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

B233241

September 2024 - July 2025

Contents

Acknowledgements

I would like to thank Dr Athina Spiliopoulou for her expertise, kindness, and infinite patience in supervising this dissertation. I thank Professor Paul McKeigue for his input regarding the background and methodology of the project. Finally, my deepest gratitude goes to my wife, for keeping all our plates spinning while I've been busy shouting at my laptop.

Statement of Originality & Contributions

I confirm that all work is my own except where indicated, and that all sources are clearly referenced. Parts of this dissertation were informed by my participation in a related project with my supervisor's research group, which included comparison of Mendelian randomisation causal estimation methods using similar methodology, i.e. data simulation and citation searching. I confirm that all simulation code, literature searches and analyses contained within this dissertation are solely my own work, produced under the appropriate guidance of my supervisor.

Word Count

Word count: 9995

1 Abstract

1.1 Background

Mendelian randomisation (MR) uses data from observational genetic studies to support causal inference between exposures and outcomes of interest. Pleiotropy - where genetic variants influence outcomes through multiple pathways - can bias MR causal estimates. Inadequate controls for pleiotropy likely contribute to high false-positive causal report rates across MR literature. MR-Hevo is a recently proposed MR methodology claiming superior handling of pleiotropy and fewer false-positive reports of causality versus the established weighted median estimator (WME) method.

1.2 Aims

To evaluate differences in causal effect estimates between WME versus MR-Hevo methods, and to establish whether these may alter conclusions drawn in real-world studies.

1.3 Methods

Outputs from each method were compared through parallel analysis of simulated data, following the published approach used to validate WME. Simulations represented plausible combinations of population parameters and assumption violations. To investigate differences between methods using real-world data, both were applied to a sample of ten highly-cited MR studies reporting WME causal effect estimates alongside sufficient data to allow replication.

1.4 Results

Using simulated data with null causal effect, MR-Hevo demonstrated lower false-positive report rates versus WME across all 24 combinations of parameters/assumption violations considered (mean false-positive report rate 0.41% versus 5.1%). Using simulated data with true causal effect, MR-Hevo demonstrated higher power to detect this, both on average (mean true-positive report rate 31% versus 28%) and in most cases considered (14 of 24). Cases with higher true- and false-positive report rates for WME versus MR-Hevo correlated with conditions biasing away from the null, suggesting MR-Hevo estimates may be more robust to assumption violations. Re-analysis of highly-cited MR studies found poor reproducibility of published WME estimates in 4 of 10 studies included. Causal effect estimates were similar in magnitude between MR-Hevo and WME; conclusions regarding presence of causality were consistent between both methods across all 10 studies.

1.5 Conclusions

Compared to WME, MR-Hevo exhibited lower false-positive report rates and less perturbation by assumption violations. Across published MR literature reporting a causal effect, re-analysis using MR-Hevo may change conclusions in a minority of cases. Future work should investigate the non-reproducibility of MR results observed.

Word count: 349

2 Introduction and Background

2.1 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. 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. Such genetic ran-

domisation should therefore enable valid causal inference, provided that assumptions of IV analysis are met[5].

2.2 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.

Typically, MR studies estimate the causal effect using several genetic instruments; the causal effect estimate derived from the jth instrument is denoted $\hat{\beta}_j$. Let:

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

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

be linear models for exposure and outcome, respectively, given a specific genetic instrument G_j, where:

- γ_0 and Γ_0 reflect intercept values corresponding to the predicted value for X and Y, respectively, when $G_j = 0$ (i.e. these are the predicted values for an individual carrying the non-effect allele 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, representing effects of variables not explicitly included in the model, e.g. confounders of the exposure-outcome relationship (U in the causal diagram), and genetic contributions outside of 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.:

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

These coefficients of gene-exposure and gene-outcome association ($\hat{\gamma}$ and $\hat{\Gamma}$) can be obtained from a genome-wide association study (GWAS), which quantifies associations between small genetic variations - known as single nucleotide polymorphism (SNP)s - and various phenotypes. Each genetic instrument selected from a GWAS 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 might be violating 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 coefficients of gene-outcome association ($\bar{\Gamma}$) are expressed as log odds ratios ($log(OR)$) for estimating the causal effect, and the causal effect estimate can be converted to an odds ratio (OR) through exponentiation, i.e.:
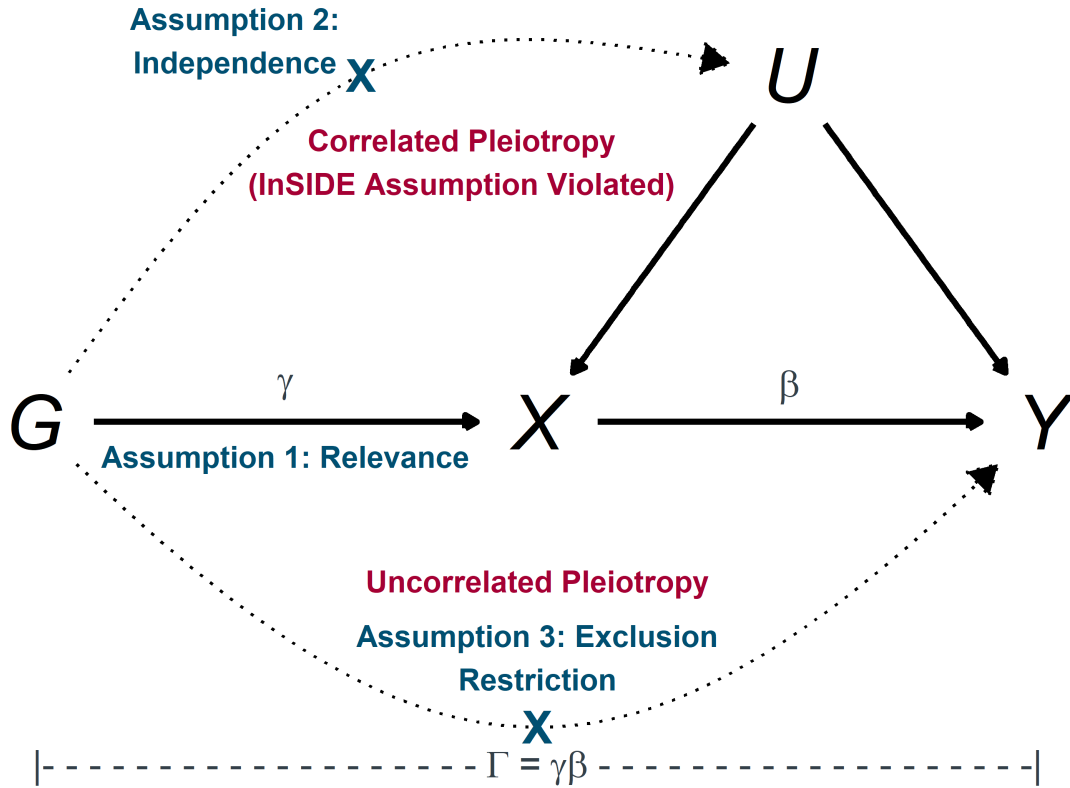
$$OR = e^{\hat{\beta}} \tag{5}$$

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]

2.3   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 their observed strength of association with exposures of interest in one or more GWAS. Sufficient gene-exposure association can be partly assured by selection using an appropriate genome-wide significance level (e.g. $p < 10^{-8}$) during this instrument selection. Statistical testing can also further 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 sample size under investigation[9]. An F-statistic of $\geq$ 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 determinants of the outcome,
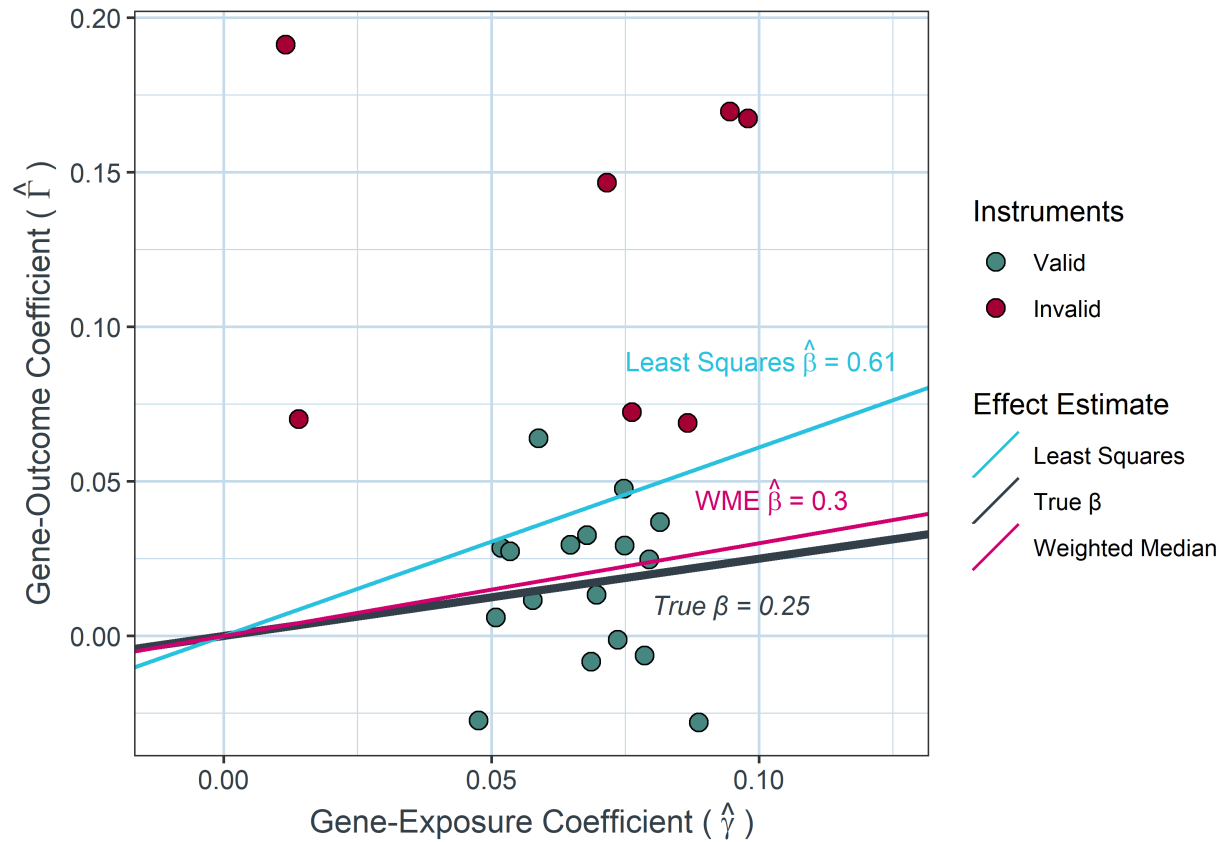
Figure 2: Simulated MR Study on 10,000 individuals using 25 genetic instruments, of which 30% are invalid (red points) and introduce directional pleiotropic effects. The true value of the exposure-outcome causal effect is 0.25 (grey line, causal effect represented by gradient). Regression using an unadjusted 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]

both known and unknown; 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 controls for population structure, 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, spurious exposure-outcome associations can be generated by confounding due to underlying differences in e.g. allele frequency, baseline disease risks etc. between ancestrally different populations; techniques such as principle components analysis can help control for such differences[9].

Exclusion restriction is a particularly troublesome issue in MR, due to so-called (horizontal) genetic pleiotropy, which is abundant among genetic variants associated with disease traits[10]. A genetic variant has "pleiotropic" effects if it influences several traits simultaneously through its involvement in multiple biological pathways; such alternative biological pathways - and the resulting pleiotropic effects - are often unknown. Pleiotropic effects open unmeasured causal pathways between a genetic instrument and the outcome (Figure 1), thus introducing bias in the MR estimate 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[11]. Where pleiotropic effects are in both positive and negative directions with a mean of zero - "balanced pleiotropy" - then they only add variance to the causal effect estimate, making inference more uncertain[12]. By contrast, "directional pleiotropy", where the mean of pleiotropic effects is non-zero, may introduce bias[8] (Figure 2). Note that, due to sampling error, the distribution of pleiotropic effects in a finite sample of genetic instruments may not reflect the true distribution across all possible genetic instruments for a given trait. For example, even where balanced pleiotropy is present across all genetic variants influencing a trait, bias could still be introduced into MR estimates of causal effect if a sample of instruments used happens (by chance) to have a non-zero mean pleiotropic effect on the outcome; the probability of this occuring reduces as the number of instruments used increases. In general, as the number of instruments used tends towards infinity, balanced pleiotropy will tend to introduce more variance than bias into causal effect estimates, and directional pleiotropy will tend to introduce bias more consistently than any additional variance it contributes.

If an additional causal pathway acts between gene G and outcome Y via a confounding factor U, then the magnitude of direct - and therefore 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 far more challenging to account for - potentially impossible using MR methods alone. Any attempt to statistically describe associations between

genes, exposures and outcomes in the presence of correlated pleiotropy will struggle to identify a unique model - different parameterisations where different proportions of overall gene-outcome association are subsumed into the confounder-outcome association may be equally plausible given available observational data. For this reason, several MR methods explicitly require an additional assumption of Instrument Strength Independent of Direct Effect (InSIDE), i.e no correlated pleiotropy to be present[13].

2.4   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. The median causal effect will provide a consistent effect estimate despite presence of pleiotropic effects, provided that <50% of included instruments are invalid. However, the simple median is an inefficient estimator - it weights estimates from all included instruments equally, without accounting for differences in precision of estimates obtained from each instrument, or by extension the factors influencing precision (e.g. sample size)[8].

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 the value of the causal effect across the range of values estimated based on each of the instruments. 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.

2.5   Issues With WME Confidence Intervals

WME calculation methods are available via several prolific MR tools: the R packages MendelianRandomization[14] and TwoSampleMR, and the MR-Base web platform[15]. WME is now commonly used either as the primary method or as a sensitivity analysis to account for pleiotropic effects in a large number of MR studies. WME's cited "breakdown level" of 50% invalid instruments suggests that, in the vast majority of cases, WME causal effect estimates should be robust to the presence of pleiotropic effects[8]. Despite this, a growing concern in the field of MR is high Type 1 error rates (i.e. frequently declaring a causal effect between exposures and outcomes where none is actually present)[16]. If WME causal estimates are indeed robust to pleiotropy

as claimed, another possible explanation for the method allowing null hypotheses to be falsely accepted would be if the confidence interval (CI) it reports for its causal estimates were inappropriately narrow.

CIs are a measure of the precision of an effect estimate, representing the range of values expected to contain the true effect given the observed variance of data around the observed average effect. Two key sources of uncertainty lead to variance in MR causal effect estimates:

1. Uncertainty around estimation of the gene-exposure and gene-outcome coefficients ($\hat{\gamma}$ and $\hat{\Gamma}$).

2. Uncertainty around the presence and magnitude of pleiotropic effects of included instruments.

All commonly available tools to calculate WME implement the original method for generating the 95% CI, using "bootstrapping" methodology to re-sample coefficients and causal estimates derived from each instrument. However, this approach only accounts for uncertainty from Point 1 (estimation of $\hat{\gamma}$ and $\hat{\Gamma}$), and may therefore under-estimate the total variance relevant to causal effect estimation. A fuller explanation of the theoretical issues with this approach, together with an overview of bootstrapping in general, is presented in Appendix B.

In brief, this method would be expected to lead to CIs which are too narrow, giving over-confidence in the causal effect estimates obtained, and therefore inflated Type 1 error rates. Given the widespread use of WME across MR literature, this issue with its CI generation could potentially have far-reaching consequences.

2.6   MR-Hevo

MR-Hevo is an R package implementing an alternative approach to MR causal estimation. It uses Bayesian methodology to estimate MR causal effects and corresponding CIs.

In general, Bayesian methods can estimate parameters in the following way[17]:

1. A "prior" is specified - this is a probability distribution which represents existing beliefs about likely values of that parameter (e.g. based on data from previous studies),
2. Further data is collected
3. The prior is updated based on the new data to form a "posterior" probability distribution, i.e. an improved representation of likely values of the parameter is created which incorporates the new information available

In this case, MR-Hevo directly addresses possible pleiotropy through use of a prior probability distribution which models pleiotropic effects of instruments on the outcome. The chosen prior distribution (a regularised

14

hierarchical horseshoe, building on the work of Piironen & Vehtari[18]) represents existing knowledge that most genetic instruments will not exert pleiotropic effects on the outcome, but some will, and these pleiotropic effects may be large. The R implementation of MR-Hevo then uses the probabilistic programming language, Stan, to directly sample the posterior probability distribution of pleiotropic effects on the outcome[19]. Essentially, this process involves using a computationally intensive, simulation based, "trial-and-error" approach using Markov chain Monte Carlo algorithms to perform inference over all model parameters and the available data[20,21].

Once the joint posterior distribution over all model parameters has been sampled, it is possible to obtain the marginal likelihood for the causal effect parameter given the data by dividing the posterior distributions by the prior. This computation marginalises over the pleiotropic effects, i.e. integrates over all possible values for these effects, thus removing their influence on other model parameters. This approach accounts for both main sources of uncertainty in MR studies, and also allows classical hypothesis testing for the maximum likelihood estimate of the causal effect, including calculation of a p-value from the marginal likelihood[19].

2.7   Aims and Objectives

The main aim of this study was to compare the outputs and resulting conclusions of MR-Hevo versus WME causal effect estimation methods in MR studies. In particular, this study aims 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 approach as its creators suggest. This will be achieved through addressing the research questions and objectives as outlined below:

Research Questions:

1. How do MR causal effect estimates from MR-Hevo differ versus the weighted median estimator?
2. Do conclusions of existing MR studies using weighted median causal effect estimation change if MR-Hevo methods are used?

Objectives:

1. Quantify the bias of MR-Hevo causal estimates for simulated data under differing sets of common assumptions, with reference to the weighted median estimator
2. Quantify the variance 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

3 Methods

3.1 Simulation Study

To establish differences in MR causal estimates of MR-Hevo relative to WME, the bias and variance of estimates from both methods were evaluated using simulated datasets with known parameter values.

3.1.1 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 Appendix C.

In brief, simulations were created based on three different scenarios, each representing a common set of assumptions about underlying data used for MR. Each scenario poses an increasing challenge to valid inference using any given MR causal estimation methodology than the previous assumption set. Without loss of generality, all instruments are simulated to have a positive effect on the exposure:

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 were simulated from a uniform distribution centered around zero, and were equally likely to be positive as negative.

2. Directional pleiotropy, InSIDE assumption satisfied - A proportion of invalid genetic instruments are present and introduce pleiotropic effects on the outcome which are in the positive direction only. The magnitude of these pleiotropic effects on the outcome is uncorrelated with the magnitude of instrument effects on the exposure.

3. Directional pleiotropy, InSIDE assumption not satisfied - A proportion of invalid genetic instruments are present and introduce pleiotropic effects on the outcome which are in the positive direction only. The magnitude of these pleiotropic effects on the outcome is correlated with the magnitude of instrument effects on the exposure through action via a confounder.

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 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 $n = 10,000$, then 20,000 genotypes were simulated - 10,000 for the cohort used to estimate gene-exposure association ($\widehat{\gamma}$), and a separate cohort of 10,000 used to estimate gene-outcome association ($\widehat{\Gamma}$). 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 alleles for each instrument via two draws from a binomial distribution, such that each simulated participant had 0, 1 or 2 effect alleles for each instrument.

3.1.2 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[15,19]. Results were aggregated per group of 1,000 simulated datasets corresponding to a particular combination of scenario and parameter values. The mean causal effect estimate, the mean standard errors/CIs of the causal effect estimate, and causality report rate (i.e. percentage of simulated studies reported as a non-null causal effect, with a 95% CI for the causal effect estimate not including 0) were reported for both WME and MR-Hevo for each combination of scenario/parameter values; these were computed over the 1,000 simulated MR studies using the same 25 genetic instruments in the same population.

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

3.2 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 reporting causal effect estimates using the WME method was re-analysed. A sample size of 10 published studies was pre-selected during study design as a pragmatic compromise between the scope of this project and the need to document frequency 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 some parameter/scenario combinations[8]. Therefore,

even this relatively small sample of 10 studies could plausibly demonstrate differences between methods if sufficient instruments with pleiotropic effects are present in included studies, and if the MR-Hevo approach is as appropriately conservative as its creators propose.

In an attempt to characterise the greatest possible impact of using MR-Hevo instead of WME (and thus potentially controlling Type 1 error better), 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, 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.

3.2.1   Citation Search

The Scopus search platform[22] 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[23]. 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 $\geq$ 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 potential 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.

3.3   Data Manipulation and Analysis

All simulations, data manipulations and data analyses were performed in R version 4.4.1 (2024-06-14 ucrt)[24].

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

For citation search data, a standardised data collection form in Microsoft Excel[25] was used 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[26]. A full list of packages used can be found in Appendix E.

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. Instruments missing values for association coefficients and/or standard error (SE)s were excluded from the main analysis, and were imputed as the most extreme values per dataset as a sensitivity analysis. It was noted during early testing that causal effect estimation functions did not operate correctly in the presence of zero-value coefficients of genetic association and/or their standard errors; such zero values were therefore re-coded as an arbitrarily low value of 10^{-100}.

3.4   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.

4 Results

4.1 Simulation Study

4.1.1 Data Simulation

Data were successfully simulated as intended. A selection of representative visualisations are presented in Figure 3. Full details of testing used to validate model outputs from parameter inputs are given in Appendix C.2. The minimum F-statistic calculated from simulated instruments across all simulations was >10, indicating that instruments were sufficiently strongly associated with exposure to meet the relevance assumption of IV analysis (Tables 1 and 2).

4.1.2 Analysis of Simulated Data

Due to a coding issue with outputs from MR-Hevo functions, the SEs for MR-Hevo causal estimates were not successfully retrieved; all 95% CIs were retrieved as intended. To aid comprehension and comparison of below results with regards to variance of causal effect estimates, it was decided to display an estimated SE for MR-Hevo. Estimated SE was calculated as $(Upper\ CI - Lower\ CI) \div (2 \times 1.96)$, i.e. approximating the bootstrap distribution as normally distributed per the expectation from central limit theorem (CLT) (see Appendix B). As such, whilst reported MR-Hevo CIs are accurate, MR-Hevo SEs are approximate only.

4.1.2.1 No Causal Effect

Across all cases where no causal effect was present (Table 1), the mean rate of reporting a causal effect (i.e. false-positive rate) for MR-Hevo was 0.41% versus 5.1% for WME. Of the 24 combinations of scenarios and parameters, MR-Hevo exhibited a favourable false-positive rate versus WME in 24 (100%).

Under Scenario 1 assumptions, the mean causal estimate (mean SE) across all parameter combinations was 0.018 (0.08) for MR-Hevo and 0.015 (0.075) for WME.

Under Scenario 2 assumptions, the mean causal estimate (mean SE) across all parameter combinations was 0.052 (0.082) for MR-Hevo and 0.031 (0.076) for WME.

Under Scenario 3 assumptions, the mean causal estimate (mean SE) across all parameter combinations was 0.05 (0.08) for MR-Hevo and 0.072 (0.078) for WME.

Overall, compared to WME, MR-Hevo estimates displayed slightly more bias away from the null in Scenarios 1 and 2, and slightly less in Scenario 3. MR-Hevo estimates exhibited fractionally more variance across all cases. For both methods, increasing proportion of invalid instruments tended to bias estimates away from the null and increase estimate variance, although the 30% invalid instrument cases in Scenarios 1 and 3
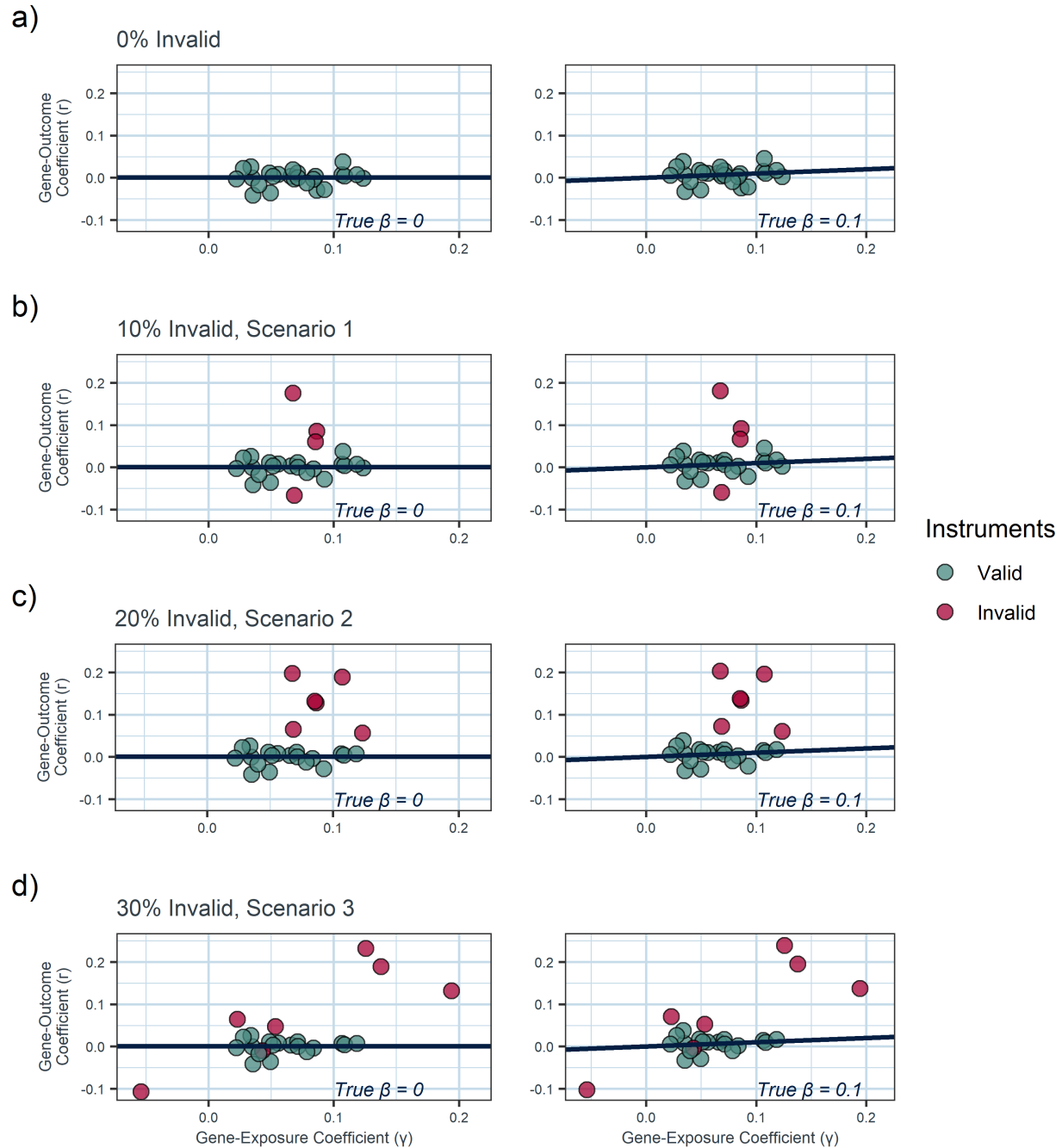
Figure 3: Plots of a representative group of simulated datasets; all simulate genetic instruments from the same index from the same random seed. Left and right columns demonstrate null and positive true causal effects, respectively; the true causal effect is represented by the gradient of the line shown. The scenario and the proportion of invalid (i.e. pleiotropic) genetic instruments changes with each row. a) 0% of instruments invalid, rendering scenario assumptions regarding invalid assumptions irrelevant. b) 10% of instruments invalid, Scenario 1: balanced pleiotropy simulated, though random sampling of these instruments has introduced some directionality; both variance and some bias are introduced to causal effect estimation. c) 20% of instruments invalid, Scenario 2: directional pleiotropy biases in the direction of the invalid instruments. d) 30% of instruments invalid, Scenario 3: directional pleiotropy and InSIDE assumption violation strongly biases towards a positive effect estimate in a manner difficult to statistically distinguish from a true causal effect.

did not follow this pattern. Increasing sample size tended to decrease both bias and variance of estimates, as expected.

4.1.2.2   Positive Causal Effect

Across all cases where a positive causal effect was present (Table 2), the mean rate of reporting a causal effect (i.e. true-positive rate) for MR-Hevo was 31% versus 28% for WME. Of the 24 combinations of scenarios and parameters, MR-Hevo exhibited a favourable true-positive rate versus WME in 14 (58%).

Under Scenario 1 assumptions, the mean causal estimate (mean SE) across all parameter combinations was 0.11 (0.08) for MR-Hevo and 0.087 (0.076) for WME.

Under Scenario 2 assumptions, the mean causal estimate (mean SE) across all parameter combinations was 0.14 (0.083) for MR-Hevo and 0.1 (0.077) for WME.

Under Scenario 3 assumptions, the mean causal estimate (mean SE) across all parameter combinations was 0.14 (0.082) for MR-Hevo and 0.15 (0.079) for WME.

Overall, MR-Hevo estimates displayed slight bias away from the null across all three scenarios, whereas WME estimates were more mixed. MR-Hevo estimates again demonstrated fractionally more variance on average than WME. For both methods, increasing proportion of invalid instruments tended to increase both size and variance of effect estimates, though again the 30% invalid instrument cases in Scenarios 1 and 3 did not follow this pattern. Increasing sample size tended to decrease both bias and variance of estimates, as expected.

Table 1: Summary of 1,000 simulated Mendelian randomisation studies per combination of scenario and parameters, all with null causal effect

N	Invalid IVs	F	R^2	Weighted Median			MR Hevo		
				Mean Estimate (Mean SE)	Mean 95% CI	Causal Report Rate	Mean Estimate (Mean SE*)	Mean 95% CI	Causal Report Rate
Scenario 1: Balanced pleiotropy, InSIDE assumption satisfied									
10,000	0%	11.7	2.8%	0.001 (0.078)	-0.15 to 0.15	0.2%	0.000 (0.061)	-0.12 to 0.12	0%
10,000	10%	11.7	2.8%	0.026 (0.086)	-0.14 to 0.19	1.5%	0.032 (0.085)	-0.13 to 0.2	0%
10,000	20%	11.7	2.8%	0.022 (0.092)	-0.16 to 0.2	2%	0.037 (0.106)	-0.17 to 0.25	0%
10,000	30%	11.7	2.8%	0.014 (0.093)	-0.17 to 0.2	1.6%	0.022 (0.116)	-0.2 to 0.25	0%
20,000	0%	26.2	3.2%	0.003 (0.056)	-0.11 to 0.11	0.3%	0.001 (0.044)	-0.09 to 0.09	0%
20,000	10%	24.5	3%	0.022 (0.062)	-0.1 to 0.14	0.5%	0.019 (0.063)	-0.1 to 0.14	0.1%
20,000	20%	24.5	3%	0.020 (0.067)	-0.11 to 0.15	1.3%	0.022 (0.077)	-0.13 to 0.18	0%
20,000	30%	24.5	3%	0.012 (0.067)	-0.12 to 0.14	0.8%	0.014 (0.084)	-0.15 to 0.18	0%
Scenario 2: Directional pleiotropy, InSIDE assumption satisfied									
10,000	0%	11.7	2.8%	0.001 (0.078)	-0.15 to 0.15	0.3%	0.000 (0.061)	-0.12 to 0.12	0%
10,000	10%	11.7	2.8%	0.020 (0.087)	-0.15 to 0.19	0.8%	0.039 (0.089)	-0.13 to 0.22	0%
10,000	20%	11.7	2.8%	0.050 (0.093)	-0.13 to 0.23	4.1%	0.098 (0.113)	-0.11 to 0.33	1.5%
10,000	30%	11.7	2.8%	0.066 (0.094)	-0.12 to 0.25	5.8%	0.126 (0.118)	-0.09 to 0.38	3.6%
20,000	0%	24.5	3%	0.004 (0.056)	-0.11 to 0.11	0.2%	0.001 (0.044)	-0.08 to 0.09	0%
20,000	10%	24.5	3%	0.016 (0.062)	-0.11 to 0.14	0.7%	0.021 (0.063)	-0.1 to 0.15	0.1%
20,000	20%	24.5	3%	0.038 (0.067)	-0.09 to 0.17	2.2%	0.054 (0.08)	-0.1 to 0.22	0.5%
20,000	30%	24.5	3%	0.050 (0.068)	-0.08 to 0.18	4.9%	0.076 (0.086)	-0.08 to 0.25	1.2%
Scenario 3: Directional pleiotropy, InSIDE assumption not satisfied									
10,000	0%	11.7	2.8%	0.001 (0.078)	-0.15 to 0.15	0.2%	0.000 (0.061)	-0.12 to 0.12	0%
10,000	10%	13.7	3.3%	0.077 (0.087)	-0.09 to 0.25	8%	0.044 (0.083)	-0.12 to 0.21	0.1%
10,000	20%	14.9	3.6%	0.144 (0.099)	-0.05 to 0.34	24.7%	0.107 (0.114)	-0.1 to 0.35	1.3%
10,000	30%	12.8	3.1%	0.103 (0.097)	-0.09 to 0.29	11.9%	0.102 (0.12)	-0.11 to 0.36	0.6%
20,000	0%	24.5	3%	0.004 (0.056)	-0.11 to 0.11	0.2%	0.001 (0.044)	-0.08 to 0.09	0%
20,000	10%	30.4	3.7%	0.061 (0.063)	-0.06 to 0.18	8.5%	0.030 (0.06)	-0.09 to 0.15	0.1%
20,000	20%	32.4	3.9%	0.111 (0.071)	-0.03 to 0.25	28.3%	0.060 (0.077)	-0.08 to 0.22	0.5%
20,000	30%	31.1	3.8%	0.079 (0.07)	-0.06 to 0.22	13.6%	0.058 (0.081)	-0.09 to 0.22	0.2%

CI: Confidence Interval, InSIDE: Instrument Strength Independent of Direct Effect, IV: Instumental Variable, SE: Standard Error, SE*: Estimated Standard Error. F and R² statistics presented are the minimum values across all simulated datasets.

Null Causal Effect (β = 0)

Table 2: Summary of 1,000 simulated Mendelian randomisation studies per combination of scenario and parameters, all with positive causal effect

N	Invalid IVs	F	R^2	Weighted Median			MR Hevo		
				Mean Estimate (Mean SE)	Mean 95% CI	Causal Report Rate	Mean Estimate (Mean SE*)	Mean 95% CI	Causal Report Rate
Scenario 1: Balanced pleiotropy, InSIDE assumption satisfied									
10,000	0%	11.7	2.8%	0.070 (0.079)	-0.08 to 0.22	4.9%	0.085 (0.061)	-0.04 to 0.21	6.2%
10,000	10%	11.7	2.8%	0.094 (0.087)	-0.08 to 0.26	11%	0.118 (0.086)	-0.05 to 0.29	12.6%
10,000	20%	11.7	2.8%	0.089 (0.093)	-0.09 to 0.27	10.3%	0.124 (0.107)	-0.08 to 0.34	5.6%
10,000	30%	11.7	2.8%	0.081 (0.094)	-0.1 to 0.27	8.7%	0.108 (0.116)	-0.12 to 0.34	1.6%
20,000	0%	24.5	3%	0.080 (0.056)	-0.03 to 0.19	21.3%	0.089 (0.044)	0 to 0.18	62.2%
20,000	10%	24.5	3%	0.098 (0.063)	-0.03 to 0.22	27.8%	0.108 (0.064)	-0.01 to 0.23	29.9%
20,000	20%	24.5	3%	0.095 (0.067)	-0.04 to 0.23	22.6%	0.113 (0.078)	-0.04 to 0.27	15%
20,000	30%	24.5	3%	0.088 (0.068)	-0.05 to 0.22	17.7%	0.104 (0.085)	-0.06 to 0.27	5.4%
Scenario 2: Directional pleiotropy, InSIDE assumption satisfied									
10,000	0%	11.7	2.8%	0.070 (0.079)	-0.08 to 0.22	5.3%	0.085 (0.061)	-0.04 to 0.21	5.9%
10,000	10%	11.7	2.8%	0.089 (0.088)	-0.08 to 0.26	9%	0.124 (0.091)	-0.05 to 0.31	11.9%
10,000	20%	11.7	2.8%	0.119 (0.094)	-0.07 to 0.3	17.7%	0.187 (0.116)	-0.02 to 0.43	32.3%
10,000	30%	11.7	2.8%	0.133 (0.095)	-0.05 to 0.32	23.3%	0.216 (0.121)	0 to 0.47	46.1%
20,000	0%	24.5	3%	0.080 (0.057)	-0.03 to 0.19	21.1%	0.089 (0.044)	0 to 0.18	62.7%
20,000	10%	24.5	3%	0.093 (0.063)	-0.03 to 0.22	24%	0.109 (0.064)	-0.01 to 0.24	29.1%
20,000	20%	24.5	3%	0.116 (0.068)	-0.02 to 0.25	35.3%	0.146 (0.082)	-0.01 to 0.31	41.2%
20,000	30%	24.5	3%	0.127 (0.069)	-0.01 to 0.26	40.7%	0.168 (0.088)	0.01 to 0.35	56.2%
Scenario 3: Directional pleiotropy, InSIDE assumption not satisfied									
10,000	0%	11.7	2.8%	0.070 (0.079)	-0.08 to 0.22	5.2%	0.085 (0.061)	-0.04 to 0.21	5.7%
10,000	10%	13.7	3.3%	0.150 (0.089)	-0.02 to 0.32	35%	0.137 (0.085)	-0.03 to 0.31	25.1%
10,000	20%	14.9	3.6%	0.213 (0.1)	0.02 to 0.41	55.8%	0.202 (0.118)	-0.01 to 0.46	45.2%
10,000	30%	12.8	3.1%	0.169 (0.099)	-0.03 to 0.36	37.1%	0.191 (0.123)	-0.03 to 0.46	29.1%
20,000	0%	24.5	3%	0.080 (0.057)	-0.03 to 0.19	21.5%	0.089 (0.044)	0 to 0.18	62.8%
20,000	10%	30.4	3.7%	0.144 (0.064)	0.02 to 0.27	66%	0.125 (0.061)	0.01 to 0.25	63%
20,000	20%	32.4	3.9%	0.189 (0.073)	0.05 to 0.33	81.5%	0.154 (0.079)	0.01 to 0.32	58.6%
20,000	30%	31.1	3.8%	0.153 (0.071)	0.01 to 0.29	60.3%	0.146 (0.083)	-0.01 to 0.32	41%

CI: Confidence Interval, InSIDE: Instrument Strength Independent of Direct Effect, IV: Instumental Variable, SE: Standard Error. F and R^2 statistics presented are the minimum values across all simulated datasets.

Positive Causal Effect (β = 0.1)

4.2 Re-Analysis of Published Data

4.2.1 Citation Search Results

A total of 103 abstracts and 54 full texts were screened to identify the 10 studies included[27–36]; these studies are summarised in Table 3. The flow diagram of study screening and selection is presented in Figure 4.

Identification

Studies from databases/registers **(n = 103)**
  Scopus (n = 103)

References removed **(n = 0)**
  Duplicates identified manually (n = 0)
  Duplicates identified by Covidence (n = 0)
  Marked as ineligible by automation tools (n = 0)
  Other reasons (n = 0)

Screening

Studies screened **(n = 103)** → Studies excluded **(n = 49)**

Studies sought for retrieval **(n = 54)** → Studies not retrieved **(n = 0)**

Studies assessed for eligibility **(n = 54)** → Studies excluded **(n = 44)**
  Review (n = 1)
  Methodology Paper (n = 1)
  <20 Genetic Instruments (n = 9)
  Instrument Regression Coefficients Not Reported (n = 31)
  Standard Error of Regression Coefficients Not Reported (n = 2)

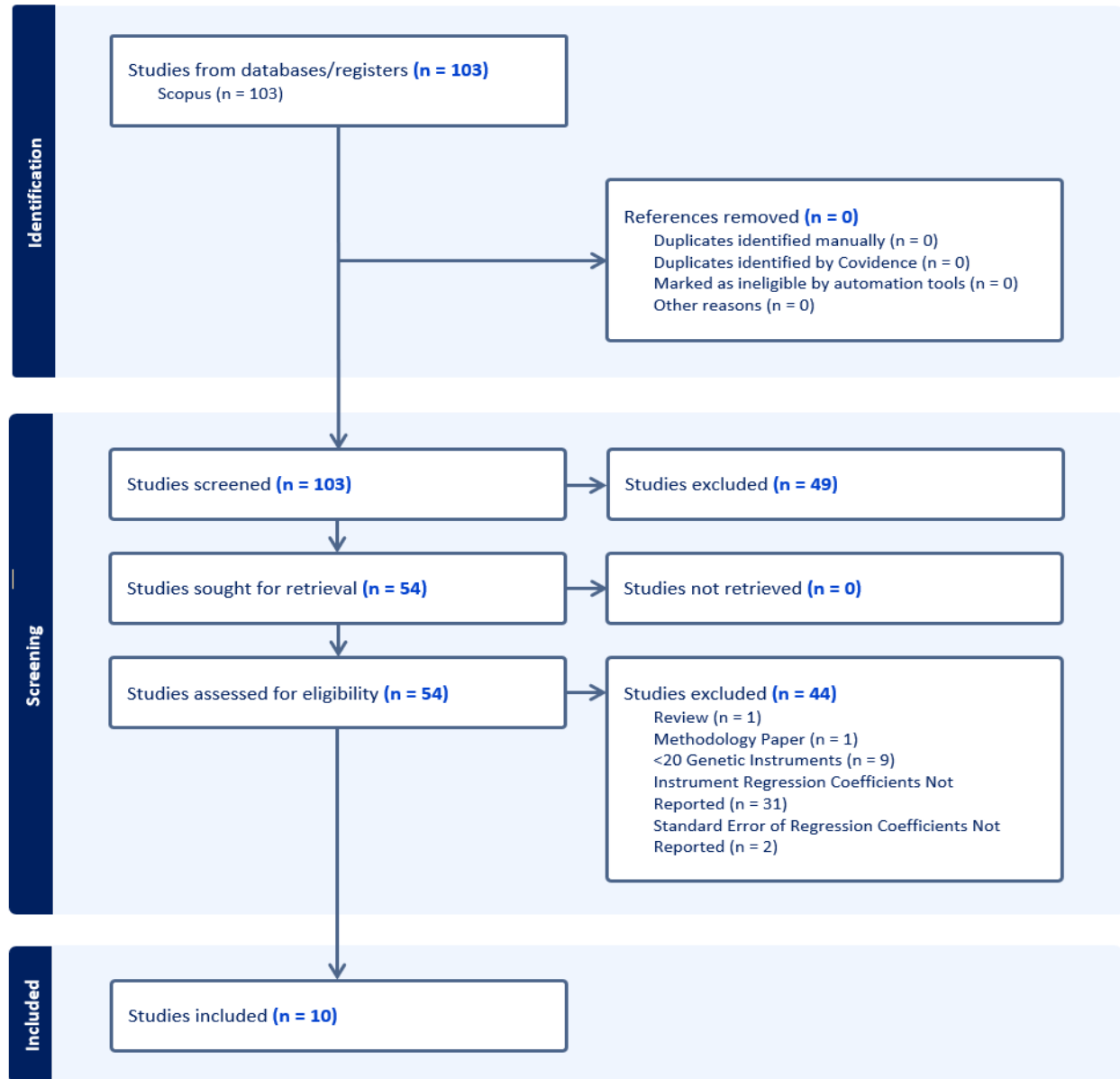**Included**

Studies included **(n = 10)**

Figure 4: Flow diagram illustrating selection of a sample of ten highly-cited two-sample Mendelian randomisation articles reporting a weighted median estimate of causal effect

Table 3: Summary of ten highly-cited two-sample Mendelian randomisation articles reporting a weighted median estimate of causal effect

Author	Citations	Association	N Instruments	Participants			Causal Effect		Causality Reported	p-value
				N		Maximum Estimated Overlap				
				Exposure	Outcome		Measure	Estimate		
Budu-Aggrey et al, 2019	182	BMI on Psoriasis	97	339,224	12,559	0%	OR	1.06 (1 to 1.12)	No	-
Carreras-Torres et al, 2017	200	Height on Pancreatic Cancer	558	253,288	15,002	19%	OR	1.14 (1 to 1.29)	No	0.05
Carter et al, 2019	199	Education on Coronary Disease	1,267	766,345	184,305	0%	OR	0.62 (0.57 to 0.67)	Yes	<0.001
Choi et al, 2019	492	Activity on Depression	24	377,234	143,265	0%	OR	1.49 (0.94 to 2.36)	No	0.08
Clift et al, 2022	129	Smoking Initiation on COVID-19 Infection	378	1,232,091	281,105	36%	OR	1.53 (1.02 to 2.28)	Yes	0.04
Ligthart et al, 2018	298	CRP on Schizophrenia	52	204,402	82,315	0%	OR	0.89 (0.81 to 0.96)	Yes	0.004
Mokry et al, 2016	199	BMI on Multiple Sclerosis	70	322,105	38,589	2.5%	OR	1.26 (0.98 to 1.62)	No	0.08
Pasman et al, 2018	328	Schizophrenia on Cannabis Use	109	150,064	184,765	0%	β	0.163 (0.067 to 0.259)	Yes	0.001
Xie et al, 2023	138	T2DM on NAFLD	526	441,016	218,792	0%	OR	1.61 (1.09 to 2.38)	Yes	<0.001
Xu et al, 2022	183	Coeliac on Gut Bifidobacterium	105	15,283	24,269	63%	OR	0.998 (0.99 to 1.005)	No	0.56

β and OR presented as: estimate (95% CI).

β: causal effect estimate, CI: Confidence Interval, OR: Odds Ratio, SE: Standard Error.

BMI: body mass index, CRP: C-reactive protein, NAFLD: non-alcoholic fatty liver disease, T2DM: type 2 diabetes mellitus

4.2.2　Re-Analysis Results

4.2.2.1　Data Validation and Re-Analysis

There were missing gene-outcome coefficients for 3 instruments from Xie et al[35], and 1 instrument in Clift et al[36] was reported as having an implausibly large gene-outcome coefficient and standard error (-1243.03 and 19161.64, respectively); these instruments were excluded from the main analyses, and included/imputed as the most extreme values per study as a worst-case sensitivity analysis, which did not change conclusions (Appendix D.2). Data were otherwise complete as expected per the descriptions in each study manuscript. A summary of the re-analysis results is presented in Table 4; estimates are presented both as β regression coefficients and ORs for discrete variables to aid comparison across studies.

4.2.2.2　Re-Analysed vs Reported WME Causal Estimates

3 of the WME estimates generated through re-analysis matched the originally reported estimates poorly (Ligthart et al[29], Carreras-Torres et al[32], Mokry et al[31]), with a >0.1 difference in re-analysis estimates of OR versus the values originally reported. Re-analysed OR upper or lower CIs were >0.1 different to reported values for 4 studies (Ligthart et al[29], Carreras-Torres et al[32], Mokry et al[31], Budu-Aggrey et al[34]). Details of instruments used in re-analysis were re-checked against the relevant manuscripts to confirm accuracy of data used, with no discrepancies found.

Overall, estimates and CIs from re-analysis of the other 6 studies (Choi et al[27], Xie et al[35], Pasman et al[28], Carter et al[30], Clift et al[36], Xu et al[33]) appeared comparable to reported values, after accounting for rounding errors from published summary data, and random variation inherent in bootstrap generation of CIs.

Compared with reported values of ORs across the 9 studies using them, the mean difference for effect estimates (SE of estimate) from the re-analysis estimate was 0.03 (0.17). For 95% CIs, the mean differences between reported and re-analysed values were 0.07 for the lower bounds and -0.04 for upper bounds, i.e. reported CIs were narrower on average than re-analysed WME CIs.

Conclusions regarding presence of a causal effect were mostly consistent: reported WME and re-analysed WME estimates were discordant in detecting a causal exposure-outcome effect for 2 studies: 1 where a previously reported causal effect was not found (Ligthart et al[29]), and 1 where a causal effect was found that had not been reported previously (Mokry et al[31]).

4.2.2.3　Re-Analysed WME vs MR-Hevo Causal Estimates

Causal effect estimates generated by MR-Hevo were >0.1 different from re-analysed WME estimates for 2 studies (Choi et al[27], Carreras-Torres et al[32]). Compared with WME values of ORs across the 9 studies using them, the mean difference for effect estimates (SE of estimate) from the MR-Hevo estimate was -0.049 (-0.058). For 95% CIs, the mean differences between MR-Hevo and WME values were -0.049 for the

lower bounds and 0.266 for upper bounds, i.e. MR-Hevo CIs were wider and slightly shifted in the negative direction on average than WME values. MR-Hevo OR upper or lower CIs were >0.1 different to WME values for 6 studies (Choi et al[27], Xie et al[35], Ligthart et al[29], Carreras-Torres et al[32], Clift et al[36], Mokry et al[31]).

Overall, estimates and CIs from MR-Hevo analysis of the other 4 studies (Pasman et al[28], Carter et al[30], Budu-Aggrey et al[34], Xu et al[33]) appeared comparable to re-analysed WME values.

Conclusions regarding presence of a causal effect were consistent: re-analysed WME estimates were discordant in detecting a causal exposure-outcome effect in 0 studies versus MR-Hevo, with both reporting a causal effect in the same 5 studies (Xie et al[35], Pasman et al[28], Carter et al[30], Clift et al[36], Mokry et al[31]).

Table 4: Re-analysis of ten highly-cited two-sample Mendelian randomisation articles reporting a weighted median estimate of causal effect, comparing results of both WME and MR-Hevo causal effect estimation methods

Author	Exposure	Outcome	SNPs	Weighted Median				MR-Hevo			
				β	SE	OR	Causality Reported	β	SE*	OR	Causality Reported
Budu-Aggrey et al	BMI	Psoriasis	97	0 (-0.29-0.29)	0.148	1 (0.75-1.34)	No	0.08 (-0.16-0.33)	0.126	1.09 (0.85-1.4)	No
Carreras-Torres et al	Height	Pancreatic Cancer	558	0 (-0.12-0.12)	0.063	1 (0.88-1.13)	No	-0.21 (-1.11-1.38)	0.636	0.81 (0.33-3.99)	No
Carter et al	Years of Education	Coronary Artery Disease	1,266	-0.46 (-0.55–0.38)	0.044	0.63 (0.58-0.69)	Yes	-0.48 (-0.54–0.42)	0.031	0.62 (0.58-0.66)	Yes
Choi et al	Self-Reported Physical Activity	Major Depressive Disorder	25	0.39 (-0.06-0.83)	0.226	1.47 (0.94-2.29)	No	0.23 (-0.2-0.64)	0.216	1.26 (0.81-1.9)	No
Clift et al	Genetically Determined Smoking Initiation	COVID-19 Infection	377	0.43 (0.03-0.82)	0.201	1.53 (1.03-2.27)	Yes	0.37 (0.1-0.63)	0.135	1.45 (1.11-1.88)	Yes
Ligthart et al	Genetically Determined CRP	Schizophrenia	52	-0.34 (-0.81-0.13)	0.240	0.71 (0.45-1.14)	No	-0.46 (-1.62-0.62)	0.570	0.63 (0.2-1.85)	No
Mokry et al	BMI	Multiple Sclerosis	70	0.34 (0.09-0.59)	0.128	1.41 (1.09-1.81)	Yes	0.34 (0.17-0.51)	0.087	1.4 (1.18-1.66)	Yes
Pasman et al	Liability to Schizophrenia	Cannabis Use	109	0.16 (0.06-0.26)	0.052	-	Yes	0.17 (0.08-0.26)	0.046	-	Yes
Xie et al	T2DM	NAFLD	523	0.48 (0.1-0.86)	0.195	1.61 (1.1-2.37)	Yes	0.51 (0.28-0.74)	0.119	1.66 (1.32-2.1)	Yes
Xu et al	Coeliac Disease	Gut Bifidobacterium	105	0 (-0.01-0)	0.004	1 (0.99-1)	No	0 (-0.01-0)	0.003	1 (0.99-1)	No

β and OR presented as: estimate (95% CI).

β: causal effect estimate, CI: Confidence Interval, OR: Odds Ratio, SE: Standard Error, SE*: Estimated Standard Error.

BMI: body mass index, CRP: C-reactive protein, NAFLD: non-alcoholic fatty liver disease, T2DM: type 2 diabetes mellitus

5 Discussion

5.1 Outputs of Methods

5.1.1 Simulation

5.1.1.1 Null Causal Effect

For a null causal effect, MR-Hevo exhibited broadly similar bias away from the null to WME on average, with both methods tending to slightly over-estimate the true value. Considering variance of estimates, the mean estimated SE for MR-Hevo was fractionally higher than for WME. For MR-Hevo, both bias and variance of estimates generally appeared to increase relatively more than those of WME with increasing proportion of invalid instruments under the assumptions of Scenarios 1 and 2, and relatively less under Scenario 3 assumptions.

Despite comparable average effect estimation from a quantitative perspective, MR-Hevo was consistently more accurate in categorical classification as causal/no causal effect, exhibiting a superior false-positive rate in all 24 parameter/scenario combinations with null causal effect. False-positive reports from WME in particular appeared to correlate with parameter/scenario combinations promoting greater pleiotropy and therefore bias away from the null, potentially suggesting that MR-Hevo estimates are relatively more robust to assumption violations than WME. As false-positive reports were still an infrequent occurrence for both methods, it is perhaps unsurprising that mean estimates/SEs don't clearly reflect this difference in Type 1 error.

These results do suggest that, despite MR-Hevo point estimates being marginally more biased away from the null on average than those of WME, MR-Hevo appears to more effectively deal with pleiotropic effects regarding reporting of causality. This primarily appears to be by appropriately reducing the reported precision of the estimate to reflect additional uncertainty, consistently reducing Type 1 error for MR-Hevo versus WME.

5.1.1.2 Positive Causal Effect

When a positive causal effect was present, MR-Hevo exhibited a slightly higher mean true-positive rate than WME both when all results were pooled, and also in the majority of parameter/scenario combinations. MR-Hevo estimates were generally more biased in the positive direction than WME; this bias was generally stable across scenarios for MR-Hevo. For WME estimates, reported causal effect magnitude appeared to correlate with assumption violations promoting bias away from the null, again potentially suggesting that this method's estimates are more susceptible to peturbation by pleiotropic effects. In this light, higher true-positive rates for WME versus MR-Hevo seen in Scenario 3 with $\geq$ 20% invalid instruments may seem a less convincing endorsement for its use.

Variance of estimates was similar to the null causal effect case, with broadly comparable average SEs/CIs across both methods, though disproportionately higher variance for MR-Hevo versus WME estimates with parameter/scenario combinations promoting bias due to pleiotropy.

Again, these findings would be in keeping with MR-Hevo reporting causality in a way more robust to assumption violations than WME, rather than WME truly being a more powerful method for detecting causal effects.

An exception to general trends is the combination of 0% invalid instruments with 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 with 10,000 participants.

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, with respect to the power of MR-Hevo methods to detect a true causal effect, sample size interacts in a non-linear way with invalid instrument proportions approaching zero. 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%.

5.1.2   Re-Analysis

As discussed in 6.1.2, the comparison of re-analysis WME causal effect estimates to those of MR-Hevo may represent the true performance of MR-Hevo on "real-world" data poorly, given the poor reproducibility of WME findings on re-analysis. This inability to reproduce published results using the corresponding published data and methods potentially represents a major issue for the field of MR - arguably greater than any difference in outcomes between causal estimation methodologies. A full discussion is outside the scope of this project; however, the following data features were noted which may explain some of this phenomenon. Of the six studies with reproducible WME estimates, all presented data rounded to three to five decimal places. By contrast, of the four non-reproducible studies, only one (Ligthart et al[29]) presented data rounded to three decimal places, with the other three reporting to one to two significant figures. It is possible that

such minor data variations are proportionally large enough to influence causal effect estimates, given that most gene-exposure and gene-outcome coefficients are numerically small in absolute terms. The author (B233241) was unable to find any literature pertaining to this, or indeed to reproducibility of MR results in general; this would seem to be a lacuna which warrants further investigation.

In studies where WME estimates were reproducible, MR-Hevo estimates and CIs matched these closely. Across all studies, reproducible or not, conclusions regarding causality matched exactly, although estimates and particularly confidence intervals differed substantially between the two methods for some studies. This is a broadly reassuring finding, suggesting that the conclusions of the most highly-cited works in the field are robust to differences in methodology. However, with thousands of MR studies now reported, as evidenced by the citation search returning 5,417 articles referencing Bowden et al[8] alone, this is not to say that MR-Hevo could not change conclusions of many published MR studies in the literature if it were used to re-analyse their data.

5.2 Results in Context

A key concern in the MR literature of late has been suspiciously high numbers of studies reporting causal effects, often where causality does not seem biologically plausible[16]. 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 an element of publication bias in favour of studies reporting statistically significant results[37,38]; 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[15] 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 exposures and outcomes which could be examined[39]. 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 them to write a paper which programmatically assessed all possible exposure-outcome associations on the platform, in an attempt to disincentivise this practice[40]:

> "…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"[41]

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"[16], 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[42].

By contrast, the group behind MR-Hevo assert that valid methods should yield valid results, regardless of the scale on which analyses are performed. Further, they argue that it is not biologically plausible that directional pleiotropy would routinely exist without the InSIDE assumption also being violated[19]. Regarding this study's simulation results, this pre-supposed coupling of direction and magnitude of pleiotropic effects would imply that Scenario 2 is essentially defunct. The most relevant assumption set regarding bias introduced by invalid instruments with pleiotropic effects would then be Scenario 3, where WME exhibited the highest false-positive causal report rates, and where MR-Hevo arguably exhibits both the greatest robustness to assumption violations, but also the largest fall in true-positive rate versus WME. In the above "everything against everthing" pre-print[40], it is significant that evidence of pleiotropy was noted in >90% of comparisons on the MR-Base platform. The presence of truly balanced pleiotropy would be expected to add only variance to causal effect estimates, and therefore could not explain a high false-positive causal report rate. Taken together, this would therefore suggest that MR-Hevo is the more suitable of the two methods to address any contributions of pleiotropic effects to the high false-positive causal report rates observed.

6   Limitations and Recommendations

6.1   Limitations

6.1.1   Simulation Study

Key objectives of this study were to evaluate the bias and variance of WME and MR-Hevo causal estimates under differing sets of assumptions, including differing proportions of invalid genetic instruments. The random seed used happened to assign similar numbers (6 vs 7) of invalid instruments to both the 20% and 30% invalid instrument cases; relevant code was checked to ensure this was not an error in specification of model parameters. As this was noted after analysis had begun, it was decided not to re-run simulations to avoid implicit multiple comparisons in analysis. However, the resulting datasets generated arguably do not represent the true differences expected between cohorts with a 10% change in valid instrument proportion. In particular, Scenario 3 simulations may have been disproportionately affected by this phenomenon. InSIDE assumption violation means each invalid instrument may introduce different proportions of variance and bias to the causal effect estimates generated. If the average variance introduced per instrument is greater than the average bias introduced per instrument, then adding a single extra invalid instrument may act to bias towards the null if its predominant effect is to reduce estimate precision. The addition of several extra instruments may be required for variance terms to average each other out and so for bias terms to predominate; this may explain why several trends in causal estimates, confidence intervals and causal report rates abruptly plateau around 20% under assumptions for Scenarios 1 and 2, and reverse under Scenario 3. Conclusions drawn from trends tracking the progression from 0% to 20% invalid instruments should be unaffected, but extrapolating trends with $\geq$ 30% invalid instruments from these data alone would seem inadvisable. If desired, simulations could be extended to progressively larger proportions of invalid instruments to investigate this possibility further.

This study had intended to exactly duplicate the simulation methodology of Bowden et al[8] to maximise comparability of results; however several barriers prevented this. As outlined in Methods, the process of genotype generation, including frequencies of effect/non-effect alleles of simulated genetic instruments, was not reported by Bowden et al[8]. In addition, Bowden et al used 10,000 simulations per meta-analysis of each combination of Scenario assumptions and simulation parameters; this was rendered impractical due to the computationally intensive nature of MR-Hevo's resampling approach. On a mid-range home desktop computer (specification found in Appendix E.2), each combination of scenario/parameters took in the order of 8 hours to process $n = 1,000$ datasets; testing indicated that the vast majority of processing time was used by MR-Hevo, rather than WME. Assuming linear $O(n)$ growth in processing time, this implies reproducing both tables would take 2.29 weeks of processing time, which was not practical with the time available. This was also realised too late in the project timeline to arrange for alternative computing capability, such as

through the University of Edinburgh's compute cluster "Eddie"[43]. Although exact replication would have been desirable, using 1,000 simulated datasets per analysis appears to have been sufficient to generate comparable mean values of the parameter estimates to those reported by Bowden et al[8]. The main effect of increasing number of datasets per meta-analysis would be to improve precision of the mean estimates of each parameter (i.e. precision of mean causal effect estimates, mean SEs and mean CIs). As the spread of these parameter estimates is not reported in Tables 1 and 2, this change is not likely to have affected conclusions from this study.

Finally, it was noted that parameters used during simulation were not representative of values observed in the ten highly-cited studies used for the re-analysis section of this project. Both number of genetic instruments and number of participants were at least an order of magnitude lower in simulations versus re-analysed studies. Additionally, due to the nature of two-sample MR studies, numbers of participants in exposure versus outcome cohorts varied substantially in several real-world studies, whereas the simulation code only considered equally sized cohorts. As mentioned in Methods, effect allele frequency was simulated around 50% in line with Bowden et al[8]; where re-analysed studies reported effect allele frequencies, these typically varied from around 10-90% across the range of genetic instruments included. The effects of varying any of these parameters, either alone or in combination, could plausibly affect the true- and false-positive report rates of any MR causal effect estimation method. Post-hoc sensitivity analyses of simulations using larger numbers of instruments and participants (chosen as the most discrepant parameters versus observed values) are presented in Appendix D.1; these analyses indicate substantial changes in study conclusions could result from such alterations to simulation parameters, with WME false-positive rates up to 100% in some cases. It may be that the original simulation parameter values reflect the majority of MR literature better than these highly-cited studies, which by their nature would be expected to include more data than is typical for the field. However, given that these parameter values are taken from the original WME exposition[8], it is also possible that they were originally chosen as a "best case scenario", intended to represent WME performance in the most favourable possible light. If this were the case, then WME could paradoxically struggle more with the increasingly comprehensive datasets now commonly used for MR studies. Extending this project using simulation parameter values more representative of real-world data would therefore seem warranted before drawing definitive conclusions regarding relative performance of both methods.

6.1.2 Re-Analysis of Published Data

As noted in Results, reproducibility of published findings from each re-analysed study was sub-optimal, despite using the same set of genetic instruments as reported in each study. The degree of divergence between published and re-analysed values was not anticipated, and therefore not accounted for in study design. It had been expected that WME re-analysis would confirm published findings more closely, allowing comparison of MR-Hevo vs re-analysed WME estimates to be a valid proxy for comparison of MR-Hevo

vs published WME results; this was only the case in six re-analysed studies. In this project, results of re-analysis were included irrespective of consistency of results with published data; however, any similar future work could employ exclusion criteria for studies whose estimates are not replicable within a specified error margin.

Even if all ten studies had given consistent estimates on re-analysis, this study would have been substantially under-powered to detect a true difference between methods. Using the mean difference in false-positive report rate observed across all scenario/parameter combinations in the simulation study (MR-Hevo = 0.41% vs WME = 5.1%, Section 4.1.2.1), to detect a difference of this size with $\alpha = 0.05$ and $1 - \beta = 0.8$ would require a sample size of around 195 studies to be re-analysed by both methods. If only 40% of studies were adequately reproducible for inclusion, the required sample size for re-analysis would increase to 488. Even with the most extreme difference in false positive report rates observed (MR-Hevo = 0.5% vs WME = 28.3%; $N = 20000$, 20% invalid instruments, Scenario 3; Table 1), the required sample size would be 27, or 68 if only 40% of studies were adequately reproducible for inclusion. As such, it is not possible to conclude that MR-Hevo methods would not change conclusions in a substantial number of studies; as stated in Methods, the intent of this project was to assess the potential impact of MR-Hevo by initially re-analysing high-impact studies, where impact was proxied by citation count.

6.2 Recommendations

The results of this project do not suggest the need to disregard every MR-derived causal association identified using WME methodology. However, the results would be compatible with a substantial absolute number of studies in the literature which may falsely report a causal effect due to use of WME. Furthermore, causal reports from MR studies may separately be untrustworthy by virtue of not being reproducible from data and methodology reported; the extent to which this affects the MR evidence base is not known. If MR-Hevo were used in place of WME to identify potential causality in future MR studies, this would be expected to lower the false-positive report rate by up to ~25% - though with a potential loss of statistical power of ~10-20% in situations where significant bias from pleiotropic effects is present. Given the field-wide concerns regarding high false-positive rates, this may be a reasonable compromise.

Following this project, the recommendations below are offered:

- Further research is required to estimate the proportion of MR literature whose results and conclusions are not reproducible from the data and methods presented. Such work would ideally investigate potential causes of such discrepancies (e.g. the contribution of rounding errors in summary results) so that guidance can be developed to prevent further non-reproducible studies being created. A less pressing research suggestion would be to evaluate the effect of varying simulation parameter values (e.g. num-

ber of participants, number of genetic instruments and effect allele frequency) on rates of true- and false-positive/true- and false-negative reports of causality from MR causal estimation methods.

- Before taking any significant action on the results of any MR study reporting causality using any methodology, attempts should be made to reproduce key findings from reported data and methods. Where this is not possible (e.g. due to data availability restrictions), consideration should be given as to whether said significant actions would still be taken if effect estimates and/or their CIs were to alter by a plausible margin of around $\beta = \pm 0.1$

- For interpretation of existing MR studies relying on WME causal effect estimation, re-analysing using MR-Hevo methods is unlikely to substantially alter the magnitude or direction of estimated causal effect. Where a WME MR study reports no evidence of causality, MR-Hevo re-analysis is unlikely to overturn this conclusion. MR-Hevo is, however, more conservative than WME when generating CIs; it is therefore likely to change overall interpretation in a significant minority of cases reported as supporting a truly causal exposure-outcome association. Re-analysis of WME MR studies may therefore be warranted as a sensitivity analysis of WME MR studies reporting causality, either where significant action is planned on the strength of the results, or where the validity of the result is questioned.

- For future MR studies looking to establish potential causal links between exposures and outcomes, use of MR-Hevo causal estimation is expected to produce a lower false-positive causal report rate than WME methods. The main disadvantages are a) a corresponding loss in power in cases with substantial pleiotropy; and b) the extra compute required, though for most applications this difference will be trivial.

7 Conclusions

This project principally aimed to establish whether the WME causal effect estimation method for MR studies produces over-confident effect estimates versus the MR-Hevo method, and whether this might affect the validity of conclusions drawn in real-world studies.

Using simulated data with known parameters, WME and MR-Hevo produced similar effect estimates and 95% CIs when averaged over all cases. However, MR-Hevo reported wider CIs than WME specifically in cases where key assumptions of IV analysis were relatively more violated; this resulted in an overall false-positive report rate an order of magnitude lower for MR-Hevo versus WME (0.41% versus 5.1%) across simulations with no causal effect present. In cases where a causal effect truly was present, the overall sensitivity was slightly higher for MR-Hevo versus WME (31% versus 28%). In a minority of cases (42%), MR-Hevo's greatly improved specificity came at the expense of a moderate drop in sensitivity, typically in the order of 5-10%, though up to 23% in the most extreme case. Higher sensitivities with WME versus MR-Hevo analysis appeared to correlate with increasing effect of parameters and assumption violations biasing away from the null; this result may therefore represent MR-Hevo estimates being more robust to assumption violations than those of WME.

Re-analysis of data from highly cited MR studies was attempted. WME causal effect estimates were found to be poorly reproducible from available data; this finding alone is concerning for the validity of conclusions for the whole field of MR, though further investigation of this phenomenon was outside the scope of this project. MR-Hevo produced 95% CIs that were wider on average than those generated by WME, and in 6 of 10 cases this difference was large enough that it could plausibly have changed conclusions for a small causal effect size. In the sample of 10 re-analysed studies, the final conclusion regarding presence of a true causal effect was not changed in any case by use of MR-Hevo versus WME methods. Characteristics of included studies were noted to diverge significantly from simulation parameters used in the original validation of WME and reproduced in this project, particularly in terms of number of participants and number of genetic instruments used.

These results suggest that MR-Hevo may be a more suitable method than WME for avoiding false-positive reports of causality between exposures and outcomes, which has been a key challenge for the field of MR of late. Within the limitations of the small sample of relatively high-quality MR studies re-analysed, it does not appear necessary to disregard all conclusions drawn from existing WME MR analyses. Given the tendency of WME to produce false-positive results, it may be warranted to re-analyse selected MR studies which report causality from WME estimates, e.g. where results are borderline suggestive of causality, where IV analysis assumptions are likely to be substantially violated, or where significant further action is planned on the basis of MR results. The main drawbacks of MR-Hevo versus WME are the more computationally intensive implementation, and a reduction in sensitivity in cases with substantial pleiotropic effects present.

Suggested directions for future work include quantifying and further characterising the non-reproducibility of MR results observed here, and evaluating MR causal estimation methods using simulation parameters representative of observed values from existing real-world studies.

8 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.   Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant Pleiotropy in Human Complex Diseases and Traits. American Journal of Human Genetics [Internet]. 2011 Nov [cited 2025 Jun 28];89(5):607–18. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3213397/

11.   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/

12.   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

13.   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

14.   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

15.   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

16. 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

17. Clyde M, Çetinkaya-Rundel M, Rundel C, Banks D, Chai C, Huang L. Chapter 2 Bayesian Inference | An Introduction to Bayesian Thinking [Internet]. 2022 [cited 2025 Jul 1]. Available from: https://statswithr.github.io/book/bayesian-inference.html

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. 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

20. Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. 2014;

21. Hamra G, MacLehose R, Richardson D. Markov Chain Monte Carlo: An introduction for epidemiologists. International Journal of Epidemiology [Internet]. 2013 Apr [cited 2025 Jul 1];42(2):627–34. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3619958/

22. Scopus - Document search | Signed in [Internet]. [cited 2025 Apr 15]. Available from: https://www.scopus.com/search/form.uri?display=basic#basic

23. Covidence [Internet]. 2025. Available from: https://www.covidence.org/

24. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/

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

26. 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.

27. Choi KW, Chen CY, Stein MB, Klimentidis YC, Wang MJ, Koenen KC, et al. Assessment of Bidirectional Relationships Between Physical Activity and Depression Among Adults: A 2-Sample Mendelian Randomization Study. JAMA Psychiatry [Internet]. 2019 Apr [cited 2025 Apr 27];76(4):399–408. Available from: https://doi.org/10.1001/jamapsychiatry.2018.4175

28. Pasman JA, Verweij KJH, Gerring Z, Stringer S, Sanchez-Roige S, Treur JL, et al. GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal effect of schizophrenia liability. Nature Neuroscience [Internet]. 2018 Sep [cited 2025 May 11];21(9):1161–70. Available from: https://www.nature.com/articles/s41593-018-0206-1

29. Ligthart S, Vaez A, Võsa U, Stathopoulou MG, Vries PS de, Prins BP, et al. Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. The American Journal of Human Genetics [Internet]. 2018 Nov [cited 2025 May 27];103(5):691–706. Available from: https://www.cell.com/ajhg/abstract/S0002-9297(18)30320-3

30. Carter AR, Gill D, Davies NM, Taylor AE, Tillmann T, Vaucher J, et al. Understanding the consequences of education inequality on cardiovascular disease: Mendelian randomisation study. BMJ [Internet]. 2019 May [cited 2025 May 27];365:l1855. Available from: https://www.bmj.com/content/365/bmj.l1855

31. Mokry LE, Ross S, Timpson NJ, Sawcer S, Davey Smith G, Richards JB. Obesity and Multiple Sclerosis: A Mendelian Randomization Study. Muraro PA, editor. PLOS Medicine [Internet]. 2016 Jun [cited 2025 May 27];13(6):e1002053. Available from: https://dx.plos.org/10.1371/journal.pmed.1002053

32. Carreras-Torres R, Johansson M, Gaborieau V, Haycock PC, Wade KH, Relton CL, et al. The Role of Obesity, Type 2 Diabetes, and Metabolic Factors in Pancreatic Cancer: A Mendelian Randomization Study. JNCI: Journal of the National Cancer Institute [Internet]. 2017 Sep [cited 2025 May 27];109(9):djx012. Available from: https://doi.org/10.1093/jnci/djx012

33. Xu Q, Ni JJ, Han BX, Yan SS, Wei XT, Feng GJ, et al. Causal Relationship Between Gut Microbiota and Autoimmune Diseases: A Two-Sample Mendelian Randomization Study. Frontiers in Immunology [Internet]. 2022 Jan [cited 2025 May 28];12. Available from: https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2021.746998/full

34. Budu-Aggrey A, Brumpton B, Tyrrell J, Watkins S, Modalsli EH, Celis-Morales C, et al. Evidence of a causal relationship between body mass index and psoriasis: A mendelian randomization study. PLOS Medicine [Internet]. 2019 Jan [cited 2025 Jun 2];16(1):e1002739. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002739

35. Xie J, Huang H, Liu Z, Li Y, Yu C, Xu L, et al. The associations between modifiable risk factors and nonalcoholic fatty liver disease: A comprehensive Mendelian randomization study. Hepatology [Internet]. 2023 Mar [cited 2025 Jun 3];77(3):949. Available from: https://journals.lww.com/hep/pages/articleviewer.aspx?year=2023&issue=03000&article=00022&type=Fulltext#T1

36. Clift AK, Ende A von, Tan PS, Sallis HM, Lindson N, Coupland CAC, et al. Smoking and COVID-19 outcomes: An observational and Mendelian randomisation study using the UK Biobank cohort. Thorax [Internet]. 2022 Jan [cited 2025 Jun 3];77(1):65–73. Available from: https://thorax.bmj.com/content/77/1/65

37. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. International Journal of Epidemiology [Internet]. 2015 Apr [cited 2025 Jun 12];44(2):512–25. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4469799/

38. Bowden J, Jackson D, Thompson SG. Modelling multiple sources of dissemination bias in meta-analysis. Statistics in Medicine [Internet]. 2010 [cited 2025 Jun 12];29(7-8):945–55. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3813

39. Balding DJ. A tutorial on statistical methods for population association studies. Nature Reviews Genetics [Internet]. 2006 Oct [cited 2025 Jun 12];7(10):781–91. Available from: https://www.nature.com/articles/nrg1916

40. Hemani G, Bowden J, Haycock P, Zheng J, Davis O, Flach P, et al. Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome [Internet]. bioRxiv; 2017 [cited 2025 Jun 12]. Available from: https://www.biorxiv.org/content/10.1101/173682v2

41. MRC IEU at University of Bristol. Noodles, No Nulls and Numb Skulls: Threats to the future of Mendelian Randomization - G. Davey Smith [Internet]. 2023 [cited 2025 Jun 12]. Available from: https://www.youtube.com/watch?v=jQmlba_B_lg

42. Munafò MR, Higgins JPT, Smith GD. Triangulating Evidence through the Inclusion of Genetically Informed Designs. Cold Spring Harbor Perspectives in Medicine [Internet]. 2021 Aug [cited 2025 Jun 12];11(8):a040659. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8327826/

43. Eddie [Internet]. Edinburgh University DRS. [cited 2025 Jun 10]. Available from: https://digitalresearchservices.ed.ac.uk/resources/eddie

44. 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

45. 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

46. 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/

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

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

49.   Gillespie C. benchmarkme: Crowd sourced system benchmarks [Internet]. 2022. Available from: https://CRAN.R-project.org/package=benchmarkme

50.   Mautner Wizentier M, Goodman MS, Bather JR. biostats101: Practical functions for biostatistics beginners [Internet]. 2024. Available from: https://CRAN.R-project.org/package=biostats101

51.   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

52.   Xie Y. bookdown: Authoring books and technical documents with r markdown [Internet]. 2024. Available from: https://github.com/rstudio/bookdown

53.   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/

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

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

56.   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

57.   Gohel D, Skintzos P. flextable: Functions for tabular reporting [Internet]. 2025. Available from: https://CRAN.R-project.org/package=flextable

58.   Yasumoto A. ftExtra: Extensions for "Flextable" [Internet]. 2024. Available from: https://CRAN.R-project.org/package=ftExtra

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

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

61. Nicholls K. gluedown: Wrap vectors in markdown formatting [Internet]. 2024. Available from: https://CRAN.R-project.org/package=gluedown

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

63. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/

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

65. 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.

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

67. 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.

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

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

70. Bengtsson H. matrixStats: Functions that apply to rows and columns of matrices (and to vectors) [Internet]. 2024. Available from: https://CRAN.R-project.org/package=matrixStats

71. Gohel D, Moog S, Heckmann M. officer: Manipulation of microsoft word and PowerPoint documents [Internet]. 2025. Available from: https://CRAN.R-project.org/package=officer

72. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2024. Available from: https://www.R-project.org/

73. 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

74. 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

75. 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

76. Stan Development Team. RStan: The R interface to Stan [Internet]. 2024. Available from: https://mc-stan.org/

77. Murdoch D. tables: Formula-driven table generation [Internet]. 2024. Available from: https://CRAN.R-project.org/package=tables

78. 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

79. 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

80. Hendtlass M. where: Vectorised substitution and evaluation [Internet]. 2024. Available from: https://CRAN.R-project.org/package=where

81. Marwick B. wordcountaddin: Word counts and readability statistics in r markdown documents [Internet]. 2024. Available from: https://github.com/benmarwick/wordcountaddin

A   Appendix: List of Abbreviations

CI confidence interval

CLT central limit theorem

GWAS genome-wide association study

IV instrumental variable

InSIDE Instrument Strength Independent of Direct Effect

MR Mendelian randomisation

OR odds ratio

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 General Method

The typical process for "bootstrap" generating an estimate, SE and CIs of a population parameter (e.g. population mean μ) from a sample x is as follows[44]:

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 also 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 CLT[45], 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 5 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, $\hat{\mu} = (\sum_{i=1}^{k} \bar{x}) \div k$
   - 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 Example: Prostate Volume

The above process is illustrated in Figure 5. 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 (red-shaded areas).
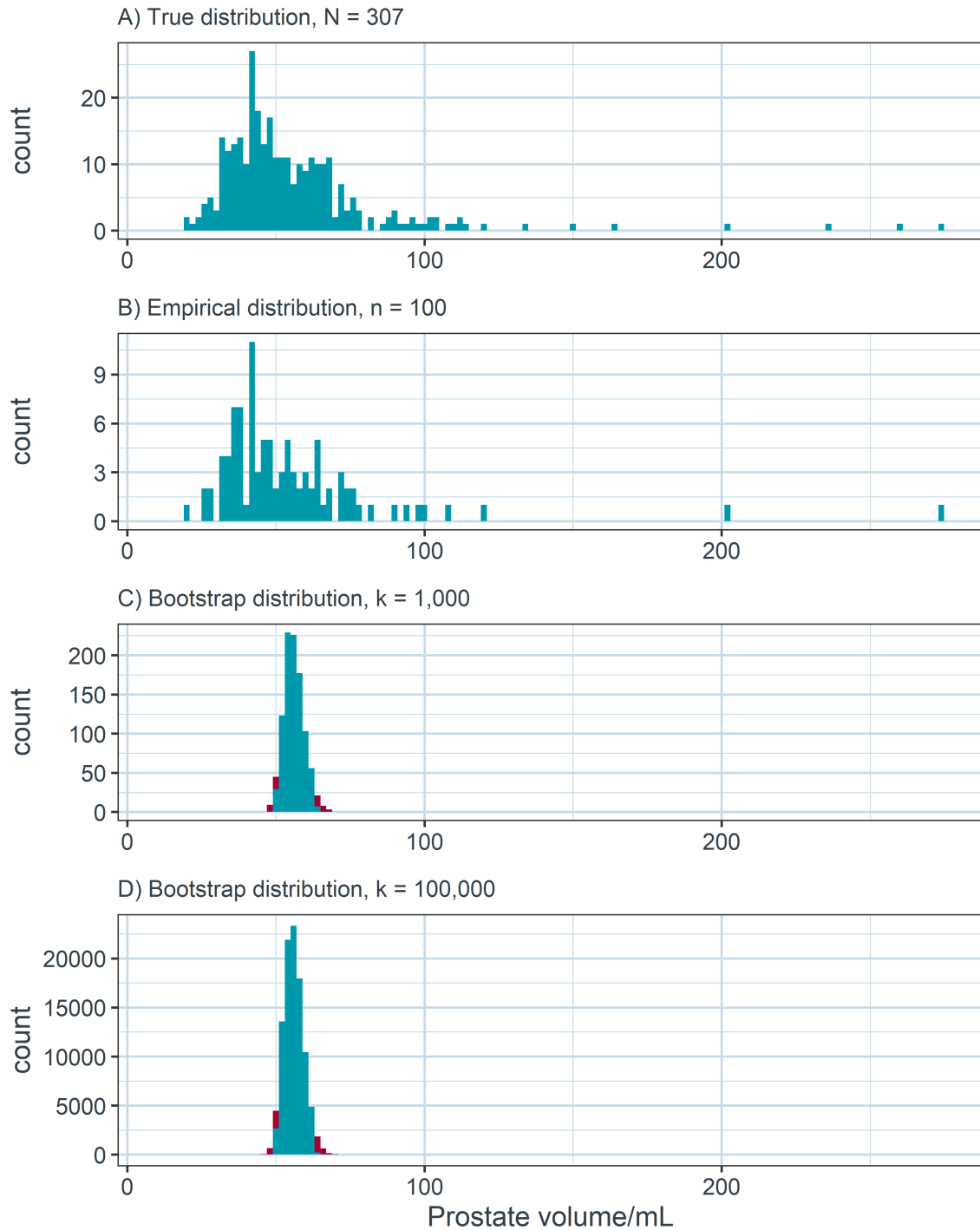
Figure 5: Histograms demonstrating distribution of prostate volumes in patients with prostatic cancer, taken from Cata et al 2011[46] via the R package `medicaldata`[47]. A) Distribution from whole study population of 307 patients with non-missing data, exhibiting right-skew. B) Distribution from random 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 (red-shaded areas).

B.3   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 what is described as a bootstrapping approach. However, the implementation of this deviates from the standard bootstrapping methodology described above.

CIs for WME are generated in the following way (see Web Appendix 2 from the original paper[8] for full code):

1. Summary data (estimates of gene-exposure coefficient $\hat{\gamma}$, gene-outcome coefficient $\hat{\Gamma}$, and their respective standard errors) are obtained for a sample x of n genetic instruments, $G_1, G_2...G_j...G_n$.

2. For each instrument G_j in x, a normal distribution is created centered on the input gene-exposure coefficient estimate $\hat{\gamma}_j$, with the SD of this distribution set equal to the SE of the estimate, $SE(\hat{\gamma}_j)$.

3. Each of these normal distributions, $\mathcal{N}(\hat{\gamma}_j, SE(\hat{\gamma}_j))$, is treated as a bootstrap population for estimating the gene-exposure coefficient and standard error for its respective instrument. Each instrument is re-sampled a default of $k = 1000$ times from its bootstrap population, forming a bootstrap distribution of k estimates of the gene-exposure coefficient for each instrument, $\hat{\gamma}_j^{*1}, \hat{\gamma}_j^{*2}...\hat{\gamma}_j^{*k}$

4. Steps $2.$ and $3.$ are repeated using gene-outcome coefficients and their SEs

5. For each instrument, a Wald estimate of causal effect, $\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$, is obtained from each of the k pairs of gene-exposure and gene-outcome coefficients, $\hat{\beta}_j^{*1}, \hat{\beta}_j^{*2}...\hat{\beta}_j^{*k}$

6. A weighted median causal estimate is calculated from each of the k sets of Wald estimates, $\hat{\beta}_{WME}^{*1}, \hat{\beta}_{WME}^{*2}...\hat{\beta}_{WME}^{*k}$, or $(\hat{\beta}_{WME}^{*i})_{i=1}^{k}$

7. The standard deviation of the k bootstrapped values of $(\hat{\beta}_{WME}^{*i})_{i=1}^{k}$ is taken as the bootstrapped standard error, SE^*, of the empiric/observed WME causal estimate, $\hat{\beta}_{WME}$. The 95% CI is then calculated as $\hat{\beta}_{WME} \pm 1.96 \times SE^*$

This approach differs from typical bootstrapping methodology in several regards. Of particular note, the instruments themselves are not re-sampled; rather, gene-exposure and gene-outcome coefficients are re-sampled from normal distributions constructed for the given set of instruments. This approach may adequately account for some random error in the estimation of these coefficients, but it does not account for the presence, distribution or magnitude of pleiotropic effects across the instruments. A re-sampling approach could account for pleiotropy-related uncertainty, but this would require re-sampling of the instruments, which is difficult to implement in practice.

Given that this method of CI generation does not account for an important source of uncertainty, it would be expected to be over-confident and produce CIs which are too narrow. Narrow CIs are less likely to include the null value, and therefore this may lead to higher rates of causality being reported when using this method,

even when no true causal effect exists. As such, incorrect CI generation may therefore contribute to a higher Type 1 error rate when using WME than might otherwise be expected.

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 ϵ_i^U, ϵ_i^X and ϵ_i^Y were each drawn independently from standard normal distributions. The genetic effects on the exposure γj 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, plus a toggle for random error terms to allow testing of data simulation (i.e. `rand_error = FALSE` sets random error/noise terms equal to zero):

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:

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`).

```r
test_extract_model <- get_models(test_data_sim)
```

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

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

```
##   Mean   :0.1   Mean   :0.3   Mean   :25
##   3rd Qu.:0.1   3rd Qu.:0.3   3rd Qu.:25
##   Max.   :0.1   Max.   :0.3   Max.   :25
##   n_participants
##   Min.   :1000
##   1st Qu.:1000
##   Median :1000
##   Mean   :1000
##   3rd Qu.:1000
##   Max.   :1000
```
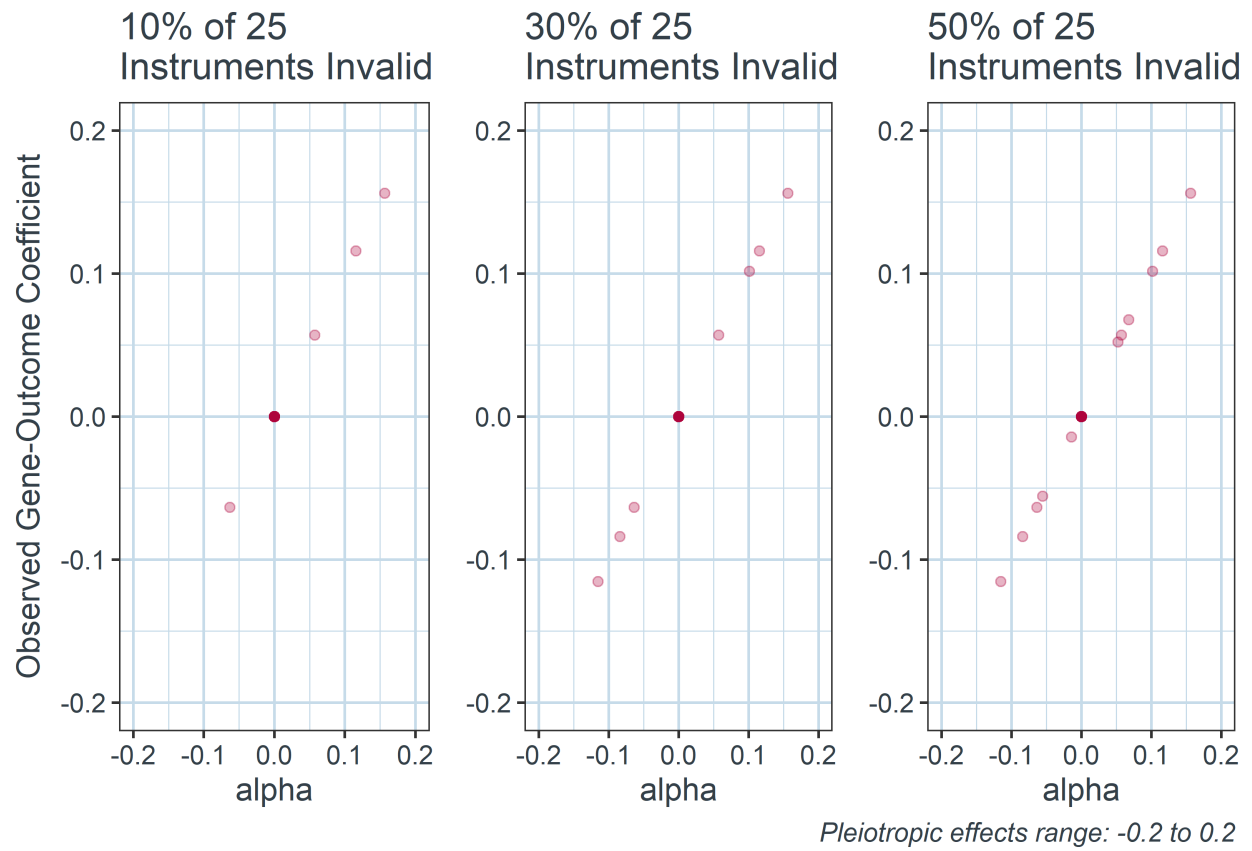
C.2 Testing Generation of Data and Models

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`

| 10% of 25 Instruments Invalid | 30% of 25 Instruments Invalid | 50% of 25 Instruments Invalid |

Pleiotropic effects range: -0.2 to 0.2

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:

Actual and Estimated Gene-Exposure Coefficients Match
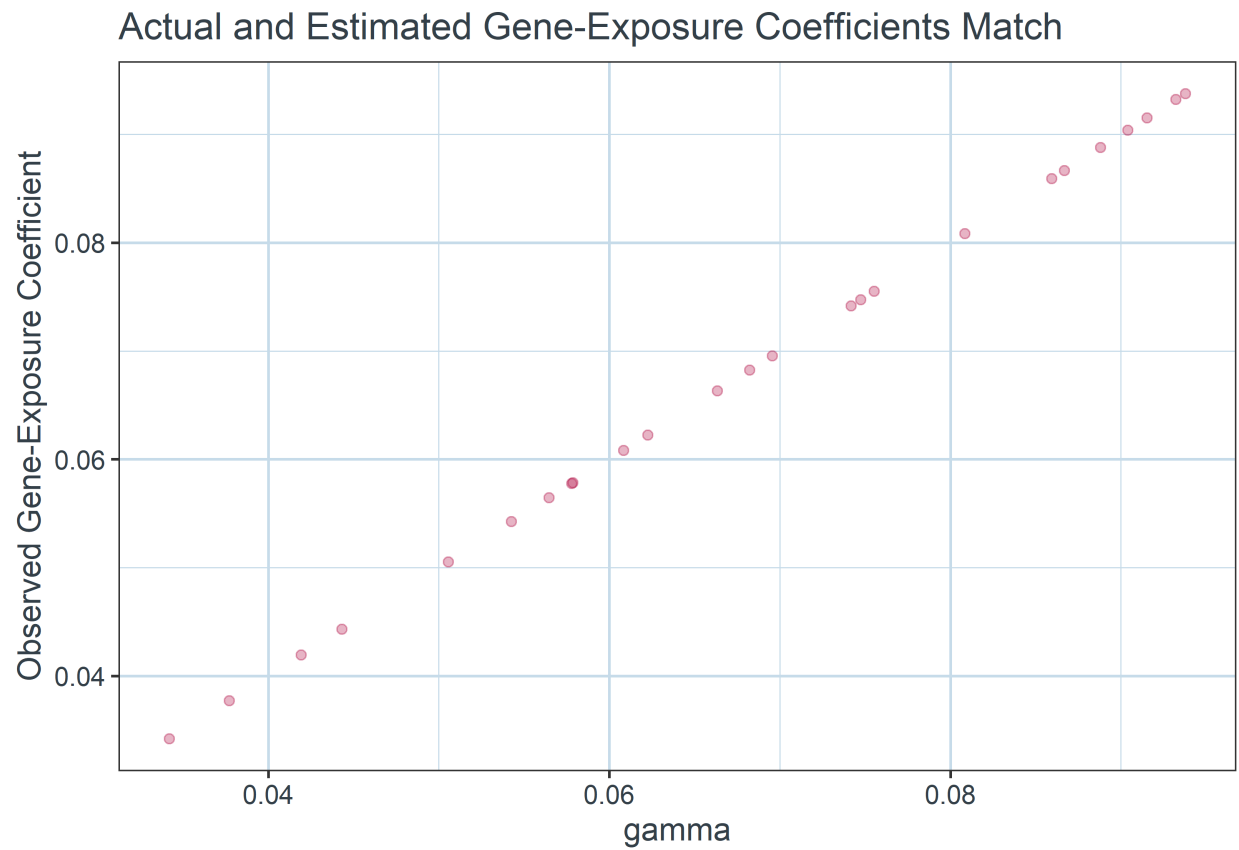
C.2.2 Gene-Exposure Coefficient Versus Gene-Outcome Coefficient Plots

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:

```r
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",
```

```r
            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 β = 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:
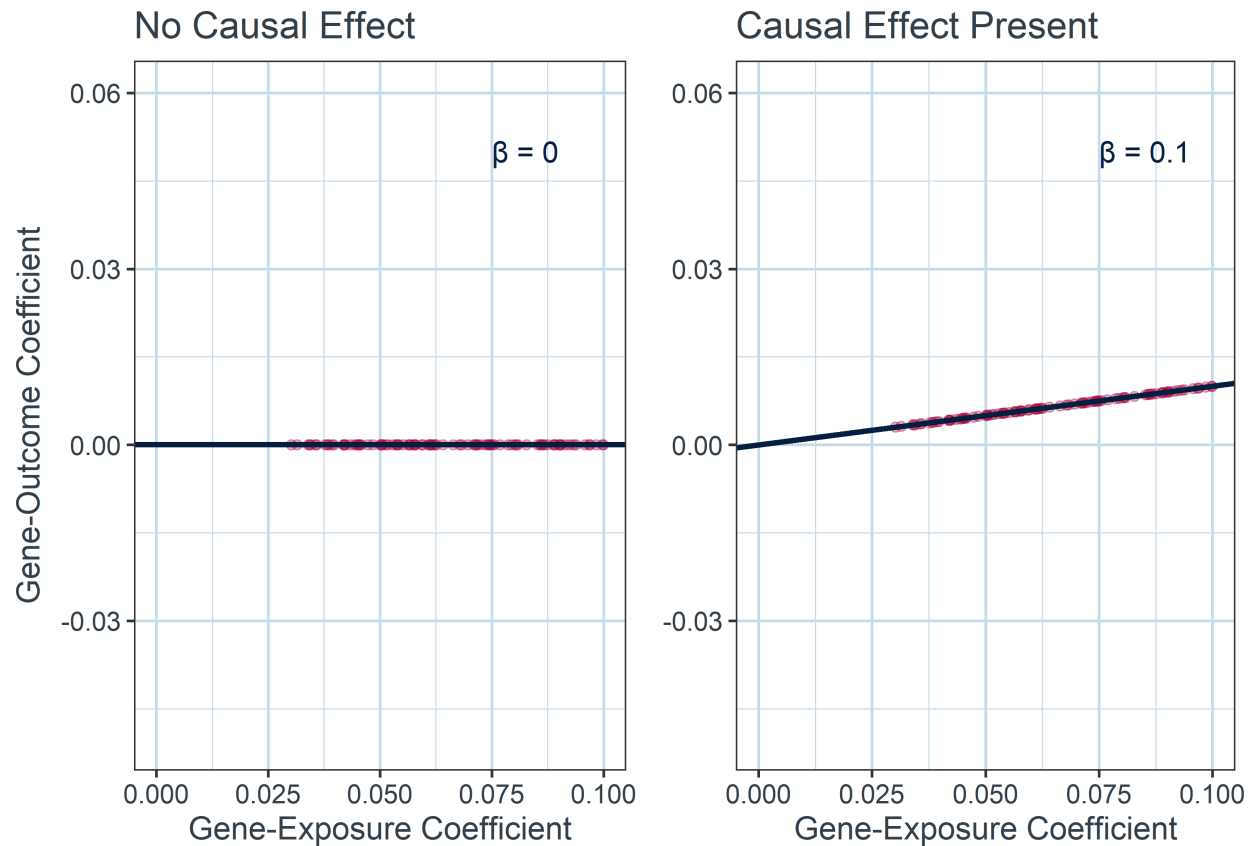
C.2.3   Random Errors

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:
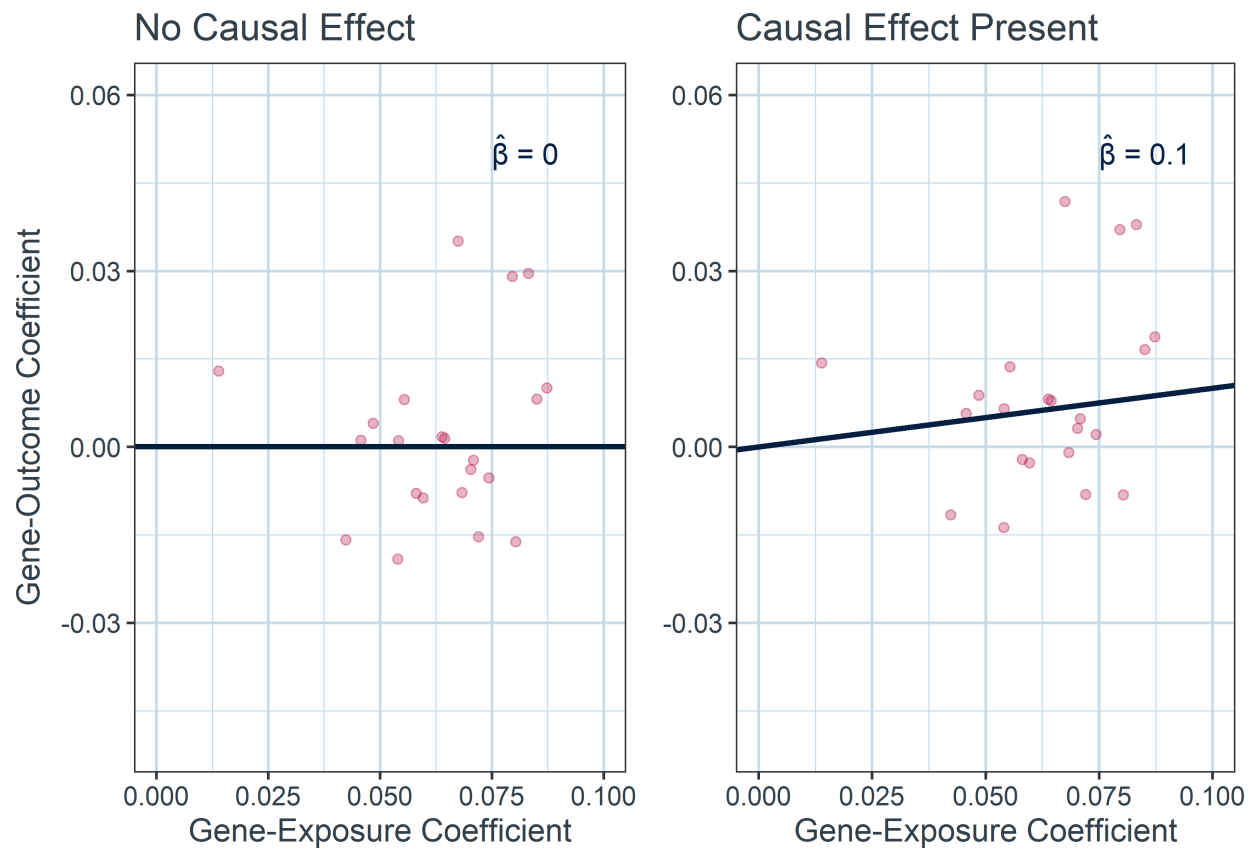
C.2.4 One versus Two Sample MR

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:
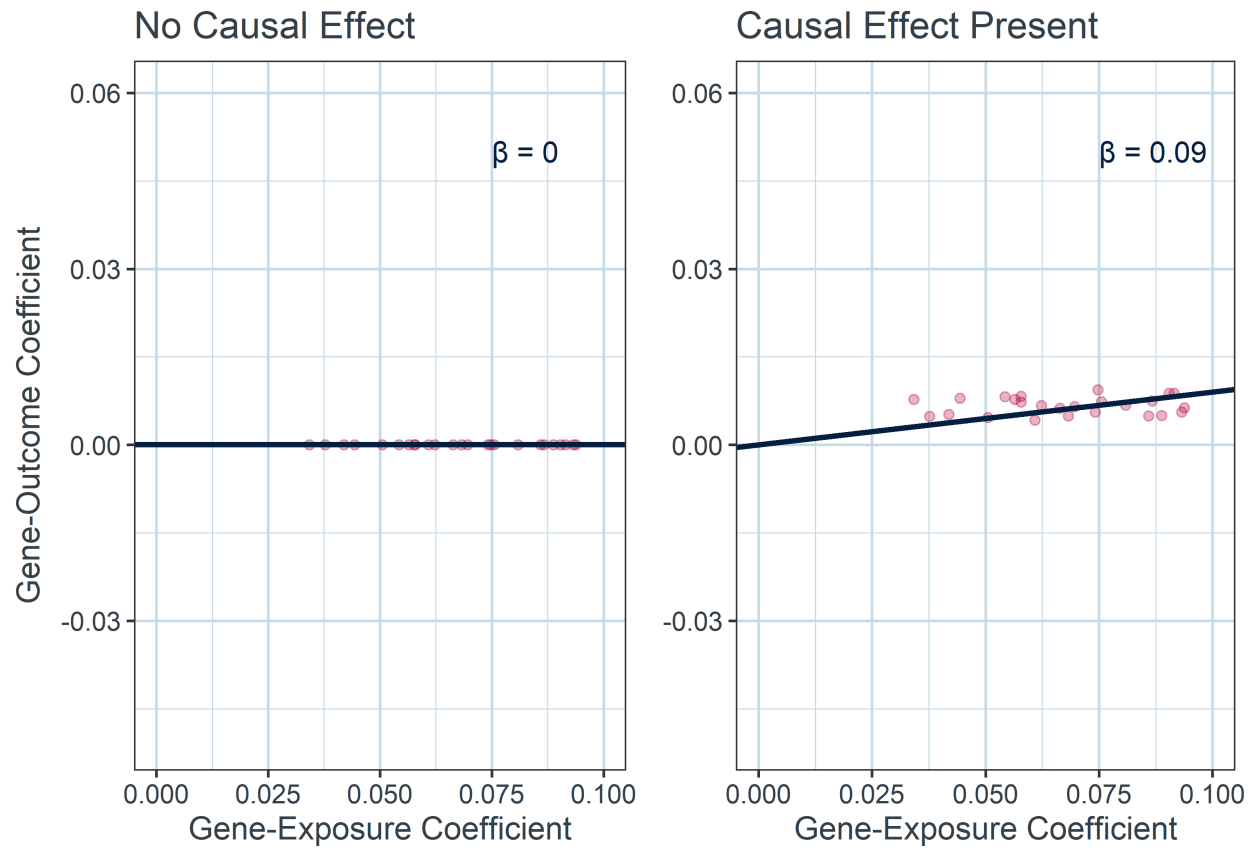
C.2.5 Invalid Instruments

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:
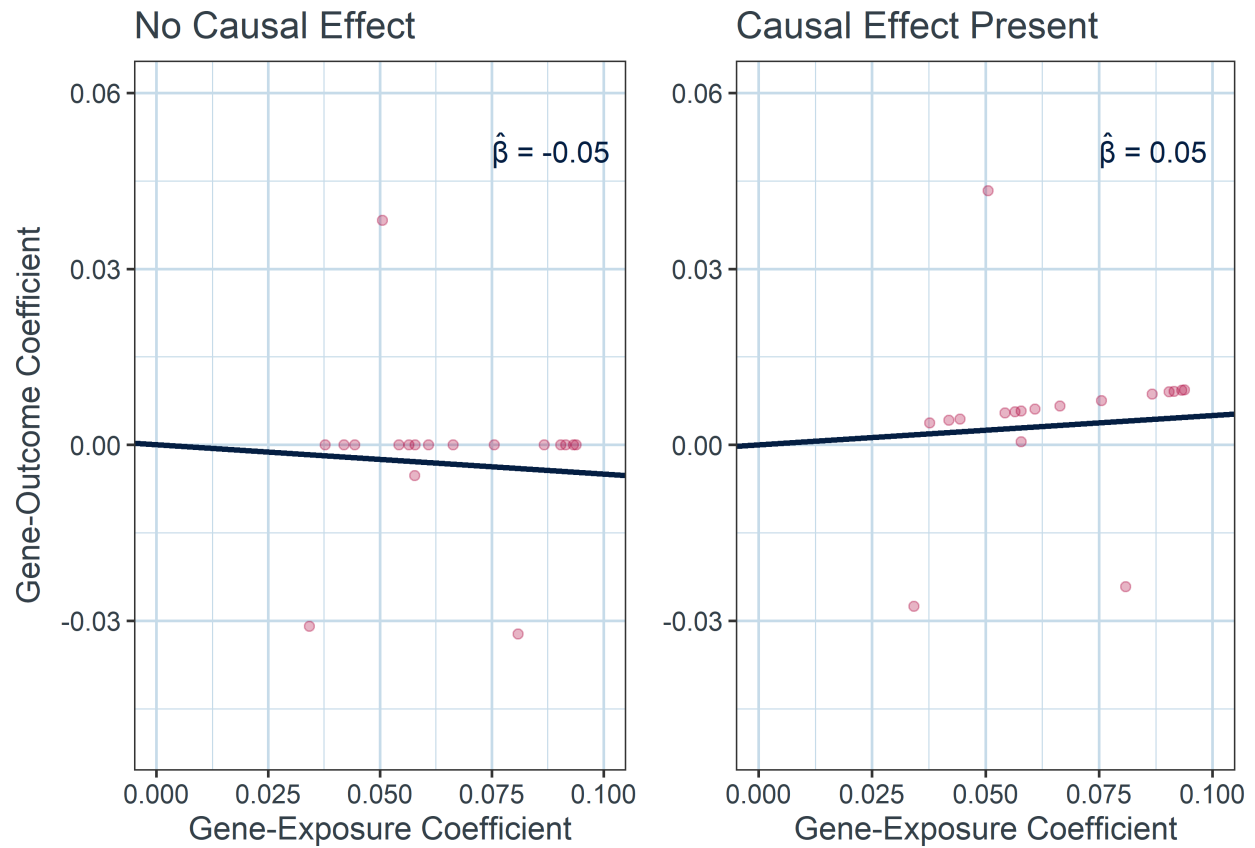
C.2.6 Balanced Versus Directional Pleiotropy

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:
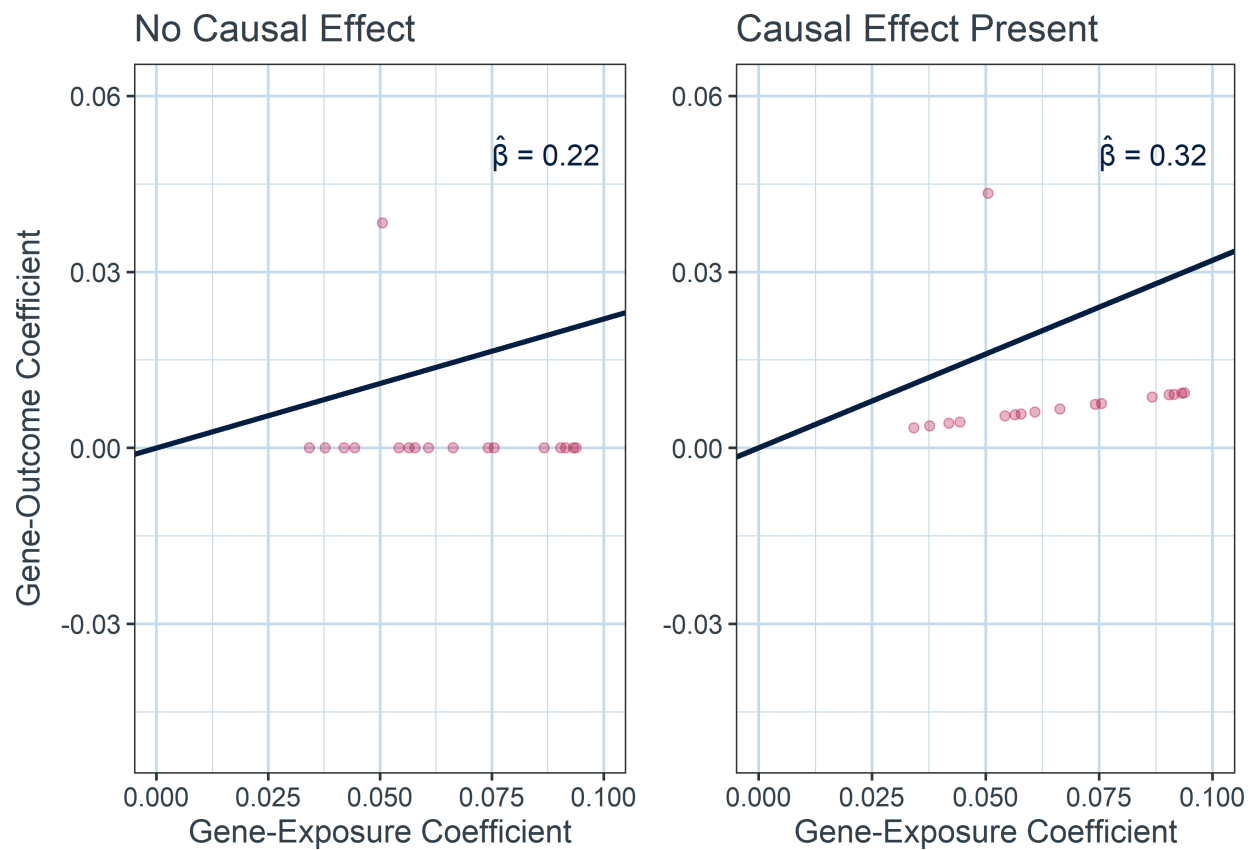
C.2.7 InSIDE Assumption and Phi

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 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:
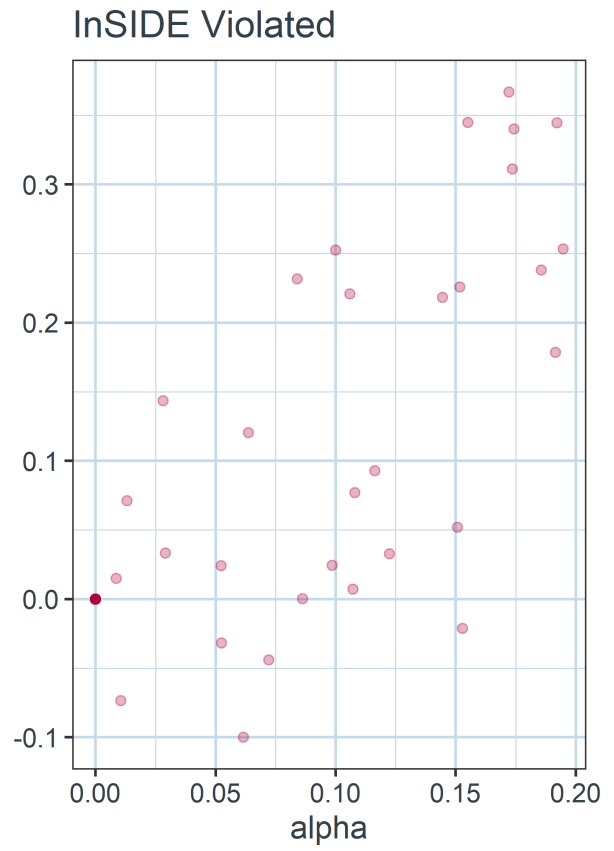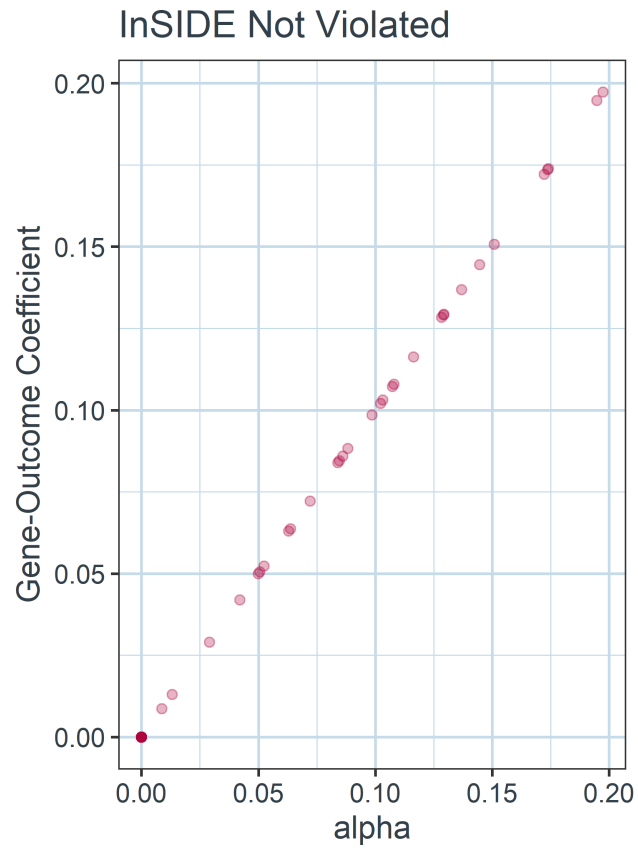
Setting `InSIDE_satisfied = TRUE` should mean `phi = 0`; `InSIDE_satisfied=FALSE` should result in `phi` $\propto$ gene-outcome coefficient, with scatter only in the positive direction of gene-outcome coefficients given the simulation only uses the `phi` term in Scenario 3 (which requires directional pleiotropy):
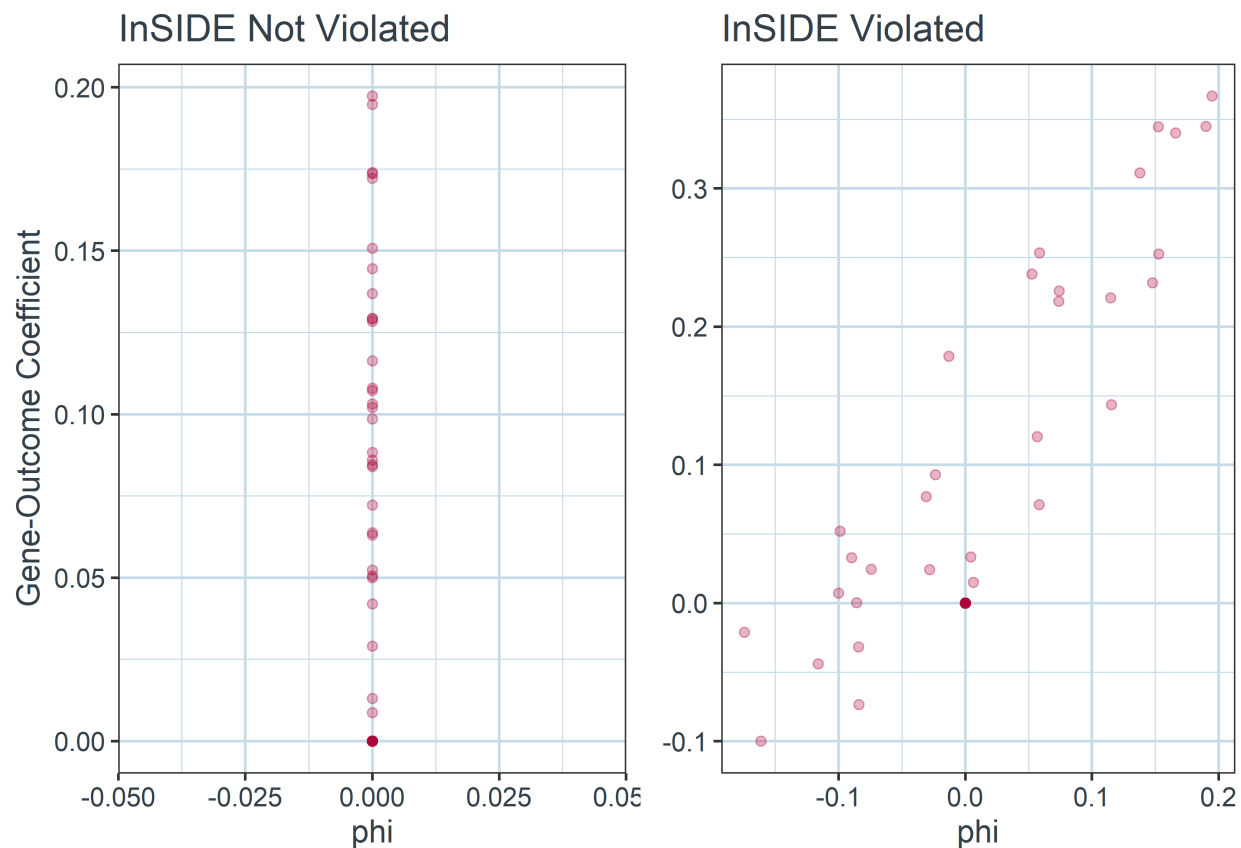
C.3 Summary Table

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]:

```r
# 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::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::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(),
```

```r
                              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
```

```r
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
```

```r
    results_tib[dataset, ]$Hevo_est <- Hevo_results$summary["theta","mean"]
    # Below used in error - actually Monte Carlo SE
    results_tib[dataset, ]$Hevo_se <- Hevo_results$summary["theta","se_mean"]
    # Below should have been used as SE - SD of bootstrap distribution
    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::here(is.double), round, 3))


  return(output_tib_row)


}
```

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

```r
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.
##
## 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.
```

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

D   Appendix: Sensitivity Analyses

D.1   Simulation Study (Post-Hoc)

D.1.1   100 Genetic Instruments

Table 5: Summary of 100 simulated Mendelian randomisation studies per combination of scenario and parameters, all with null causal effect and using 20,000 participants with 100 genetic instruments

N	Invalid IVs	F	R^2	Weighted Median			MR Hevo		
				Mean Estimate (Mean SE)	Mean 95% CI	Causal Report Rate	Mean Estimate (Mean SE*)	Mean 95% CI	Causal Report Rate
Scenario 1: Balanced pleiotropy, InSIDE assumption satisfied									
20,000	0%	24.1	10.8%	-0.003 (0.031)	-0.06 to 0.06	1%	-0.001 (0.021)	-0.04 to 0.04	0%
20,000	10%	24.1	10.8%	-0.002 (0.031)	-0.06 to 0.06	2%	-0.001 (0.024)	-0.05 to 0.05	1%
20,000	20%	24.1	10.8%	-0.005 (0.033)	-0.07 to 0.06	1%	-0.001 (0.028)	-0.06 to 0.05	1%
20,000	30%	24.1	10.8%	-0.003 (0.036)	-0.07 to 0.07	1%	-0.002 (0.036)	-0.07 to 0.07	0%
Scenario 2: Directional pleiotropy, InSIDE assumption satisfied									
20,000	0%	24.1	10.8%	-0.003 (0.031)	-0.06 to 0.06	1%	-0.001 (0.021)	-0.04 to 0.04	0%
20,000	10%	24.1	10.8%	-0.001 (0.031)	-0.06 to 0.06	1%	0.004 (0.025)	-0.04 to 0.05	1%
20,000	20%	24.1	10.8%	0.013 (0.033)	-0.05 to 0.08	2%	0.023 (0.03)	-0.04 to 0.08	1%
20,000	30%	24.1	10.8%	0.035 (0.036)	-0.04 to 0.11	11%	0.063 (0.041)	-0.01 to 0.15	23%
Scenario 3: Directional pleiotropy, InSIDE assumption not satisfied									
20,000	0%	24.1	10.8%	-0.003 (0.031)	-0.06 to 0.06	1%	-0.001 (0.021)	-0.04 to 0.04	0%
20,000	10%	26.9	11.9%	0.017 (0.031)	-0.04 to 0.08	3%	0.009 (0.024)	-0.04 to 0.06	1%
20,000	20%	31.9	13.8%	0.048 (0.034)	-0.02 to 0.11	24%	0.018 (0.029)	-0.04 to 0.08	1%
20,000	30%	45.8	18.7%	0.170 (0.045)	0.08 to 0.26	100%	0.049 (0.038)	-0.02 to 0.12	11%

CI: Confidence Interval, InSIDE: Instrument Strength Independent of Direct Effect, IV: Instumental Variable, SE: Standard Error, SE*:

Estimated Standard Error. F and R² statistics presented are the minimum values across all simulated datasets.

Null Causal Effect (β = 0)

Table 6: Summary of 100 simulated Mendelian randomisation studies per combination of scenario and parameters, all with positive causal effect and using 20,000 participants with 100 genetic instruments

N	Invalid IVs	F	R^2	Weighted Median			MR Hevo		
				Mean Estimate (Mean SE)	Mean 95% CI	Causal Report Rate	Mean Estimate (Mean SE*)	Mean 95% CI	Causal Report Rate
Scenario 1: Balanced pleiotropy, InSIDE assumption satisfied									
20,000	0%	24.1	10.8%	0.072 (0.031)	0.01 to 0.13	71%	0.085 (0.021)	0.04 to 0.13	100%
20,000	10%	24.1	10.8%	0.072 (0.032)	0.01 to 0.13	67%	0.084 (0.025)	0.04 to 0.13	100%
20,000	20%	24.1	10.8%	0.069 (0.033)	0 to 0.13	59%	0.083 (0.029)	0.03 to 0.14	97%
20,000	30%	24.1	10.8%	0.070 (0.036)	0 to 0.14	48%	0.084 (0.037)	0.01 to 0.16	70%
Scenario 2: Directional pleiotropy, InSIDE assumption satisfied									
20,000	0%	24.1	10.8%	0.072 (0.031)	0.01 to 0.13	71%	0.085 (0.021)	0.04 to 0.13	100%
20,000	10%	24.1	10.8%	0.074 (0.032)	0.01 to 0.14	70%	0.089 (0.026)	0.04 to 0.14	100%
20,000	20%	24.1	10.8%	0.087 (0.033)	0.02 to 0.15	80%	0.108 (0.031)	0.05 to 0.17	100%
20,000	30%	24.1	10.8%	0.109 (0.037)	0.04 to 0.18	91%	0.151 (0.042)	0.07 to 0.24	100%
Scenario 3: Directional pleiotropy, InSIDE assumption not satisfied									
20,000	0%	24.1	10.8%	0.072 (0.031)	0.01 to 0.13	71%	0.085 (0.021)	0.04 to 0.13	100%
20,000	10%	26.9	11.9%	0.090 (0.032)	0.03 to 0.15	86%	0.095 (0.025)	0.05 to 0.14	100%
20,000	20%	31.9	13.8%	0.127 (0.035)	0.06 to 0.2	100%	0.108 (0.03)	0.05 to 0.17	100%
20,000	30%	45.8	18.7%	0.245 (0.045)	0.16 to 0.33	100%	0.143 (0.039)	0.07 to 0.22	100%

CI: Confidence Interval, InSIDE: Instrument Strength Independent of Direct Effect, IV: Instumental Variable, SE: Standard Error, SE*: Estimated Standard Error. F and R² statistics presented are the minimum values across all simulated datasets.

Positive Causal Effect (β = 0.1)

D.1.2 200,000 Participants

Table 7: Summary of 100 simulated Mendelian randomisation studies per combination of scenario and parameters, all with null causal effect and using 200,000 participants with 25 genetic instruments

N	Invalid IVs	F	R^2	Weighted Median			MR Hevo		
				Mean Estimate (Mean SE)	Mean 95% CI	Causal Report Rate	Mean Estimate (Mean SE*)	Mean 95% CI	Causal Report Rate
Scenario 1: Balanced pleiotropy, InSIDE assumption satisfied									
200,000	0%	245.6	3%	0.001 (0.018)	-0.03 to 0.04	0%	0.000 (0.014)	-0.03 to 0.03	0%
200,000	10%	245.6	3%	0.006 (0.02)	-0.03 to 0.05	1%	0.002 (0.02)	-0.04 to 0.04	0%
200,000	20%	245.6	3%	0.006 (0.021)	-0.04 to 0.05	1%	0.002 (0.025)	-0.05 to 0.05	0%
200,000	30%	245.6	3%	0.004 (0.021)	-0.04 to 0.05	1%	0.001 (0.027)	-0.05 to 0.06	0%
Scenario 2: Directional pleiotropy, InSIDE assumption satisfied									
200,000	0%	245.6	3%	0.001 (0.018)	-0.03 to 0.04	0%	0.000 (0.014)	-0.03 to 0.03	0%
200,000	10%	245.6	3%	0.004 (0.02)	-0.04 to 0.04	1%	0.002 (0.02)	-0.04 to 0.04	0%
200,000	20%	245.6	3%	0.011 (0.021)	-0.03 to 0.05	2%	0.006 (0.025)	-0.04 to 0.06	0%
200,000	30%	245.6	3%	0.015 (0.022)	-0.03 to 0.06	6%	0.008 (0.027)	-0.04 to 0.06	0%
Scenario 3: Directional pleiotropy, InSIDE assumption not satisfied									
200,000	0%	245.6	3%	0.001 (0.018)	-0.03 to 0.04	0%	0.000 (0.014)	-0.03 to 0.03	0%
200,000	10%	299.7	3.6%	0.020 (0.02)	-0.02 to 0.06	8%	0.004 (0.02)	-0.04 to 0.04	0%
200,000	20%	333.1	4%	0.037 (0.023)	-0.01 to 0.08	32%	0.008 (0.025)	-0.04 to 0.06	0%
200,000	30%	296.7	3.6%	0.026 (0.022)	-0.02 to 0.07	17%	0.007 (0.027)	-0.04 to 0.06	0%

CI: Confidence Interval, InSIDE: Instrument Strength Independent of Direct Effect, IV: Instumental Variable, SE: Standard Error, SE*: Estimated Standard Error. F and R² statistics presented are the minimum values across all simulated datasets.

Null Causal Effect (β = 0)

Table 8: Summary of 100 simulated Mendelian randomisation studies per combination of scenario and parameters, all with positive causal effect and using 200,000 participants with 25 genetic instruments

N	Invalid IVs	F	R^2	Weighted Median			MR Hevo		
				Mean Estimate (Mean SE)	Mean 95% CI	Causal Report Rate	Mean Estimate (Mean SE*)	Mean 95% CI	Causal Report Rate
Scenario 1: Balanced pleiotropy, InSIDE assumption satisfied									
200,000	0%	245.6	3%	0.088 (0.018)	0.05 to 0.12	100%	0.092 (0.014)	0.06 to 0.12	100%
200,000	10%	245.6	3%	0.092 (0.02)	0.05 to 0.13	100%	0.094 (0.021)	0.05 to 0.14	100%
200,000	20%	245.6	3%	0.093 (0.021)	0.05 to 0.13	100%	0.097 (0.026)	0.05 to 0.15	100%
200,000	30%	245.6	3%	0.089 (0.022)	0.05 to 0.13	100%	0.095 (0.028)	0.04 to 0.15	100%
Scenario 2: Directional pleiotropy, InSIDE assumption satisfied									
200,000	0%	245.6	3%	0.088 (0.018)	0.05 to 0.12	100%	0.092 (0.014)	0.06 to 0.12	100%
200,000	10%	245.6	3%	0.090 (0.02)	0.05 to 0.13	100%	0.094 (0.021)	0.05 to 0.14	100%
200,000	20%	245.6	3%	0.099 (0.022)	0.06 to 0.14	100%	0.100 (0.026)	0.05 to 0.15	100%
200,000	30%	245.6	3%	0.102 (0.022)	0.06 to 0.15	100%	0.103 (0.028)	0.05 to 0.16	100%
Scenario 3: Directional pleiotropy, InSIDE assumption not satisfied									
200,000	0%	245.6	3%	0.088 (0.018)	0.05 to 0.12	100%	0.092 (0.014)	0.06 to 0.12	100%
200,000	10%	299.7	3.6%	0.114 (0.021)	0.07 to 0.16	100%	0.102 (0.021)	0.06 to 0.14	100%
200,000	20%	333.1	4%	0.129 (0.024)	0.08 to 0.18	100%	0.101 (0.026)	0.05 to 0.15	100%
200,000	30%	296.7	3.6%	0.111 (0.023)	0.07 to 0.16	100%	0.097 (0.028)	0.04 to 0.15	100%

CI: Confidence Interval, InSIDE: Instrument Strength Independent of Direct Effect, IV: Instumental Variable, SE: Standard Error, SE*: Estimated Standard Error. F and R² statistics presented are the minimum values across all simulated datasets.

Positive Causal Effect (β = 0.1)

D.2 Re-Analysis Study

Table 9: Sensitivity re-analyses using inclusion of outliers and imputation of missing values as most extreme values from the respective dataset

| Author | Re-Analysis | Exposure | Outcome | SNPs | Weighted Median | | | | MR-Hevo | | | |
					β	SE	OR	Causality Reported	β	SE	OR	Causality Reported
Clift et al	Outlier Included	Genetically Determined Smoking Initiation	COVID-19 Infection	378	0.43 (0.04-0.81)	0.195	1.53 (1.05-2.24)	Yes	0.37 (0.11-0.64)	0.134	1.45 (1.11-1.89)	Yes
Xie et al	NAs imputed - max values	T2DM	NAFLD	526	0.48 (0.08-0.87)	0.202	1.61 (1.09-2.4)	Yes	0.52 (0.29-0.75)	0.118	1.68 (1.33-2.12)	Yes
Xie et al	NAs imputed - min values	T2DM	NAFLD	526	0.48 (0.09-0.87)	0.198	1.61 (1.09-2.38)	Yes	0.52 (0.29-0.75)	0.118	1.68 (1.33-2.12)	Yes

β and OR presented as: estimate (95% CI).

β: causal effect estimate, CI: Confidence Interval, OR: Odds Ratio, SE: Standard Error, SE*: Estimated Standard Error.

NAFLD: non-alcoholic fatty liver disease, T2DM: type 2 diabetes mellitus

E Appendix: R Packages Used

E.1 Package Citations

This work was completed using R version 4.4.1[24] with the following R packages: acronymsdown v. 0.11.1[48], benchmarkme v. 1.0.8[49], biostats101 v. 0.1.1[50], bookdown v. 0.41[51,52], car v. 3.1.3[53], cowplot v. 1.1.3[54], crayon v. 1.5.3[55], devtools v. 2.4.5[56], flextable v. 0.9.9[57], ftExtra v. 0.6.4[58], ggdag v. 0.2.13[59], gghighlight v. 0.4.1[60], gluedown v. 1.0.9[61], grateful v. 0.2.12[62], grid v. 4.4.1[63], here v. 1.0.1[64], infer v. 1.0.8[65], kableExtra v. 1.4.0[66], knitr v. 1.50[67–69], matrixStats v. 1.4.1[70], medicaldata v. 0.2.0[47], officer v. 0.6.10[71], parallel v. 4.4.1[72], rmarkdown v. 2.29[73–75], rstan v. 2.32.6[76], tables v. 0.9.31[77], tidyverse v. 2.0.0[26], TwoSampleMR v. 0.6.8[78,79], where v. 1.0.0[80], wordcountaddin v. 0.3.0.9000[81].

E.2 Session Information

CPU: 13th Gen Intel(R) Core(TM) i7-13700H, 20 cores
RAM: 16.9 GB

```
##  setting  value
##  version  R version 4.4.1 (2024-06-14 ucrt)
##  os       Windows 11 x64 (build 26100)
##  system   x86_64, mingw32
##  ui       RStudio
##  language (EN)
##  collate  English_United Kingdom.utf8
##  ctype    English_United Kingdom.utf8
##  tz       Europe/London
##  date     2025-07-04
##  rstudio  2025.05.0+496 Mariposa Orchid (desktop)
##  pandoc   3.4 @ C:/Program Files/RStudio/resources/app/bin/quarto/bin/tools/ (via rmarkdown)

## # A tibble: 35 x 5
##    package   ondiskversion loadedversion date  source
##    <chr>     <chr>         <chr>         <chr> <chr>
##  1 acronyms~ 0.11.1        0.11.1        2025~ Githu~
##  2 benchmar~ 1.0.8         1.0.8         2022~ CRAN ~
##  3 biostats~ 0.1.1         0.1.1         2024~ CRAN ~
##  4 bookdown  0.41          0.41          2024~ CRAN ~
```

```
##  5 cowplot    1.1.3        1.1.3         2024~ CRAN ~
##  6 dplyr      1.1.4        1.1.4         2023~ CRAN ~
##  7 flextable  0.9.9        0.9.9         2025~ CRAN ~
##  8 forcats    1.0.0        1.0.0         2023~ CRAN ~
##  9 ftExtra    0.6.4        0.6.4         2024~ CRAN ~
## 10 ggdag      0.2.13       0.2.13        2024~ CRAN ~
## 11 gghighli~  0.4.1        0.4.1         2023~ CRAN ~
## 12 ggplot2    3.5.2        3.5.2         2025~ CRAN ~
## 13 gluedown   1.0.9        1.0.9         2024~ CRAN ~
## 14 grateful   0.2.12       0.2.12        2025~ CRAN ~
## 15 here       1.0.1        1.0.1         2020~ CRAN ~
## 16 infer      1.0.8        1.0.8         2025~ CRAN ~
## 17 kableExt~  1.4.0        1.4.0         2024~ CRAN ~
## 18 knitr      1.50         1.50          2025~ CRAN ~
## 19 koRpus     0.13.8       0.13-8        2021~ CRAN ~
## 20 koRpus.l~  0.1.4        0.1-4         2020~ CRAN ~
## 21 lubridate  1.9.3        1.9.3         2023~ CRAN ~
## 22 magrittr   2.0.3        2.0.3         2022~ CRAN ~
## 23 medicald~  0.2.0        0.2.0         2021~ CRAN ~
## 24 officer    0.6.10       0.6.10        2025~ CRAN ~
## 25 purrr      1.0.2        1.0.2         2023~ CRAN ~
## 26 readr      2.1.5        2.1.5         2024~ CRAN ~
## 27 rstan      2.32.6       2.32.6        2024~ CRAN ~
## 28 StanHead~  2.32.10      2.32.10       2024~ CRAN ~
## 29 stringr    1.5.1        1.5.1         2023~ CRAN ~
## 30 sylly      0.1.6        0.1-6         2020~ CRAN ~
## 31 tables     0.9.31       0.9.31        2024~ CRAN ~
## 32 tibble     3.2.1        3.2.1         2023~ CRAN ~
## 33 tidyr      1.3.1        1.3.1         2024~ CRAN ~
## 34 tidyverse  2.0.0        2.0.0         2023~ CRAN ~
## 35 TwoSampl~  0.6.8        0.6.8         2024~ https~
```