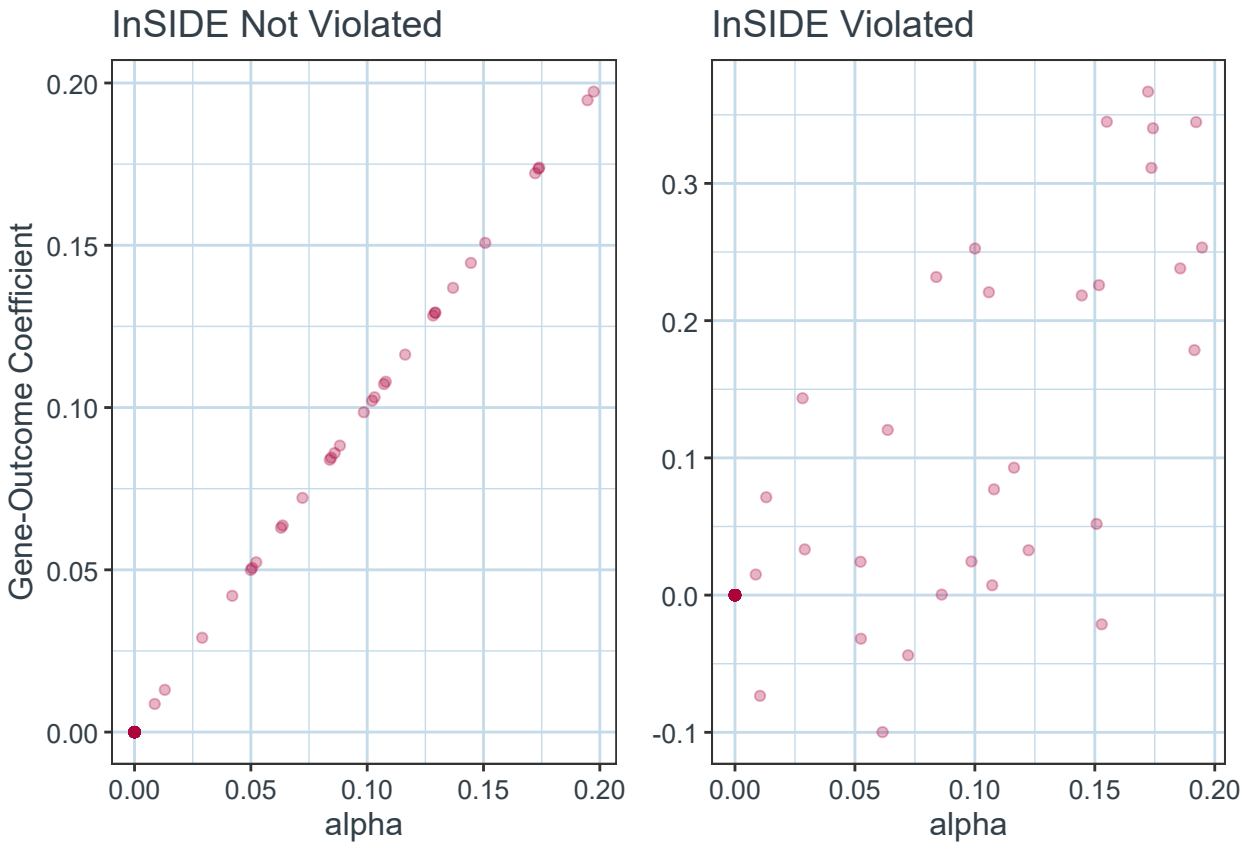
C.2.7 InSIDE Assumption and Phi

The variable phi represents additional pleiotropic effects of each invalid instrument when the InSIDE assumption is not satisfied. The InSIDE assumption states that the gene-exposure association is not correlated with the pleiotropic path gene-outcome path of any invalid genetic instruments. This assumption can be violated if e.g.:

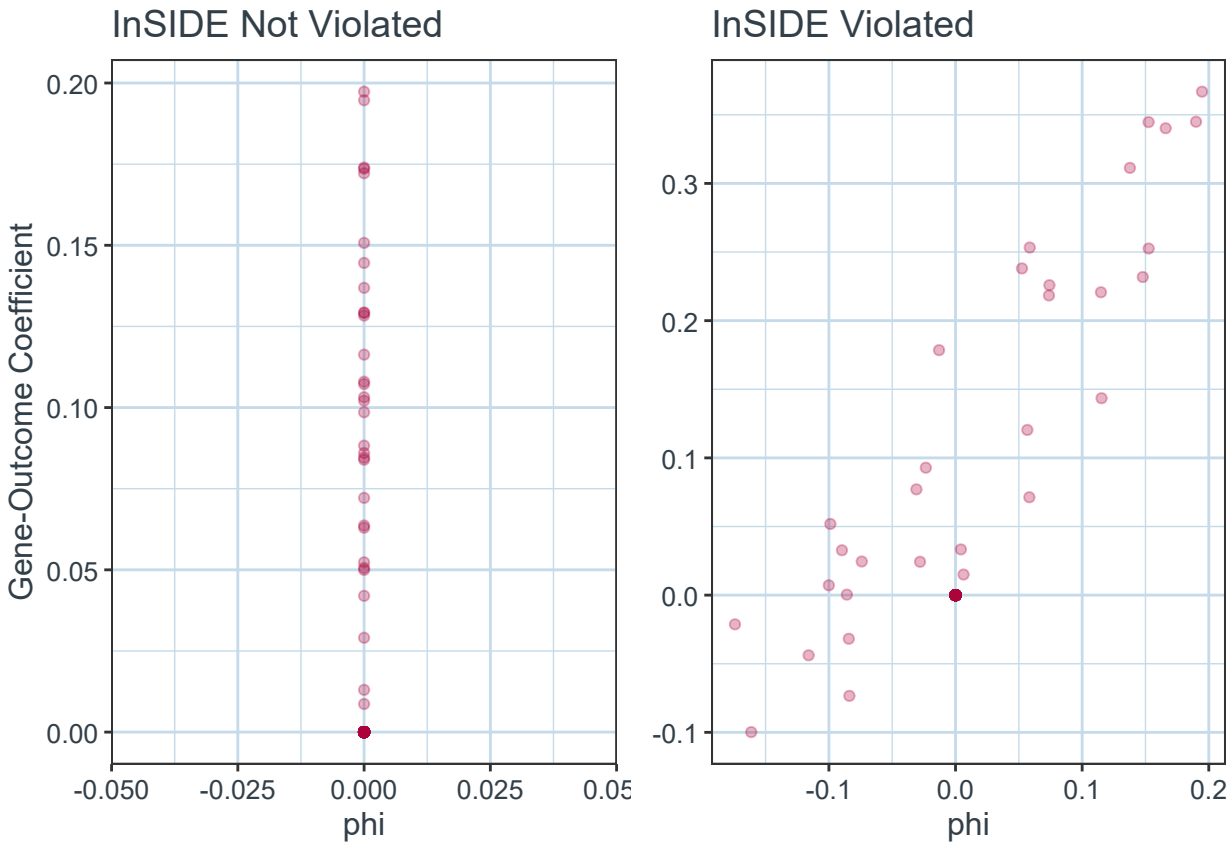
- several invalid genetic instruments influence the outcome via the same pleiotropic path
- several invalid genetic instruments are related to the same (unmeasured) confounders of the exposure:outcome relationship, aka correlated pleiotropy.

As such, when the InSIDE assumption is violated, even "strong" instruments (i.e. those with a strong gene-exposure relationship) may not allow accurate estimation of the true causal effect, as pleiotropic effects may scale with instrument strength. If pleiotropic effects are balanced, InSIDE assumption violation may lead to greater imprecision in causal effect estimation; if pleiotropic effects are directional, InSIDE assumption violation may lead to bias.

Bowden et al[8] modeled phi as the pleiotropic effects of unmeasured genetic confounders of the exposure:outcome relationship. Phi adds additional error to causal effect estimation in scenarios with directional pleiotropic effects (`0 < alpha < 0.2`) and InSIDE assumption violation. As such, switching `InSIDE_satisfied` from `TRUE` to `FALSE` should add scatter to the linear association expected when plotting alpha versus gene-outcome coefficients with random error terms set to zero:



Setting `InSIDE_satisfied = TRUE` should mean `phi = 0`; `InSIDE_satisfied=FALSE` should result in `phi` $\propto$ gene-outcome coefficient, with scatter only in the positive direction of gene-outcome coefficients given the model also requires directional pleiotropy before `phi` is used:



C.3 Summary Table

A function (`get_summary_MR_tib_row`) was written to take models generated from each simulated dataset, estimate causal effect using both weighted median and MR-Hevo methodologies, then output a summary formatted as per Tables 2 & 3 in Bowden et al[8]:

```
# Load WME functions
library(TwoSampleMR)

# Load RStan - needed for MR-Hevo
library(rstan)


# Run local copy of MR-Hevo functions
# Not using full package due to conflicts with Windows
source(here("Script", "Hevo", "functions.mrhevo.R"))

# Standard set-up for RStan
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE, save_dso = TRUE)


# Compile model for MR-Hevo
mr.stanmodel <- stan_model(file= here("Script",
                                      "Hevo",
                                      "MRHevo_summarystats.stan"),
                           model_name="MRHevo.summarystats",
                           verbose=FALSE,
                           save_dso = TRUE,
                           auto_write = TRUE)

get_summary_MR_tib_row <- function(model_list){


  # Create output tibble in same format as Table 2/3 from
  # Bowden et al
  output_tib_row <- tibble(N = as.integer(),
                           Prop_Invalid = as.double(),
                           F_stat = as.double(),
                           R2_stat = as.double(),
                           WME_Av = as.double(),
                           WME_SE = as.double(),
                           WME_Pos_Rate = as.double(),
                           Hevo_Av = as.double(),
                           Hevo_SE = as.double(),
                           Hevo_Pos_Rate = as.double())

  n_datasets <- length(model_list)

  # Create blank tibble to receive results of Weighted
  # Median Estimator function from MR-Base

  results_tib <-  tibble(WME_est = as.double(),
                         WME_se = as.double(),
```

```
                        WME_pval = as.double(),
                        WME_nsnp = as.integer(),
                        Hevo_est = as.double(),
                        Hevo_se = as.double(),
                        Hevo_sd = as.double(),
                        Hevo_est_lower_CI = as.double(),
                        Hevo_est_upper_CI = as.double(),
                        Hevo_causal_detected = as.logical()
)


# Run WME and MR-Hevo for each dataset
for(dataset in 1:n_datasets){

  # Stored as individual vectors for MR-Hevo/RStan - not
  # Tidyverse compatible
  coeff_G_X_vect <- model_list[[dataset]]$coeff_G_X
  coeff_G_Y_vect <- model_list[[dataset]]$coeff_G_Y
  coeff_G_X_SE_vect <- model_list[[dataset]]$coeff_G_X_SE
  coeff_G_Y_SE_vect <- model_list[[dataset]]$coeff_G_Y_SE
  prop_invalid <- min(model_list[[dataset]]$prop_invalid)
  F_stat <- min(model_list[[dataset]]$F_stat)
  R2_stat <- min(model_list[[dataset]]$R2_stat)
  n_instruments <- max(model_list[[dataset]]$Instrument)
  n_participants <- min(model_list[[dataset]]$n_participants)


  # N.B. MR-Hevo terminology vs WME paper/other code:
  # alpha = effects of instruments on exposure, i.e. coeff_G_X
  # beta = pleiotropic effects of instruments on outcome, i.e. alpha in WME
  # gamma = effects of instruments on outcome, i.e. coeff_G_Y
  # theta = causal effect X on Y, i.e. b

  # Results from weighted median estimator method
  WME_results <- mr_weighted_median(b_exp = coeff_G_X_vect,
                                    b_out = coeff_G_Y_vect,
                                    se_exp = coeff_G_X_SE_vect,
                                    se_out = coeff_G_Y_SE_vect,
                                    parameters = list(nboot = 1000))

  # Results from MR-Hevo method
  Hevo_results<- run_mrhevo.sstats(alpha_hat = coeff_G_X_vect,
                                   se.alpha_hat = coeff_G_X_SE_vect,
                                   gamma_hat = coeff_G_Y_vect,
                                   se.gamma_hat = coeff_G_Y_SE_vect) %>%
   summary()


  # Extract WME Results
  results_tib[dataset, ]$WME_est <- WME_results$b
  results_tib[dataset, ]$WME_se <- WME_results$se
  results_tib[dataset, ]$WME_pval <- WME_results$pval
  results_tib[dataset, ]$WME_nsnp <- WME_results$nsnp
```

```
    # Extract MR-Hevo Results
    results_tib[dataset, ]$Hevo_est <- Hevo_results$summary["theta","mean"]
    results_tib[dataset, ]$Hevo_se <- Hevo_results$summary["theta","se_mean"]
    results_tib[dataset, ]$Hevo_sd <- Hevo_results$summary["theta","sd"]
    results_tib[dataset, ]$Hevo_est_lower_CI <- Hevo_results$summary["theta","2.5%"]
    results_tib[dataset, ]$Hevo_est_upper_CI <- Hevo_results$summary["theta","97.5%"]

  }

  # Add causality Boolean to MR-Hevo
  results_tib <- results_tib %>%
   mutate(Hevo_est_causal_detected = (Hevo_est_lower_CI > 0  | Hevo_est_upper_CI < 0))


  output_tib_row <- results_tib %>%
    summarise(N = n_participants,
              Prop_Invalid = prop_invalid,
              F_stat = mean(F_stat),
              R2_stat = mean(R2_stat),
              WME_Av = mean(WME_est),
              WME_SE = mean(WME_se),
              WME_Pos_Rate = length(WME_pval[WME_pval < 0.05]) / n_datasets,
              Hevo_Av = mean(Hevo_est),
              Hevo_SE = mean(Hevo_se),
              Hevo_Lower_CI = mean(Hevo_est_lower_CI),
              Hevo_Upper_CI = mean(Hevo_est_upper_CI),
              Hevo_Pos_Rate = sum(Hevo_est_causal_detected) / n_datasets
    ) %>%
    mutate(across(where(is.double), round, 3))

  return(output_tib_row)

}
```

```
test_tib_summ_MR_data <-  get_simulated_MR_data(n_participants = 10000,
                                                n_instruments = 25,
                                                n_datasets = 2,
                                                prop_invalid = 0.1,
                                                beta_val = 0.1,
                                                causal_effect = TRUE,
                                                rand_error = TRUE,
                                                two_sample = TRUE,
                                                balanced_pleio = TRUE,
                                                InSIDE_satisfied = TRUE)

test_tib_summ_MR_models <- get_models(test_tib_summ_MR_data)

test_tib_summ_MR_row <- get_summary_MR_tib_row(test_tib_summ_MR_models)
```

```
##
## CHECKING DATA AND PREPROCESSING FOR MODEL 'MRHevo.summarystats' NOW.
##
## COMPILING MODEL 'MRHevo.summarystats' NOW.
```

N	Prop_Invalid	F_stat	R2_stat	WME_Av	WME_SE	WME_Pos_Rate	Hevo_Av	Hevo_SE	Hevo_Lower_CI	Hevo_Upper_CI	Hevo_Pos_Rate
10000	0.1	14.739	0.036	0.091	0.09	0.5	0.121	0.001	-0.052	0.302	0

```
##
## STARTING SAMPLER FOR MODEL 'MRHevo.summarystats' NOW.

##
## CHECKING DATA AND PREPROCESSING FOR MODEL 'MRHevo.summarystats' NOW.
##
## COMPILING MODEL 'MRHevo.summarystats' NOW.
##
## STARTING SAMPLER FOR MODEL 'MRHevo.summarystats' NOW.
```

```
test_tib_summ_MR_row %>%
  kable() %>%
  kable_styling(latex_options="scale_down")
```

D Appendix: R Packages Used

D.1 Package Citations

This work was completed using R version 4.4.3[19] with the following R packages: acronymsdown v. 0.11.1[26], bookdown v. 0.43[27,28], car v. 3.1.3[29], cowplot v. 1.1.3[30], crayon v. 1.5.3[31], devtools v. 2.4.5[32], ggdag v. 0.2.13[33], gghighlight v. 0.4.1[34], grateful v. 0.2.12[35], grid v. 4.4.3[36], here v. 1.0.1[37], infer v. 1.0.8[38], kableExtra v. 1.4.0[39], knitr v. 1.50[40–42], matrixStats v. 1.5.0[43], medicaldata v. 0.2.0[25], parallel v. 4.4.3[44], rmarkdown v. 2.29[45–47], rstan v. 2.32.7[48], tidyverse v. 2.0.0[21], TwoSampleMR v. 0.6.16[49,50], wordcountaddin v. 0.3.0.9000[51].

D.2 Session Information

```
##  setting  value
##  version  R version 4.4.3 (2025-02-28 ucrt)
##  os       Windows 11 x64 (build 26100)
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  English_United Kingdom.utf8
##  ctype    English_United Kingdom.utf8
##  tz       Europe/London
##  date     2025-06-08
##  pandoc   3.4 @ C:/Program Files/RStudio/resources/app/bin/quarto/bin/tools/ (via rmarkdown)
##  quarto   NA @ C:\\PROGRA~1\\RStudio\\resources\\app\\bin\\quarto\\bin\\quarto.exe
```

```
## # A tibble: 22 x 5
##    package      ondiskversion loadedversion date       source
##    <chr>        <chr>         <chr>         <chr>      <chr>
##  1 acronymsdown 0.11.1        0.11.1        2025-06-07 Github (rchaput/acronyms~
##  2 bookdown     0.43          0.43          2025-04-15 CRAN (R 4.4.3)
##  3 cowplot      1.1.3         1.1.3         2024-01-22 CRAN (R 4.4.3)
##  4 dplyr        1.1.4         1.1.4         2023-11-17 CRAN (R 4.4.3)
##  5 forcats      1.0.0         1.0.0         2023-01-29 CRAN (R 4.4.3)
##  6 gghighlight  0.4.1         0.4.1         2023-12-16 CRAN (R 4.4.3)
##  7 ggplot2      3.5.2         3.5.2         2025-04-09 CRAN (R 4.4.3)
##  8 grateful     0.2.12        0.2.12        2025-04-30 CRAN (R 4.4.3)
##  9 here         1.0.1         1.0.1         2020-12-13 CRAN (R 4.4.3)
## 10 infer        1.0.8         1.0.8         2025-04-14 CRAN (R 4.4.3)
## 11 kableExtra   1.4.0         1.4.0         2024-01-24 CRAN (R 4.4.3)
## 12 lubridate    1.9.4         1.9.4         2024-12-08 CRAN (R 4.4.3)
## 13 medicaldata  0.2.0         0.2.0         2021-08-16 CRAN (R 4.4.3)
## 14 purrr        1.0.4         1.0.4         2025-02-05 CRAN (R 4.4.3)
## 15 readr        2.1.5         2.1.5         2024-01-10 CRAN (R 4.4.3)
## 16 rstan        2.32.7        2.32.7        2025-03-10 CRAN (R 4.4.3)
## 17 StanHeaders  2.32.10       2.32.10       2024-07-15 CRAN (R 4.4.3)
## 18 stringr      1.5.1         1.5.1         2023-11-14 CRAN (R 4.4.3)
## 19 tibble       3.2.1         3.2.1         2023-03-20 CRAN (R 4.4.3)
## 20 tidyr        1.3.1         1.3.1         2024-01-24 CRAN (R 4.4.3)
## 21 tidyverse    2.0.0         2.0.0         2023-02-22 CRAN (R 4.4.3)
## 22 TwoSampleMR  0.6.16        0.6.16        2025-06-05 https://mrcieu.r-univers~
```