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: 9996

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)[?]. 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[?].

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[?]. 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)[?].

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

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[?]:

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

These assumptions are represented graphically in Figure 1.

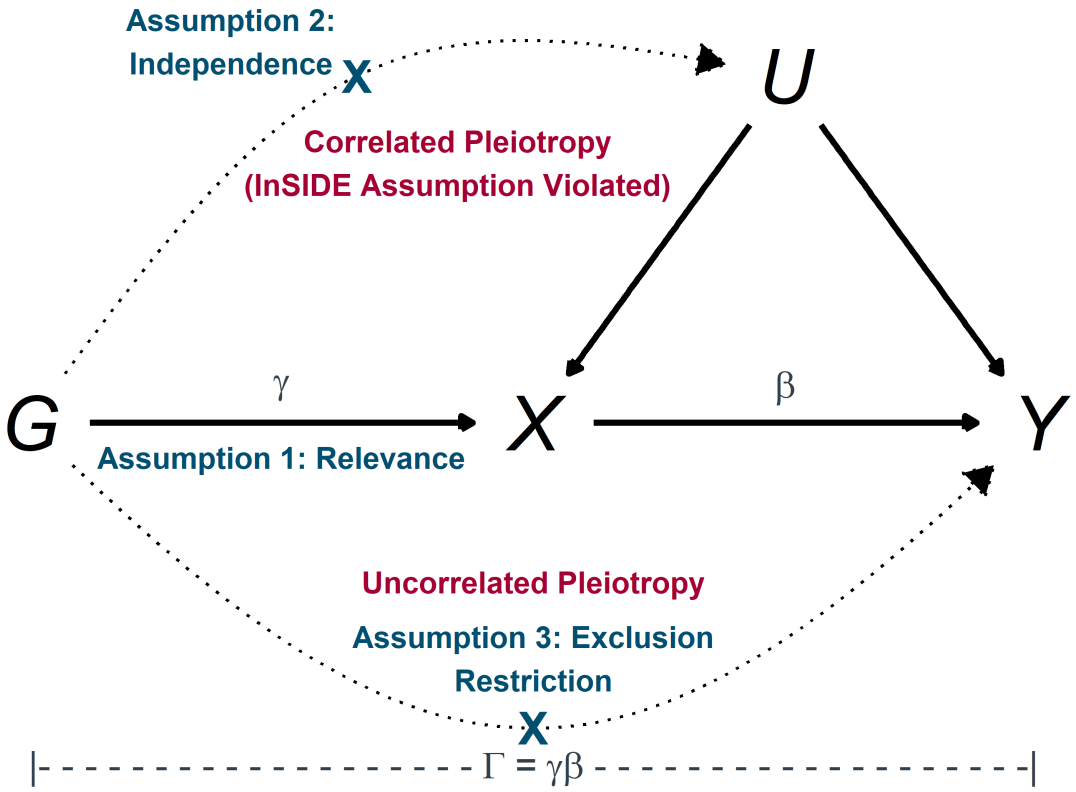

Figure 1: Causal diagram illustrating the relationships between genetic instrument G, exposure X, outcome Y and confounders of the exposure-outcome relationship U in Mendelian randomisation studies. Blue text & crosses represent key assumptions to ensure valid inference of causal effect of X on Y using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

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

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

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[?]. An F-statistic of ≥ 10 is generally considered to represent a strong enough gene-exposure association to consider a genetic instrument for use[?].

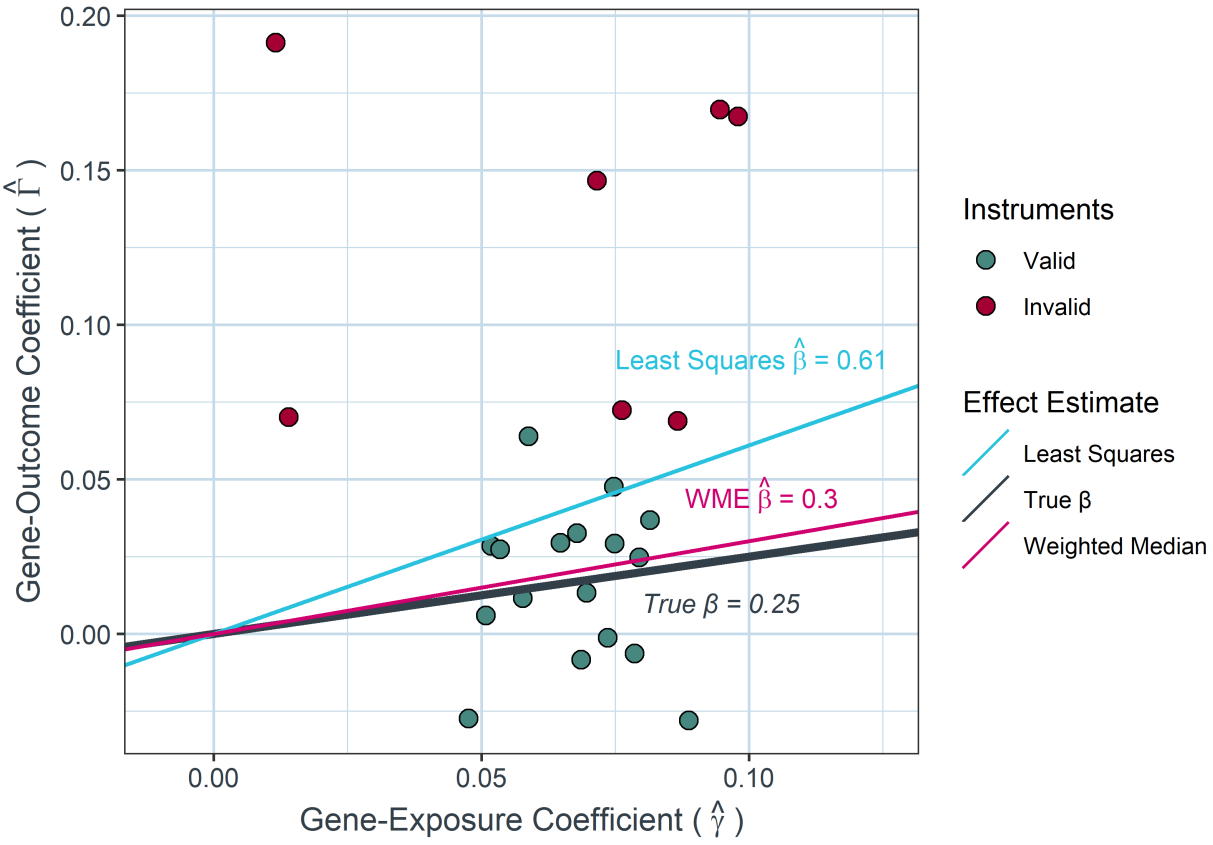

Figure 2: Simulated MR Study on 10,000 individuals using 25 genetic instruments, of which 30% are invalid (red points) and introduce directional pleiotropic effects. The true value of the exposure-outcome causal effect is 0.25 (grey line, causal effect represented by gradient). Regression using an unadjusted 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[?]

The assumptions of independence and exclusion restriction depend on all determinants of the outcome, 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[?].

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[?].

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[?]. 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[?]. 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[?]. By contrast, "directional pleiotropy", where the mean of pleiotropic effects is non-zero, may introduce bias[?] (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[?].

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[?].

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)[?].

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[?]. 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`[?] and `TwoSampleMR`, and the MR-Base web platform[?]. 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[?]. 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)[?]. 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 **??**.

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[?]:

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 hierarchical horseshoe, building on the work of Piironen & Vehtari[?]) 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[?]. 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[?,?].

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[?].

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[?]. Full details of simulation reproduction, including code and validation of outputs, is presented in Appendix **??**.

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 ($\hat{\gamma}$), and a separate cohort of 10,000 used to estimate gene-outcome association ($\hat{\Gamma}$). Parameter values for effect allele frequency were not specified by Bowden et al, though initial testing showed values around 0.5 produced WME causal effect estimates closest to published values when other parameters were matched[?]. 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[?,?]. 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[?] 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[?].

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

Inclusion criteria:

- Original two-sample MR study
- Able to determine samples' ancestry sufficient to establish presence/potential degree of participant overlap between groups
- Reporting $\geq$ 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)[?].

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

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

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 **??**. 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 **??**). 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 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 1), 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.

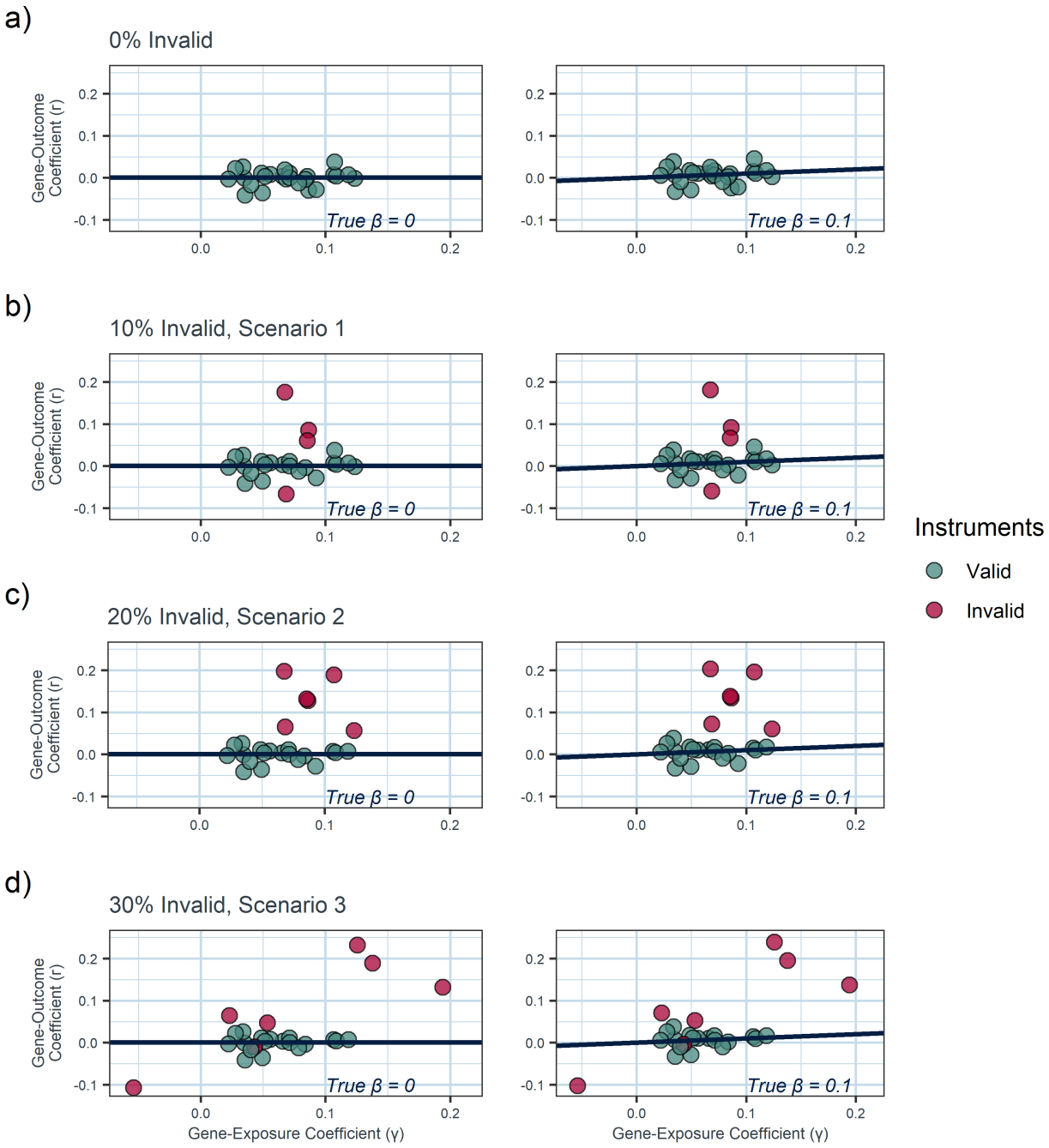

Figure 3: Plots of a representative group of simulated datasets; all simulate genetic instruments from the same index from the same random seed. Left and right columns demonstrate null and positive true causal effects, respectively; the true causal effect is represented by the gradient of the line shown. 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.

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 R^2 statistics presented are the minimum values across all simulated datasets.
Null Causal Effect ($\beta = 0$)

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 ($\beta = 0.1$)