Causal Effect Estimation in Mendelian Randomisation Studies -

Evaluating a Novel Bayesian Approach To Genetic Pleiotropy

Versus Established Weighted Median Methodology

B233241

September 2024 - June 2025

**Contents**

Acknowledgements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

Statement of Originality . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

Word Count . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

**1 Introduction and Background 4**

1.1 Introduction to Mendelian Randomisation (MR) . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.2 Causal Effect Estimation in MR . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

1.3 Violations to Assumptions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

1.4 Weighted Median Estimator (WME) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

1.5 Issues With WME CIs . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

1.6 MR-Hevo . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

1.7 Aims and Objectives . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

**2 Methods 10**

2.1 Simulation Study . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

2.2 Re-Analysis of Published Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.3 Data Manipulation and Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.4 Ethical Approval . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**3 Results 13**

3.1 Simulation Study . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

3.2 Re-Analysis of Published Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

1

**4 Discussion 22**

4.1 Performance of Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 22

4.2 Results in Context . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23

**5 Limitations and Recommendations 25**

5.1 Limitations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 25

5.2 Recommendations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26

**6 References 27**

**A Appendix: List of Abbreviations 33**

**B Appendix: Bootstrapping 34**

B.1 Bootstrapping - General Method . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

B.2 Bootstrapping - Example: Prostate Volume . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

B.3 Bootstrapping - Relevance to WME . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36

**C Appendix: Simulation Code 37**

C.1 Generating Data and Models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

C.2 Testing Generation of Data and Models . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

C.3 Summary Table . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50

**D Appendix: R Packages Used 54**

D.1 Package Citations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54

D.2 Session Information . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 54

2

**Acknowledgements**

I would like to acknowledge

**Contributions**

Mine others

**Statement of Originality**

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

**Word Count**

Word count: 8371

3

**1 Introduction and Background**

**1.1 Introduction to Mendelian Randomisation (MR)**

Epidemiology is the study of determinants and distribution of disease across populations; a common epidemiological

study aim is therefore to seek evidence as to whether a given exposure (e.g. cigarette smoking) may

cause a given outcome (e.g. lung cancer)1. Logistics limit experimental interventions across large groups,

so insights into associations between exposures and outcomes are gleaned from observational data of people

in the population of interest. Comparing health outcomes between individuals with different levels of a

particular exposure may highlight potential links, e.g. higher cancer incidence in those who smoke more is

consistent with a causal role for cigarettes in carcinogenesis1.

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 observed2. 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 bias4. Statistical methods can quantify the probability that any observed outcome differences

could have occurred by chance, and thereafter any outcome differences can be interpreted as caused by

exposure differences. As allocation and receipt of exposures is known to precede outcome measurements,

reverse causality is impossible.

In MR, naturally occurring genetic variants - “genetic instruments” – are chosen based on their known

association to an exposure of interest. Provided that assumptions of IV analysis are met, random assignment

of alleles (i.e. variants of a given gene) from parents to offspring during meiosis creates randomisation

analagous to that performed for an RCT – both measured and unmeasured confounders should be distributed

evenly between the groups created, allowing valid causal inference after other sources of bias and random

variation are accounted for5.

**1.2 Causal Effect Estimation in MR**

At its simplest, the relationship between two continuous variables - an exposure 𝑋 and outcome 𝑌 - can be

represented as a linear model:

𝑌 = 𝛼 + 𝛽𝑋 + 𝜖 (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:

• 𝛽 = 0 implies no causal link between exposure and outcome

• 𝛽 > 0 implies 𝑋 causes 𝑌

4

• 𝛽 < 0 implies 𝑋 prevents 𝑌

To estimate a causal effect using a genetic variant in an IV analysis, three key assumptions must be met6:

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 20167

Typically, MR studies estimate causal effect using a set of several genetic instruments; the causal effect

estimate derived from the 𝑗𝑡ℎ instrument is denoted 𝛽𝑗̂ . Each estimate 𝛽𝑗̂ acknowledges there will be specific

effects on the observed values of exposure and outcome given the presence of that specific genetic variant 𝐺𝑗

under study, i.e. 𝛽𝑗̂ is based on the instrument-conditioned exposure 𝑋|𝐺𝑗 and the instrument-conditioned

outcome 𝑌 |𝐺𝑗. These observed values of exposure and outcome can be described by their own linear models:

5

𝑋|𝐺𝑗 = 𝛾0 + 𝛾𝑗𝐺𝑗 + 𝜖𝑋𝑗 (2)

𝑌 |𝐺𝑗 = Γ0 + Γ𝑗𝐺𝑗 + 𝜖𝑌𝑗 (3)

where, for exposure and outcome respectively:

• 𝛾0 and Γ0 reflect base values without any influence from the effect allele of the genetic variant

• 𝛾𝑗 and Γ𝑗 are coefficients of association with the genetic variant, representing the extent to which an

effect allele of 𝐺𝑗 will perturb the value of 𝑋 or 𝑌 versus the non-effect allele

• 𝜖𝑋𝑗 and 𝜖𝑌𝑗 are error terms, containing contributions from confounders of the exposure-outcome relationship

(𝑈 in the causal diagram), and all genetic variants except 𝐺𝑗.

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.:

𝛽𝑗̂ = Γ̂𝑗 𝛾𝑗̂(4)

These coefficients of gene-exposure and gene-outcome association (𝛾̂ and Γ̂) can be obtained from a genomewide 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 𝛽̂ from any given MR method will typically seek to pool effect estimates from several instruments so as to minimise effects of any invalid instruments included, e.g. by removing/down-weighting contributions of genetic instruments which violate one or more assumptions. This is equivalent to plotting all estimated coefficients of gene-outcome association (Γ̄) versus all estimated coefficients of gene-exposure association (𝛾̄) for the set of instruments, then using the gradient of a regression line through the points as the causal effect estimate 𝛽;̂ picking an MR methodology is analogous to choosing the method to draw the line of best fit (Figure 2). For binary outcomes, the causal effect estimate can be converted to an odds ratio (OR) through exponentiation, i.e.:

𝑂𝑅 = 𝑒𝛽̂ (5)

**1.3 Violations to Assumptions**

In practice, only the relevance assumption can be directly tested and proven. Typically, genetic variants for

MR studies are selected as instruments based on 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 𝑟2

statistic, representing the proportion of variance in the exposure explained by the genotype, and the related

𝐹 -statistic, which additionally accounts for the the sample size under investigation9. An 𝐹 -statistic of ≥ 10 is

generally considered to represent a strong enough gene-exposure association to consider a genetic instrument

for use2.

The assumptions of independence and exclusion restriction depend on all possible confounders of the

exposure-outcome association, both measured and unmeasured; as such, these can never be proven absolutely.

Various methods have been proposed to quantify and account for violations of these two additional

assumptions, including the weighted median estimator, described below8.

6

Figure 2: Simulated MR Study on 10,000 individuals using 25 genetic instruments, of which 30% are invalid

(red points) and introduce directional pleiotropic effects. The true value of the exposure-outcome causal

effect is 0.25 (grey line, causal effect represented by gradient). Regression using an unajusted least-squares

linear model (light blue line) results in a biased estimate in the positive direction due to the influence of the

invalid instruments. Using the Weighted Median Estimator method (pink line) attenuates the effects of the

invalid instruments, resulting in an estimate closer to the true value. Adapted from Bowden et al 20168

7

The main methods to avoid violations of the independence assumption relate to appropriate selection of

populations studied to avoid confounding due to ancestry or population stratification. For example, in twosample

MR studies, where gene-exposure and gene-outcome coefficients are estimated from two separate

GWAS studies, it is recommended to select GWAS studies performed in similar population groups (e.g. both

in Western Europeans). This practice helps avoid spurious exposure-outcome associations being generated

by confounding due to underlying differences in e.g. allele frequency, baseline disease risks etc between

ancestrally different populations9.

Exclusion restriction is a particularly universal issue in MR, due to so-called (horizontal) genetic pleiotropy,

where a single genetic variant may have multiple “pleiotropic” effects – i.e. it may influence several traits

simultaneously. Such pleiotropic effects may be unknown and open unmeasured causal pathways between

a genetic instrument and the outcome (Figure 1), thus potentially biasing MR estimates of the association

between exposure and outcome. As pleiotropy influences outcome separate to the path involving the exposure

of interest, the term “direct effects” is also used10. Where pleiotropic effects are in both positive and negative

directions with a mean of zero - “balanced pleiotropy” - then they only add noise to causal effect estimation11.

By contrast, “directional pleiotropy”, where the mean of pleiotropic effects is non-zero, may introduce bias8

(Figure 2).

If such an additional causal pathway acts between gene 𝐺 and outcome 𝑌 via a confounding factor 𝑈 , then

the magnitude of direct/overall effects of 𝐺 on 𝑌 will correlate with the effects of 𝐺 on 𝑋 (i.e. Γ ∝ 𝛾),

and “correlated pleiotropy” is present. If an additional causal pathway acts directly between gene 𝐺 and

outcome 𝑌 independent of both exposure 𝑋 and confounders 𝑈 , this results in “uncorrelated pleiotropy”

(Figure 1). Both correlated and uncorrelated pleiotropy can introduce bias which distorts the estimate of the

true causal effect. In general, correlated pleiotropy is more challenging to account for; several MR methods

explicitly require an additional assumption of Instrument Strength Independent of Direct Effect (InSIDE),

i.e no correlated pleiotropy to be present12.

**1.4 Weighted Median Estimator (WME)**

A common approach to produce exposure-outcome causal effect estimates robust to violations of the exclusion

restriction assumption is the weighted median estimator (WME) method, proposed by Bowden et al8.

In WME analysis, several genetic instruments are used to estimate the exposure-outcome causal effect 𝛽.̂

Each instrument is known to be associated with the exposure of interest, but an unknown proportion of

these instruments may be invalid due to pleiotropic genetic effects. Any instrument linked to an outcome

via multiple pleiotropic causal pathways will exhibit a less consistent gene-outcome association than a relationship

mediated by a single pathway; this results in larger variance in causal estimates derived from

invalid/pleiotropic genetic instruments versus estimates from valid instruments.

WME therefore assigns a weight to each genetic instrument’s estimate of the causal effect according to the

inverse of the variance of the estimate; these weighted effect estimates are used to construct a cumulative

distribution function for probability of true causal effect size across the range of estimated values. The 50th

percentile of this distribution can then be taken as a “weighted median estimate” of the true causal effect,

theoretically producing consistent causal estimates even if up to 50% of the included information comes from

invalid instruments8. An example of WME attenuating the effects of invalid instruments is shown in Figure

2.

**1.5 Issues With WME CIs**

WME calculation methods are available via several prolific MR tools: the R packages “MendelianRandomization”13

and “TwoSampleMR”, and the MR-Base web platform14. However, these implement the original authors’

suggested process of generating 95% confidence intervals for WME, which deviates from accepted re-sampling

methodology:

8

“We found the bootstrap confidence interval…too conservative. However, the bootstrap standard

error… gave more reasonable coverage using a normal approximation (estimate ±1.96 x standard

error) to form a 95% confidence interval”8

This modification, explicitly aiming to boost estimate precision artificially, would be expected to lead to a

high Type 1 error rate, which has been a growing concern in the field of late15. The theoretical issues with

this approach, and the fundamentals of bootstrapping in general, are covered in Appendix B.

**1.6 MR-Hevo**

MR-Hevo is an R package which uses more typical Bayesian methodology to estimate MR causal effects and

corresponding 95% confidence intervals. It uses the probabilistic programming language, Stan, to directly

sample the posterior probability distribution of pleiotropic effects on the outcome, rather than making

untested assumptions about the shape of this distribution as current WME implementations do16 (Appendix

B).

MR-Hevo incorporates several additional features which its creators claim further aid valid causal inference.

Most MR methods can only account for one genetic variant per genetic locus (i.e. per location in the genome).

If multiple variants exist at a given locus, generally only one can be selected as an instrument for further

MR analysis. HR-Hevo handles multiple instruments per genetic locus via scalar construction, essentially

assigning a “score” to each locus based on the variant(s) present, thus incorporating more information than

if closely grouped variants had been discarded16. As MR-Hevo is based on a Bayesian approach, it generates

estimates which incorporate relevant existing information generated by prior studies, increasing the amount

of data informing each estimate. In this case, MR-Hevo bases estimates on a prior probability distribution

which reflects existing knowledge that most individual genetic instruments will have only small effects on

complex traits17,18, further aiding biologically plausible inference regarding distribution of pleiotropic effects.

**1.7 Aims and Objectives**

The main aim of this study will be to demonstrate if the WME approach gives over-confident causal estimates

in the presence of pleiotropy, and whether this issue is more correctly handled by the MR-Hevo Bayesian

approach. This will be achieved through addressing the research questions and objectives as outlined below:

**Research Questions:**

1. How does MR-Hevo perform versus the weighted median estimator when estimating causal effects in

MR studies?

2. Do conclusions of existing MR studies using weighted median causal effect estimation change if MRHevo

methods are used?

**Objectives:**

1. Quantify the accuracy and precision of MR-Hevo causal estimates for simulated data under differing

sets of common assumptions, with reference to the weighted median estimator

2. Evaluate the consistency of MR-Hevo causal estimates for simulated data under differing sets of common

assumptions, with reference to the weighted median estimator

3. Compare the conclusions drawn from MR-Hevo causal effect estimation versus the weighted median

estimator on real-world data

9

**2 Methods**

**2.1 Simulation Study**

To establish the performance of MR-Hevo causal estimation relative to WME, the accuracy, precision and

consistency of both methods were evaluated using simulated datasets with known parameter values.

**2.1.1 Data Simulation**

To aid comparability with existing methods and literature, the simulation methodology of the original WME

exposition was reproduced based on published models and parameters in Appendix 3 of its supplementary

materials8. 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, and each increasingly challenging to the performance of

any given MR causal estimation methodology:

1. Balanced pleiotropy, InSIDE assumption satisfied - A proportion of invalid genetic instruments are

present and introduce pleiotropic effects uncorrelated with the instrument strength; these pleiotropic

effects are equally likely to be positive as negative with a mean value = 0, thus introducing noise into

the estimation of causal effect.

2. Directional pleiotropy, InSIDE assumption satisfied - A proportion of invalid genetic instruments are

present and introduce pleiotropic effects uncorrelated with the instrument strength; these pleiotropic

effects are positive only, with a mean value > 0, thus biasing the causal effect estimate in a positive

direction.

3. Directional pleiotropy, InSIDE assumption not satisfied - A proportion of invalid genetic instruments

are present and introduce pleiotropic effects correlated with the instrument strength through action

via a confounder; these pleiotropic effects are positive only, with a mean value > 0, thus potentially

biasing the causal effect estimate in a positive direction to an even greater extent than Scenario 2.

1,000 simulated datasets of participant-level data were generated for every combination of each scenario and

each the following simulation parameters:

• Proportion of invalid instruments: 0%, 10%, 20% or 30%

• Number of participants: 𝑛 = 10, 000 or 𝑛 = 20, 000

• Causal effect: null (𝛽 = 0) or positive (𝛽 = 0.1)

The same set of 25 simulated genetic instruments were used across all datasets, with the status of each

as valid/invalid determined by random draw per instrument at the start of each simulation run of 1,000

datasets.

Genotypes were simulated as for a two-sample setting: where number of particpants was 𝑛 = 10, 000, 20,000

genotypes were simulated - 10,000 for the cohort used to estimate gene-exposure association (𝛾̂), and a

separate cohort of 10,000 used to estimate gene-outcome association (Γ̂). 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 matched8. As such,

effect allele frequencies were assigned per instrument from a uniform distribution between 0.4 to 0.6. Each

effect allele freqency thus generated per instrument was then used as a probability to assign each simulated

participant effect alleles for each instrument via two draws from a binomial distribution.

10

**2.1.2 Analysis of Simulated Data**

Each dataset generated was analysed using both WME and MR-Hevo methods, via functions from the

TwoSampleMR and mrhevo packages, respectively14,16. Results were aggregated per group of 1,000 simulated

datasets corresponding to a particular combination of scenario and parameter values. This resulted in one

meta-analysis reported per combination of scenario/parameter values, each including 1,000 simulated MR

studies using the same 25 genetic instruments in the same population. Aggregated measures for both WME

and MR-Hevo per meta-analysis were mean causal effect estimate; mean standard errors/confidence interval

(CI)s 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.

Results of the above aggregations were tabulated as per Tables 2 and 3 of Bowden et al8 to allow direct

comparisons of both methods versus each other and versus the published characteristics of existing MR

causal estimation methods.

**2.2 Re-Analysis of Published Data**

To investigate the potential implications of any differences in performance between WME and MR-Hevo

methods, a selection of published studies resporting causal effect estimates using the WME method was reanalysed.

A sample size of 10 published studies was decided as a pragmatic compromise between the scope

of this study and the need to check consistency of any observed differences. In the original Bowden et al

simulation studies, the WME causal estimation method was shown to generate a false-positive report rate of

≥ 30% with relatively minor violations of relevant assumptions8; therefore, even this relatively small sample

of 10 studies might be expected to demonstrate differences between methods if the MR-Hevo approach is as

appropriately conservative as its creators propose.

To estimate the upper bound of the potential impact of MR-Hevo versus existing WME methodology, studies

were chosen for re-analysis based on their number of citations in the wider MR literature. Compared to

studies with few or no citations, highly-cited studies would be expected to have a larger impact on their

respective fields if their conclusions were to change. In addition, highly-cited works will typically have

been submitted to more scrutiny than less-cited works - both during peer review whilst under consideration

by journals likely to produce highly-cited works, and from the wider scientific community following the

widespread dissemination evidenced by a high citation count. As such, it would be expected that highlycited

works are likely to be free of significant methodological flaws which may impede interpretation of any

re-analysis.

**2.2.1 Citation Search**

The Scopus search platform19 was used on 15/04/2025 to retrieve all articles citing the original weighted

median estimator exposition paper8. 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 platform20. Abstracts were screened by a single reviewer (B233241),

starting with the most cited article and proceeding in descending order of citation count, against the following

inclusion and exclusion criteria:

Inclusion criteria:

• Original two-sample MR study

• Able to determine samples’ ancestry sufficient to establish presence/potential degree of participant

overlap between groups

• Reporting ≥ 20 human genetic instruments relating to exposure

• Reports details of effect/non-effect alleles

11

• 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.

**2.3 Data Manipulation and Analysis**

All simulations, data manipulations and data analyses were performed in R version 4.4.1 (2024-06-14 ucrt)21.

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

For citation search data, a standardised data collection form was Microsoft Excel22 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 packages23.

A full list of packages used can be found in Appendix D.

Data were manually screened at summary level and relevant features were extracted. Data were checked for

completeness, consistency, duplicate values and plausibility. Data were transformed to an appropriate data

type, and encoding of genetic variables was standardised into a single format. Missing values for association

coefficients and standard error (SE)s were imputed as the mean value calculated per dataset. It was noted

during early testing that causal effect estimation functions did not operate correctly in the presence of zerovalue

coefficients of genetic association and/or their standard errors; such zero values were therefore re-coded

as an arbitrarily low value of 10−100.

**2.4 Ethical Approval**

The protocol for this work has been reviewed and approved by the Usher Masters Research Ethics Group

(UMREG) at the University of Edinburgh, Ethics ID: UM241126. Due to the nature of the project, using

simulated and publically available data only, no significant ethical issues were foreseen, and sponsorship was

deemed unnecessary by the UMREG reviewing panel.

12

**3 Results**

**3.1 Simulation Study**

**3.1.1 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 𝐹 -statistic calculated from simulated instruments was >10, indicating that they were sufficiently

strongly associated with exposure to meet the relevance assumption of IV analysis (Tables 1 and 2).

**3.1.2 Analysis of Simulated Data**

**3.1.2.1 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%).

For both MR-Hevo and WME methods, false-positive report rates generally increased with an increasing

proportion of invalid instruments up to around 20% invalid IVs. As assumption violations progressively

presented greater data variability and bias across scenarios 1 to 3, both MR-Hevo and WME methods

tended to exhibit higher false-positive report rates, though this progression was noticably attenuated for

MR-Hevo versus WME, particularly under the assumptions of Scenario 3. Both trends across invalid instrument

proportions and scenarios were somewhat attenuated by increasing sample size from 10,000 to 20,000

participants for both methods.

The mean causal effect estimate (mean reported 95% CIs) across all cases was 0.04 (-0.11 to 0.2) for MRHevo

and 0.039 (-0.11 to 0.19) for WME. For SE, the mean (range) SE of causal effect estimates across all

cases was 0.0012 (0 to 0.002) for MR-Hevo and 0.076 (0.056 to 0.099) for WME.

Causal effect estimates, width of CIs and SE all tended to increase slightly for each method, both with an

increasing proportion of invalid instruments up to 20% invalid IVs, and as assumption violations progressively

presented greater data variability and bias across scenarios 1 to 3. For both these trends, MR-Hevo estimates

tended to be more affected than those from WME, in contrast to the false-positive report rates, though MRHevo

causal effect estimates were once more relatively less affected by Scenario 3 assumptions. Again, both

trends across differing scenarios and invalid instrument proportions were somewhat attenuated by increasing

sample size from 10,000 to 20,000 participants for both methods.

**3.1.2.2 Positive Causal Effect**

Across all cases where no causal effect was present (Table 2), the mean rate of reporting a causal effect

(i.e. sensitivity) for MR-Hevo was 31% versus 28% for WME. Of the 24 combinations of scenarios and

parameters, MR-Hevo exhibited a favourable sensitivity versus WME in 10 (42%).

For both MR-Hevo and WME methods, causal report rates increased with an increasing proportion of invalid

instruments up to around 20% invalid IVs, though this was more consistent for WME versus MR-Hevo. As

assumption violations progressively presented greater data variability and bias across scenarios 1 to 3, both

MR-Hevo and WME methods tended to exhibit higher causal report rates. Both trends across differing

scenarios and invalid instrument proportions were somewhat attenuated by increasing sample size from

10,000 to 20,000 participants for both methods, which also generally increased sensitivity for each method.

The mean causal effect estimate (mean reported 95% CIs) across all cases was 0.13 (-0.025 to 0.3) for MRHevo

and 0.11 (-0.039 to 0.26) for WME. For SE, the mean (range) SE of causal effect estimates across all

cases was 0.0013 (0.001 to 0.002) for MR-Hevo and 0.077 (0.056 to 0.1) for WME.

Causal effect estimates, width of CIs and SE all tended to increase slightly for each method with an increasing

proportion of invalid instruments up to 20-30% invalid IVs; MR-Hevo estimates tended to be more affected

13

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 introduces noise around the causal effect. 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.

14

by proportion of invalid instruments compared to WME estimates. As assumption violations progressively

presented greater data variability and bias across scenarios 1 to 3, WME causal estimates tended to increase

across all three; MR-Hevo estimates increased when switching from Scenario 1 to Scenario 2, but were

relatively unaffected in Scenario 3 versus Scenario 2. Again, trends across invalid instrument proportions

were somewhat attenuated by increasing sample from 10,000 to 20,000 participants for both methods, though

the effects of sample size on trends across scenarios was less obvious.

15

Table 1: Summary of 1,000 simulated Mendelian randomisation

studies per combination of scenario and parameters, all with null

causal effect

Weighted MR

Median Hevo

Mean Estimate Mean Causal Mean Estimate Mean Causal

Report Report

*N* Invalid

IVs *F R2*

(Mean

SE) 95% CI

Rate

(Mean

SE) 95% CI

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 (0.001) -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 (0.001) -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 (0.002) -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 (0.002) -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) -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 (0.001) -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 (0.001) -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 (0.001) -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 (0.001) -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 (0.001) -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 (0.002) -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 (0.002) -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) -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 (0.001) -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 (0.001) -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 (0.002) -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 (0.001) -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 (0.001) -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 (0.002) -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 (0.002) -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) -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 (0.001) -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 (0.001) -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 (0.001) -0.09 to 0.22 0.2%

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

Null Causal Effect (β = 0)

16

Table 2: Summary of 1,000 simulated Mendelian randomisation

studies per combination of scenario and parameters, all with positive

causal effect

Weighted MR

Median Hevo

Mean Estimate Mean Causal Mean Estimate Mean Causal

Report Report

*N* Invalid

IVs *F R2*

(Mean

SE) 95% CI

Rate

(Mean

SE) 95% CI

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 (0.001) -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 (0.001) -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 (0.002) -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 (0.002) -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 (0.001) 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 (0.001) -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 (0.001) -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 (0.001) -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 (0.001) -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 (0.001) -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 (0.002) -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 (0.002) 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 (0.001) 0 to 0.18 62.7%

20,000 10% 24.5 3% 0.093 (0.063) -0.03 to 0.22 24% 0.109 (0.001) -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 (0.001) -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 (0.002) 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 (0.001) -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 (0.001) -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 (0.002) -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 (0.002) -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 (0.001) 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 (0.001) 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 (0.001) 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 (0.001) -0.01 to 0.32 41%

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

Positive Causal Effect (β = 0.1)

17

**3.2 Re-Analysis of Published Data**

**3.2.1 Citation Search Results**

A total of 110 abstracts and 54 full texts were screened to identify the 10 studies included24–33; these studies

are summarised in Table 3. The flow diagram of study screening and selection is presented in Figure 4.

Figure 4: Flow diagram illustrating selection of sample of ten highly-cited two-sample Mendelian randomisation

articles reporting a weighted median estimate of casual effect

18

Table 3: Summary of ten highly-cited two-sample Mendelian randomisation

articles reporting a weighted median estimate of casual

effect

Participants Causal

*N* Maximum Effect

Estimated

Author Citations Association *N*

Instruments

Exposure Outcome

Overlap

Measure Estimate

Causality

Reported *p*-value

Budu-Aggrey

et al, 2019

182 BMI vs Psoriasis 97 339,224 12,559 0% OR 1.06 (1 to 1.12) No -

Carreras-

Torres et al,

2017

200

Height vs

Pancreatic Cancer

558 253,288 15,002 19% OR 1.14 (1 to 1.29) No 0.05

Carter et al,

2019

199

Education vs

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 vs

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

vs 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 vs

Schizophrenia

52 204,402 82,315 0% OR

0.89 (0.81 to

0.96)

Yes 0.004

Mokry et al,

2016

199

BMI vs 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 vs

Cannabis Use

102 150,064 184,765 0% *β*

0.163 (0.067 to

0.259)

Yes 0.001

Xie et al, 2023 138 T2DM vs NAFLD 449 441,016 218,792 0% OR

1.61 (1.09 to

2.38)

Yes <0.001

Xu et al, 2022 183

Coeliac vs 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

19

**3.2.2 Re-Analysis Results**

**3.2.2.1 Data Validation and Re-Analysis**

There were missing gene-outcome coefficients for three instruments from Xie et al32, and one instrument

in Clift et al33 was reported as having an implausibly large gene-outcome coefficient and standard error

(-1243.03 and 19161.64, respectively); these were imputed as the respective mean value per study. 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 odds ratio (OR)s

to aid comparison across studies.

**3.2.2.2 Re-Analysed vs Reported WME Causal Estimates**

3 of the WME estimates generated through re-analysis matched the originally reported estimates poorly

(Ligthart et al26, Carreras-Torres et al29, Mokry et al28), 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 al26, Carreras-Torres et al29, Mokry et al28, Budu-Aggrey et al31). 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 al24, Xie et al32, Pasman et

al25, Carter et al27, Clift et al33, Xu et al30) 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 al26), and 1 where a causal effect was found that

had not been reported previously (Mokry et al28).

**3.2.2.3 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 al24, Carreras-Torres et al29). Compared with WME values of ORs across the 9 studies

using them, the mean difference for effect estimates (SE of estimate) from the re-analysis estimate was -

0.046 (-0.084). For 95% CIs, the mean differences between MR-Hevo and WME values were -0.044 for the

lower bounds and -0.002 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 al24, Xie et al32, Ligthart et al26, Carreras-Torres et al29, Clift et al33, Mokry

et al28).

Overall, estimates and CIs from MR-Hevo analysis of the other 4 studies (Pasman et al25, Carter et al27,

Budu-Aggrey et al31, Xu et al30) 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 al32, Pasman et al25, Carter et al27, Clift et al33, Mokry et al28).

20

Table 4: Re-analysis of ten highly-cited two-sample Mendelian randomisation

articles reporting a weighted median estimate of casual

effect, comparing results of both WME and MR-Hevo causal effect

estimation methods

Weighted

Median

MR-Hevo

Author Exposure Outcome SNPs

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

(-0.17-0.33)

0.002

1.08

(0.84-1.39)

No

Carreras-Torres

et al

Height Pancreatic Cancer 558 0 (-0.13-0.13) 0.066 1 (0.88-1.14) No

-0.28

(-1.34-0.5)

0.513

0.76

(0.26-1.64)

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)

0.000

0.62

(0.58-0.66)

Yes

Choi et al

Self-Reported Physical

Activity

Major Depressive

Disorder

25

0.39

(-0.06-0.83)

0.227

1.47

(0.94-2.29)

No

0.22

(-0.23-0.65)

0.004

1.25

(0.8-1.91)

No

Clift et al

Genetically Determined

Smoking Initiation

COVID-19 Infection 378

0.43

(0.02-0.84)

0.209

1.53

(1.02-2.31)

Yes 0.37 (0.1-0.64) 0.001 1.45 (1.1-1.9) Yes

Ligthart et al Genetically Determined CRP Schizophrenia 29

-0.41

(-0.88-0.08)

0.245

0.67

(0.41-1.08)

No

-0.38

(-1.24-0.54)

0.008

0.68

(0.29-1.72)

No

Mokry et al BMI Multiple Sclerosis 70

0.34

(0.09-0.59)

0.129

1.41

(1.09-1.81)

Yes

0.34

(0.16-0.52)

0.001

1.41

(1.17-1.67)

Yes

Pasman et al Liability to Schizophrenia Cannabis Use 109

0.16

(0.06-0.26)

0.050

1.18

(1.07-1.3)

Yes

0.17

(0.08-0.26)

0.001

1.18

(1.08-1.29)

Yes

Xie et al T2DM NAFLD 526

0.48

(0.09-0.87)

0.198

1.61

(1.09-2.38)

Yes

0.51

(0.28-0.75)

0.002

1.67

(1.32-2.13)

Yes

Xu et al Coeliac Disease Gut Bifidobacterium 105 0 (-0.01-0) 0.004 1 (0.99-1) No 0 (-0.01-0) 0.000 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.

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

21

**4 Discussion**

**4.1 Performance of Methods**

**4.1.1 Simulation**

**4.1.1.1 Null Causal Effect**

For a null causal effect, MR-Hevo exhibited comparable accuracy to WME, with both methods tending to

slightly over-estimate the true value. Considering precision, the mean SE for MR-Hevo was almost 2 orders

of magnitude smaller than for WME, though mean CIs were similarly consistent between methods despite

this, reflecting differing methodology used in CI construction (see Appendix B). Despite near identical quantitative

effect estimation, MR-Hevo was consistently more accurate in categorical classification as causal/no

causal effect, exhibiting a superior false-positive rate in all 24 meta-analyses with null causal effect. This

performance lends credence to the claims of MR-Hevo’s creators regarding the importance of CI construction

method being dissociated from SE estimation16.

Regarding consistency, there is a trend towards both methods reporting a wider CI with conditions promoting

a greater effects of invalid instruments, as might be expected. However, MR-Hevo CIs appear to widen slightly

more than for WME, both with increasing proportion of invalid instruments and progressive violation of IV

assumptions. This does suggest that MR-Hevo deals with pleiotropic effects more conservatively than WME,

namely by appropriately reducing the reported precision of the estimate to reflect additional uncertainty.

Greater consistency of effect estimation by MR-Hevo versus WME was particularly marked when moving

between different scenarios representing different assumption violations.

**4.1.1.2 Positive Causal Effect**

When a positive causal effect was present, MR-Hevo exhibited a slightly higher mean sensitivity than WME

when all results were pooled, but on per meta-analysis basis it was out-performed in the majority of cases.

The accuracy of MR-Hevo versus WME was very slightly lower overall, again tending to over-estimate the

true causal effect. Precision was similar to the null causal effect case, with a comparable CI for both methods,

but a much smaller SE for MR-Hevo.

Regarding consistency, there are again trends towards both methods reporting wider CIs with a greater

proportion of invalid instruments and/or greater violations of IV assumptions; again, this broadening of

CIs is more marked for MR-Hevo than for WME. An exception to general trends is the combination of 0%

invalid instruments 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. This parameter combination appears to drive

the somewhat discordant summary of sensitivity results.

It is not clear why this combination is associated with such a high causal report rate for MR-Hevo. If MRHevo

performed particularly well versus WME in the absence of invalid instruments, this would be expected

to hold in the 10,000 participants case also. Similarly, if MR-Hevo were particularly sensitive to the difference

in sample size versus WME, larger discrepancies would be expected between the two methods with other

parameter/scenario combinations when transitioning between 10,000 to 20,000 participants. Differences in

assumption violations between scenarios do not affect this result, as assumption violations are only relevant

to invalid instruments, of which there are none in the 0% invalid case. This unexpected result may be an

aberrant feature of the particular datasets generated, which could be investigated by re-running the analysis

from a different random seed. Alternatively, it may be that, when using MR-Hevo methods, sample size

interacts in a non-linear way for and invalid instrument proportions approaching zero with respect to the

method’s power. If the causal report rate for this parameter/scenario combination remained high after data

simulation with a different seed, this possibility could be next investigated using simulated datasets with

invalid instrument proportions between 0-10%.

22

**4.1.2 Re-Analysis**

As discussed in 5.1.2, 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 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 al26) 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 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 different methodologies. However, with thousands of MR studies now reported, as evidenced by the

5,417 articles referencing Bowden et al8 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.

**4.2 Results in Context**

A key concern in the MR literature of late has been of suspiciously high numbers of studies reporting causal

effects, often in cases where causality does not seem biologically plausible15. It is against this backdrop that

the creators of MR-Hevo introduce their approach as a potential solution, and it is worth considering this

wider context before assessing the relative merits of each method.

Several factors may be driving high positive report rates observed in published MR studies. As with other

academic fields, there is likely to be an element of publication bias in favour of studies reporting statistically

significant results34,35; 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 package14 facilitate production of

MR studies at scale. Genetic studies without a plausible hypothetical basis are at high baseline risk of false

positives due to implicit multiple comparisons, given the number of potential genes and/or phenotypes which

could be examined36. 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

practice37:

“…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”38

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”15, 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 finding39.

23

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 violated16. 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 improvement in

consistency but fall in sensitivity versus WME. In the above “everything against everthing” pre-print37, it

is significant that evidence of pleiotropy was noted in >90% of comparisons on the MR-Base platform. The

prescence of balanced pleiotropy would be expected to add only noise 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.

24

**5 Limitations and Recommendations**

**5.1 Limitations**

**5.1.1 Simulation Study**

Key objectives of this study were to evaluate the accuracy, precision and consistency of MR-Hevo causal

estimation 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 noise and bias

to the causal effect estimates generated. If the average noise 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 noise 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 speculating

on trends with ≥ 30% invalid instruments from these data alone would seem inadvisable. If desired, these

simulation studies 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 al8 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 al8. 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 [Session\_Information]), each meta-analysis took in the order of 8 hours to

process 𝑛 = 1, 000 datasets; assuming 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”40. 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 al8. 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.

**5.1.2 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 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 employee

exclusion criteria for studies whose estimates are not replicable within a specified error margin.

25

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 abserved across all scenarion/parameter combinations in the simulation study (MR-Hevo = 0.41% vs

WME = 5.1%, 3.1.2.1), to detect a difference of this size with 𝛼 = 0.05 and 1 − 𝛽 = 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 might not change conclusions in a substantial number of studies; as stated in Methods, the intent

of this project was to help delineate the upper bound of the potential effect of MR-Hevo on the field of MR.

**5.2 Recommendations**

The results of this study 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 around 25% - though with a potential loss of statistical

power of 10-20%. Given the field-wide concerns regarding high false-positive rates, this may be a reasonable

compromise.

Following this study, 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.

• 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 𝛽 = ±0.1

• For interpretation of existing MR studies relying on WME causal effect estimation, re-analysing using

MR-Hevo methods is unlikely to 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, which seems a

worthwhile trade-off given high false positive rates in the MR literature more broadly; and b) the extra

compute required, though for most applications this difference will be trivial.

26

**6 References**

1. Coggon D, Rose G, Barker D. Chapter 1. What is epidemiology? | The BMJ. In: The BMJ |

The BMJ: Leading general medical journal Research Education Comment [Internet]. 2003 [cited

2025 Apr 29]. Available from: https://www.bmj.com/about-bmj/resources-readers/publications/

epidemiology-uninitiated/1-what-epidemiology

2. Martens EP, Pestman WR, Boer A de, Belitser SV, Klungel OH. Instrumental Variables: Application

and Limitations. Epidemiology [Internet]. 2006 May [cited 2025 Apr 29];17(3):260.

Available from: https://journals.lww.com/epidem/fulltext/2006/05000/instrumental\_variables\_

\_application\_and.10.aspx

3. Hernán MA, Robins JM. Instruments for Causal Inference: An Epidemiologist’s Dream? Epidemiology

[Internet]. 2006 Jul [cited 2025 Apr 29];17(4):360. Available from: https://journals.lww.com/

epidem/fulltext/2006/07000/instruments\_for\_causal\_inference\_\_an.4.aspx#JCL1-2

4. Stel VS, Dekker FW, Zoccali C, Jager KJ. Instrumental variable analysis. Nephrology Dialysis

Transplantation [Internet]. 2013 Jul [cited 2025 Apr 29];28(7):1694–9. Available from: https:

//doi.org/10.1093/ndt/gfs310

5. Davies NM, Holmes MV, Smith GD. Reading Mendelian randomisation studies: A guide, glossary,

and checklist for clinicians. BMJ [Internet]. 2018 Jul [cited 2025 Jan 7];362:k601. Available from:

https://www.bmj.com/content/362/bmj.k601

6. Lousdal ML. An introduction to instrumental variable assumptions, validation and estimation.

Emerging Themes in Epidemiology [Internet]. 2018 Jan [cited 2025 Apr 29];15:1. Available from:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5776781/

7. Burgess S, Bowden J, Fall T, Ingelsson E, Thompson SG. Sensitivity Analyses for Robust Causal

Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. Epidemiology

(Cambridge, Mass) [Internet]. 2016 Nov [cited 2024 Oct 22];28(1):30. Available from: https://pmc.

ncbi.nlm.nih.gov/articles/PMC5133381/

8. Bowden J, Smith GD, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization

with Some Invalid Instruments Using a Weighted Median Estimator. Genetic Epidemiology [Internet].

2016 Apr [cited 2024 Oct 22];40(4):304. Available from: https://pmc.ncbi.nlm.nih.gov/articles/

PMC4849733/

9. Richmond RC, Smith GD. Mendelian Randomization: Concepts and Scope. Cold Spring Harbor

Perspectives in Medicine [Internet]. 2022 Jan [cited 2024 Oct 22];12(1):a040501. Available from:

https://pmc.ncbi.nlm.nih.gov/articles/PMC8725623/

10. Hemani G, Bowden J, Smith GD. Evaluating the potential role of pleiotropy in Mendelian randomization

studies. Human Molecular Genetics [Internet]. 2018 May [cited 2024 Oct 23];27(R2):R195.

Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6061876/

11. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting

for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. Nature

Genetics [Internet]. 2020 Jul [cited 2025 May 22];52(7):740–7. Available from: https://www.nature.

com/articles/s41588-020-0631-4

27

12. Grant AJ, Burgess S. A Bayesian approach to Mendelian randomization using summary statistics

in the univariable and multivariable settings with correlated pleiotropy. The American Journal of

Human Genetics [Internet]. 2024 Jan [cited 2025 May 20];111(1):165–80. Available from: https:

//www.cell.com/ajhg/abstract/S0002-9297(23)00433-0

13. Yavorska OO, Burgess S. MendelianRandomization: An R package for performing Mendelian randomization

analyses using summarized data. International Journal of Epidemiology [Internet]. 2017

Dec [cited 2024 Oct 23];46(6):1734–9. Available from: https://doi.org/10.1093/ije/dyx034

14. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform

supports systematic causal inference across the human phenome. Loos R, editor. eLife [Internet].

2018 May [cited 2025 Jan 7];7:e34408. Available from: https://doi.org/10.7554/eLife.34408

15. Stender S, Gellert-Kristensen H, Smith GD. Reclaiming mendelian randomization from the deluge of

papers and misleading findings. Lipids in Health and Disease [Internet]. 2024 Sep [cited 2025 Jan

7];23(1):286. Available from: https://doi.org/10.1186/s12944-024-02284-w

16. McKeigue PM, Iakovliev A, Spiliopoulou A, Erabadda B, Colhoun HM. Inference of causal and

pleiotropic effects with multiple weak genetic instruments: Application to effect of adiponectin on

type 2 diabetes [Internet]. medRxiv; 2024 [cited 2024 Oct 23]. Available from: https://www.medrxiv.

org/content/10.1101/2023.12.15.23300008v2

17. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size

distribution from genome-wide association studies and implications for future discoveries. Nature

genetics [Internet]. 2010 Jun [cited 2024 Oct 23];42(7):570. Available from: https://pmc.ncbi.nlm.

nih.gov/articles/PMC4615599/

18. Piironen J, Vehtari A. Sparsity information and regularization in the horseshoe and other shrinkage

priors [Internet]. arXiv; 2017 [cited 2024 Oct 23]. Available from: http://arxiv.org/abs/1707.01694

19. Scopus - Document search | Signed in [Internet]. [cited 2025 Apr 15]. Available from: https://www.

scopus.com/search/form.uri?display=basic#basic

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

21. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria:

R Foundation for Statistical Computing; 2025. Available from: https://www.R-project.org/

22. Microsoft Corporation. Microsoft Excel [Internet]. 2018. Available from: https://office.microsoft.

com/excel

23. 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.

24. 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

25. 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

28

26. 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

27. 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

28. 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

29. 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

30. 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

31. 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

32. 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

33. 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

34. 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/

35. Bowden J, Jackson D, Thompson SG. Modelling multiple sources of dissemination bias in metaanalysis.

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

36. 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

29

37. 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

38. 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

39. 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/

40. Eddie [Internet]. Edinburgh University DRS. [cited 2025 Jun 10]. Available from: https://

digitalresearchservices.ed.ac.uk/resources/eddie

41. 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

42. 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

43. 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/

44. Higgins P. medicaldata: Data package for medical datasets [Internet]. 2021. Available from: https:

//CRAN.R-project.org/package=medicaldata

45. Chaput R. acronymsdown: Acronyms and glossaries support for RMarkdown [Internet]. 2025. Available

from: https://github.com/rchaput/acronymsdown

46. 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

47. Xie Y. bookdown: Authoring books and technical documents with r markdown [Internet]. 2025.

Available from: https://github.com/rstudio/bookdown

48. 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/

49. Wilke CO. cowplot: Streamlined plot theme and plot annotations for “ggplot2” [Internet]. 2024.

Available from: https://CRAN.R-project.org/package=cowplot

50. Csárdi G. crayon: Colored terminal output [Internet]. 2024. Available from: https://CRAN.Rproject.

org/package=crayon

51. 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

30

52. Gohel D, Skintzos P. flextable: Functions for tabular reporting [Internet]. 2025. Available from:

https://CRAN.R-project.org/package=flextable

53. Yasumoto A. ftExtra: Extensions for “Flextable” [Internet]. 2024. Available from: https://CRAN.Rproject.

org/package=ftExtra

54. Barrett M. ggdag: Analyze and create elegant directed acyclic graphs [Internet]. 2024. Available

from: https://CRAN.R-project.org/package=ggdag

55. Yutani H. gghighlight: Highlight lines and points in “ggplot2” [Internet]. 2023. Available from:

https://CRAN.R-project.org/package=gghighlight

56. Nicholls K. gluedown: Wrap vectors in markdown formatting [Internet]. 2024. Available from: https:

//CRAN.R-project.org/package=gluedown

57. Rodriguez-Sanchez F, Jackson CP. grateful: Facilitate citation of R packages [Internet]. 2024. Available

from: https://pakillo.github.io/grateful/

58. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria:

R Foundation for Statistical Computing; 2025. Available from: https://www.R-project.org/

59. Müller K. here: A simpler way to find your files [Internet]. 2020. Available from: https://CRAN.Rproject.

org/package=here

60. 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.

61. Zhu H. kableExtra: Construct complex table with “kable” and pipe syntax [Internet]. 2024. Available

from: https://CRAN.R-project.org/package=kableExtra

62. 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.

63. Xie Y. Dynamic documents with R and knitr [Internet]. 2nd ed. Boca Raton, Florida: Chapman;

Hall/CRC; 2015. Available from: https://yihui.org/knitr/

64. Xie Y. knitr: A general-purpose package for dynamic report generation in R [Internet]. 2025. Available

from: https://yihui.org/knitr/

65. Bengtsson H. matrixStats: Functions that apply to rows and columns of matrices (and to vectors)

[Internet]. 2025. Available from: https://CRAN.R-project.org/package=matrixStats

66. 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

67. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria:

R Foundation for Statistical Computing; 2025. Available from: https://www.R-project.org/

68. 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

69. 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

31

70. 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

71. Stan Development Team. RStan: The R interface to Stan [Internet]. 2025. Available from: https:

//mc-stan.org/

72. Murdoch D. tables: Formula-driven table generation [Internet]. 2024. Available from: https://CRAN.

R-project.org/package=tables

73. 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

74. 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

75. Marwick B. wordcountaddin: Word counts and readability statistics in r markdown documents [Internet].

2025. Available from: https://github.com/benmarwick/wordcountaddin

32

**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

33

**B Appendix: Bootstrapping**

**B.1 Bootstrapping - 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 follows41:

1. A sample, 𝑥, of 𝑛 individuals is selected from a total population, 𝑋, of 𝑁 individuals

2. This sample 𝑥 is then treated as the “bootstrap population”; the empirical distribution of values in

the 𝑛 individuals in the bootstrap population is taken to be broadly representative of the distribution

of values in the underlying population 𝑋 of 𝑁 individuals

3. A “bootstrap sample”, 𝑥∗, is then obtained by re-sampling individuals from the bootstrap population

with replacement 𝑛 times per bootstrap sample, i.e. the new bootstrap sample comprises 𝑛 sampled

individuals, 𝑥∗

1, 𝑥∗

2, ...𝑥∗

𝑛. As such, individuals from the original bootstrap population 𝑥 may contribute

once, more than once or not at all to each bootstrap sample 𝑥∗.

4. A total of 𝑘 bootstrap samples are generated, 𝑥∗1, 𝑥∗2, ...𝑥∗𝑘, and the statistic of interest (e.g. sample

mean 𝑥)̄ is estimated in each individual sample, 𝑥∗̄ 𝑖, giving the complete set of 𝑥∗̄ 1, 𝑥∗̄ 2, ...𝑥∗̄ 𝑖...𝑥∗̄ 𝑘.

5. The set of 𝑘 statistics are combined to form a “bootstrap distribution”; as expected from central limit

theorem (CLT)42, 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, 𝜇)̂ is taken as the mean

of the bootstrap distribution of 𝑘 estimates, (Σ𝑘

𝑖=1 𝑥)̄ ÷ 𝑛

• the CIs are taken as the values at the appropriate centiles at the edges of the sampling

distribution, e.g. a 95% CI would be generated using values at the 2.5th and 97.5th centiles

• the SE of the estimate is taken as the standard deviation (SD) of the sampling distribution,

given by √ 1

𝑘−1 Σ𝑘

𝑖=1(𝑥𝑖̄ − 𝜇)̂ 2

**B.2 Bootstrapping - Example: Prostate Volume**

The above process is illustrated in 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.

34

Figure 5: Histograms demonstrating distribution of prostate volumes in patients with prostatic cancer, taken

from Cata et al 201143 via the R package medicaldata44. 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.

35

**B.3 Bootstrapping - Relevance to WME**

In current implementations of WME, the WME estimate of the causal effect (𝛽𝑊̂ 𝑀𝐸) is calculated as described

in Bowden et al8, and the 95% CI is generated separately using bootstrapping, though notably not

using the method described above.

The bootstrapping process begins similarly, with re-sampling undertaken (a default of 𝑘 = 1000 times) to

generate 𝑘 bootstrap samples 𝑥∗1, 𝑥∗2, ...𝑥∗𝑘. Each individual bootstrap sample 𝑥∗𝑖 is used to estimate the

causal effect using the WME method 𝛽∗̂ 𝑖

𝑊𝑀𝐸, and thus a bootstrap distribution of 𝑘 values of WME is

created, 𝛽∗̂ 1

𝑊𝑀𝐸, 𝛽̂∗2

𝑊𝑀𝐸....𝛽̂∗𝑘

𝑊𝑀𝐸.

At this stage, however, the bootstrap distribution is then assumed to be approximately normally distributed

without verifying this assumption. The 95% CI of the bootstrap estimate is then calculated as 1.96 SDs of

the bootstrap distribution either side of the mean estimate, i.e. 𝛽𝑊̂ 𝑀𝐸 ± 1.96 × 𝑆𝐸. This approach may be

problematic for several reasons.

Although CLT leads us to expect that the bootstrap distribution will approach normality as the number

of bootstrap iterations 𝑘 increases, the extent to which this occurs for a given 𝑘 may depend on the inital

distribution of values in the population 𝑋, and so also on the distribution in the sample/bootstrap population

𝑥. If the true distribution of values is very non-normal, as may be the case for traits determined by complex

genetic and environmental influences, it may take relatively more bootstrap iterations for the bootstrap

distribution to become sufficiently normal to assume mean and SD accurately describe it.

Additionally, the bootstrap SE is inversely proportional to the number of bootstrap iterations 𝑘, as opposed

to the usual standard error (given by 𝑆𝐸 = 𝑆√𝐷

𝑛 ), which is inversely proportional to the square root of the

sample size 𝑛. It is therefore possible to generate smaller SEs by increasing the number of bootstrap samples

obtained. This may lead to false confidence in estimates generated despite potential issues with initial sample

𝑥, e.g. if it too small, or sampled in such a way that it is not representative of the underlying population

𝑋. Although such issues are inherent to any bootstrapping approaches, the usual method of generating

bootstrapped CIs detailed above uses more information (i.e. using the entire bootstrap distribution) to generate

these values than the parameter-based 𝑒𝑠𝑡𝑖𝑚𝑎𝑡𝑒 ± 1.96 × 𝑆𝐸 method (i.e. using approximate summary

statistics to represent the distribution). The usual method of bootstrap CI generation may therefore be

expected to highlight any variation or uncertainty present more readily than the parameter-based approach;

this would be represented as wider CIs.

36

**C Appendix: Simulation Code**

**C.1 Generating Data and Models**

The data generating model used was from Appendix 3 of Bowden et al8; the relevant section describing their

model is reproduced below:

*“…*

𝑈𝑖 =

𝐽Σ

𝑗=1

𝜙𝑗𝐺𝑖𝑗 + 𝜖𝑈 𝑖

(6)

𝑋𝑖 =

𝐽Σ

𝑗=1

𝛾𝑗𝐺𝑖𝑗 + 𝑈𝑖 + 𝜖𝑋 𝑖

(7)

𝑌𝑖 =

𝐽Σ

𝑗=1

𝛼𝑗𝐺𝑖𝑗 + 𝛽𝑋𝑖 + 𝑈𝑖 + 𝜖𝑌 𝑖

(8)

*for participants indexed by* 𝑖 = 1, ..., 𝑁 *, and genetic instruments indexed by* 𝑗 = 1, ..., 𝐽 *.*

*The error terms* 𝜖𝑈 𝑖

, 𝜖𝑋 𝑖

*and* 𝜖𝑌 𝑖

*were each drawn independently from standard normal distributions.*

*The genetic effects on the exposure j are drawn from a uniform distribution between 0.03 and 0.1.*

*Pleiotropic effects* 𝛼𝑗 *and* 𝜙𝑗 *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* 𝛼𝑗 *parameter was drawn from a*

*uniform distribution between −0.2 and 0.2.*

*2. In Scenario 2 (directional pleiotropy, InSIDE satisfied), the* 𝛼𝑗 *parameter was drawn from a*

*uniform distribution between 0 and 0.2.*

*3. In Scenario 3 (directional pleiotropy, InSIDE not satisfied), the* 𝜙𝑗 *parameter was drawn from*

*a uniform distribution between −0.2 and 0.2.*

*The causal effect of the exposure on the outcome was either* 𝛽𝑋 = 0 *(null causal effect) or*

𝛽𝑋 = 0.1 *(positive causal effect). A total of 10 000 simulated datasets were generated for*

*sample sizes of N = 10 000 and 20 [sic] participants. Only the summary data, that is genetic*

*associations with the exposure and with the outcome and their standard errors as estimated by*

*univariate regression on the genetic instruments in turn, were used by the analysis methods. In*

*the two-sample setting, data were generated on 2N participants, and genetic associations with the*

*exposure were estimated in the first N participants, and genetic associations with the outcome in*

*the second N participants.”*8

To reproduce this model, code was written in R to generate the relevant participant level data. First, a

function (get\_simulated\_MR\_data) was written which included parameters specified by Bowden et al, and

also to allow testing of data simulation:

This initial simulation function generated data in the following format:

37

*# 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:

38

These models generated estimates of the coefficient of gene-exposure association (coeff\_G\_X), coefficient

of gene-outcome association (coeff\_G\_Y), and the relevant standard errors of these estimates. The values

of parameters inputted were also returned to aid in further testing of data/model generation, i.e. actual

gene-exposure associations (gamma), pleiotropic effects of invalid instruments (alpha), additional pleiotropic

effects when InSIDE assumption not satified (phi), causal effect of exposure on outcome (beta) and the

proportion of invalid genetic instruments with pleiotropic effects on the outcome (prop\_invalid).

test\_extract\_model <- **get\_models**(test\_data\_sim)

**summary**(test\_extract\_model[[1]])

## dataset Instrument coeff\_G\_X coeff\_G\_X\_SE

## Min. :1 Min. : 1 Min. :0.03419 Min. :6.659e-17

## 1st Qu.:1 1st Qu.: 7 1st Qu.:0.05645 1st Qu.:6.807e-17

## Median :1 Median :13 Median :0.06823 Median :6.873e-17

## Mean :1 Mean :13 Mean :0.06791 Mean :6.889e-17

## 3rd Qu.:1 3rd Qu.:19 3rd Qu.:0.08594 3rd Qu.:6.935e-17

## Max. :1 Max. :25 Max. :0.09379 Max. :7.313e-17

## gamma F\_stat R2\_stat coeff\_G\_Y

## Min. :0.03419 Min. :6.224e+27 Min. :1 Min. :-0.109067

## 1st Qu.:0.05645 1st Qu.:6.224e+27 1st Qu.:1 1st Qu.: 0.004951

## Median :0.06823 Median :6.224e+27 Median :1 Median : 0.006555

## Mean :0.06791 Mean :6.224e+27 Mean :1 Mean : 0.013297

## 3rd Qu.:0.08594 3rd Qu.:6.224e+27 3rd Qu.:1 3rd Qu.: 0.008890

## Max. :0.09379 Max. :6.224e+27 Max. :1 Max. : 0.162629

## coeff\_G\_Y\_SE alpha phi beta

## Min. :0.001550 Min. :-0.115363 Min. :0 Min. :0.1

## 1st Qu.:0.001579 1st Qu.: 0.000000 1st Qu.:0 1st Qu.:0.1

## Median :0.001591 Median : 0.000000 Median :0 Median :0.1

## Mean :0.001598 Mean : 0.006717 Mean :0 Mean :0.1

## 3rd Qu.:0.001624 3rd Qu.: 0.000000 3rd Qu.:0 3rd Qu.:0.1

## Max. :0.001653 Max. : 0.156224 Max. :0 Max. :0.1

## prop\_invalid n\_instruments n\_participants

## Min. :0.3 Min. :25 Min. :1000

## 1st Qu.:0.3 1st Qu.:25 1st Qu.:1000

## Median :0.3 Median :25 Median :1000

## Mean :0.3 Mean :25 Mean :1000

## 3rd Qu.:0.3 3rd Qu.:25 3rd Qu.:1000

## Max. :0.3 Max. :25 Max. :1000

39

**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 ≠ 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 ≠ 0, with values spread uniformly between

alpha\_min and alpha\_max

40

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:

41

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

For the next phase of testing, a function (plot\_GY\_GX) was written to plot the coefficients for gene-exposure

versus gene-outcome as estimated using the previously created linear models:

plot\_GY\_GX <- **function**(model\_tib,

plot\_title = **as.character**(NA),

x\_min = 0, *# set x-axis limits*

x\_max = 0.1,

y\_min = **-**0.05, *# set x-axis limits*

y\_max = 0.06,

beta\_x = 0.075, *# set beta-hat position*

beta\_y = 0.05,

hat\_offset = 0.003

){

model\_tib **%>%**

**mutate**(Gradient = **round**(**coefficients**(**lm**(coeff\_G\_Y **~** 0 **+** coeff\_G\_X)[1], 5),

digits = 2)) **%>%**

**plot\_template**() **+** *# pre-formatted plot template - call to ggplot with UoE colours*

**aes**(x = coeff\_G\_X, y = coeff\_G\_Y) **+**

**geom\_point**(colour = edin\_bright\_red\_hex, alpha = 0.3) **+**

**geom\_abline**(**aes**(intercept = 0,

slope = Gradient),

size = 1,

colour = edin\_uni\_blue\_hex) **+**

**geom\_text**(**aes**(label = **paste0**("\U03B2 = ", **as.character**(Gradient))), *#beta*

x = beta\_x, *# labels with gradient (causal effect estimate)*

y = beta\_y,

colour = edin\_uni\_blue\_hex,

hjust = 0,

data = . **%>% slice\_head**()*# prevent over-printing*

) **+**

*#label = expression("True" ~ hat(beta)~ "= 0.25"),*

**annotate**("text",

x = beta\_x, *# add hat to beta*

y = beta\_y **+** hat\_offset,

label = **paste**("\U02C6"),

colour = edin\_uni\_blue\_hex,

hjust = **-**0.4,

vjust = 0.9

) **+**

**labs**(title = plot\_title,

x = "Gene-Exposure Coefficient",

y = "Gene-Outcome Coefficient") **+**

**xlim**(x\_min, x\_max) **+**

**ylim**(y\_min, y\_max)

}

42

With random error terms set to 0 (rand\_error = FALSE) and no causal effect present, a graph of geneexposure

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:

43

**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:

44

**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:

45

**C.2.5 Invalid Instruments**

Where invalid instruments are present (i.e. prop\_invalid ≠ 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:

46

**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:

47

**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 geneexposure

relationship) may not allow accurate estimation of the true causal effect, as pleiotropic effects may

scale with instrument strength. If pleiotropic effects are balanced, InSIDE assumption violation may lead

to greater imprecision in causal effect estimation; if pleiotropic effects are directional, InSIDE assumption

violation may lead to bias.

Bowden et al8 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:

48

Setting InSIDE\_satisfied = TRUE should mean phi = 0; InSIDE\_satisfied=FALSE should result in phi

∝ gene-outcome coefficient, with scatter only in the positive direction of gene-outcome coefficients given the

model also requires directional pleiotropy before phi is used:

49

**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 al8:

*# Load WME functions*

**library**(TwoSampleMR)

*# Load RStan - needed for MR-Hevo*

**library**(rstan)

*# Run local copy of MR-Hevo functions*

*# Not using full package due to conflicts with Windows*

**source**(**here**("Script", "Hevo", "functions.mrhevo.R"))

*# Standard set-up for RStan*

**options**(mc.cores = parallel**::detectCores**())

**rstan\_options**(auto\_write = TRUE, save\_dso = TRUE)

*# Compile model for MR-Hevo*

mr.stanmodel <- **stan\_model**(file= **here**("Script",

"Hevo",

"MRHevo\_summarystats.stan"),

model\_name="MRHevo.summarystats",

verbose=FALSE,

save\_dso = TRUE,

auto\_write = TRUE)

get\_summary\_MR\_tib\_row <- **function**(model\_list){

*# Create output tibble in same format as Table 2/3 from*

*# Bowden et al*

output\_tib\_row <- **tibble**(N = **as.integer**(),

Prop\_Invalid = **as.double**(),

F\_stat = **as.double**(),

R2\_stat = **as.double**(),

WME\_Av = **as.double**(),

WME\_SE = **as.double**(),

WME\_Pos\_Rate = **as.double**(),

Hevo\_Av = **as.double**(),

Hevo\_SE = **as.double**(),

Hevo\_Pos\_Rate = **as.double**())

n\_datasets <- **length**(model\_list)

*# Create blank tibble to receive results of Weighted*

*# Median Estimator function from MR-Base*

results\_tib <- **tibble**(WME\_est = **as.double**(),

WME\_se = **as.double**(),

50

WME\_pval = **as.double**(),

WME\_nsnp = **as.integer**(),

Hevo\_est = **as.double**(),

Hevo\_se = **as.double**(),

Hevo\_sd = **as.double**(),

Hevo\_est\_lower\_CI = **as.double**(),

Hevo\_est\_upper\_CI = **as.double**(),

Hevo\_causal\_detected = **as.logical**()

)

*# Run WME and MR-Hevo for each dataset*

**for**(dataset **in** 1**:**n\_datasets){

*# Stored as individual vectors for MR-Hevo/RStan - not*

*# Tidyverse compatible*

coeff\_G\_X\_vect <- model\_list[[dataset]]**$**coeff\_G\_X

coeff\_G\_Y\_vect <- model\_list[[dataset]]**$**coeff\_G\_Y

coeff\_G\_X\_SE\_vect <- model\_list[[dataset]]**$**coeff\_G\_X\_SE

coeff\_G\_Y\_SE\_vect <- model\_list[[dataset]]**$**coeff\_G\_Y\_SE

prop\_invalid <- **min**(model\_list[[dataset]]**$**prop\_invalid)

F\_stat <- **min**(model\_list[[dataset]]**$**F\_stat)

R2\_stat <- **min**(model\_list[[dataset]]**$**R2\_stat)

n\_instruments <- **max**(model\_list[[dataset]]**$**Instrument)

n\_participants <- **min**(model\_list[[dataset]]**$**n\_participants)

*# N.B. MR-Hevo terminology vs WME paper/other code:*

*# alpha = effects of instruments on exposure, i.e. coeff\_G\_X*

*# beta = pleiotropic effects of instruments on outcome, i.e. alpha in WME*

*# gamma = effects of instruments on outcome, i.e. coeff\_G\_Y*

*# theta = causal effect X on Y, i.e. b*

*# Results from weighted median estimator method*

WME\_results <- **mr\_weighted\_median**(b\_exp = coeff\_G\_X\_vect,

b\_out = coeff\_G\_Y\_vect,

se\_exp = coeff\_G\_X\_SE\_vect,

se\_out = coeff\_G\_Y\_SE\_vect,

parameters = **list**(nboot = 1000))

*# Results from MR-Hevo method*

Hevo\_results<- **run\_mrhevo.sstats**(alpha\_hat = coeff\_G\_X\_vect,

se.alpha\_hat = coeff\_G\_X\_SE\_vect,

gamma\_hat = coeff\_G\_Y\_vect,

se.gamma\_hat = coeff\_G\_Y\_SE\_vect) **%>%**

**summary**()

*# Extract WME Results*

results\_tib[dataset, ]**$**WME\_est <- WME\_results**$**b

results\_tib[dataset, ]**$**WME\_se <- WME\_results**$**se

results\_tib[dataset, ]**$**WME\_pval <- WME\_results**$**pval

results\_tib[dataset, ]**$**WME\_nsnp <- WME\_results**$**nsnp

51

*# Extract MR-Hevo Results*

results\_tib[dataset, ]**$**Hevo\_est <- Hevo\_results**$**summary["theta","mean"]

results\_tib[dataset, ]**$**Hevo\_se <- Hevo\_results**$**summary["theta","se\_mean"]

results\_tib[dataset, ]**$**Hevo\_sd <- Hevo\_results**$**summary["theta","sd"]

results\_tib[dataset, ]**$**Hevo\_est\_lower\_CI <- Hevo\_results**$**summary["theta","2.5%"]

results\_tib[dataset, ]**$**Hevo\_est\_upper\_CI <- Hevo\_results**$**summary["theta","97.5%"]

}

*# Add causality Boolean to MR-Hevo*

results\_tib <- results\_tib **%>%**

**mutate**(Hevo\_est\_causal\_detected = (Hevo\_est\_lower\_CI **>** 0 **|** Hevo\_est\_upper\_CI **<** 0))

output\_tib\_row <- results\_tib **%>%**

**summarise**(N = n\_participants,

Prop\_Invalid = prop\_invalid,

F\_stat = **mean**(F\_stat),

R2\_stat = **mean**(R2\_stat),

WME\_Av = **mean**(WME\_est),

WME\_SE = **mean**(WME\_se),

WME\_Pos\_Rate = **length**(WME\_pval[WME\_pval **<** 0.05]) **/** n\_datasets,

Hevo\_Av = **mean**(Hevo\_est),

Hevo\_SE = **mean**(Hevo\_se),

Hevo\_Lower\_CI = **mean**(Hevo\_est\_lower\_CI),

Hevo\_Upper\_CI = **mean**(Hevo\_est\_upper\_CI),

Hevo\_Pos\_Rate = **sum**(Hevo\_est\_causal\_detected) **/** n\_datasets

) **%>%**

**mutate**(**across**(**where**(is.double), round, 3))

**return**(output\_tib\_row)

}

test\_tib\_summ\_MR\_data <- **get\_simulated\_MR\_data**(n\_participants = 10000,

n\_instruments = 25,

n\_datasets = 2,

prop\_invalid = 0.1,

beta\_val = 0.1,

causal\_effect = TRUE,

rand\_error = TRUE,

two\_sample = TRUE,

balanced\_pleio = TRUE,

InSIDE\_satisfied = TRUE)

test\_tib\_summ\_MR\_models <- **get\_models**(test\_tib\_summ\_MR\_data)

test\_tib\_summ\_MR\_row <- **get\_summary\_MR\_tib\_row**(test\_tib\_summ\_MR\_models)

##

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

##

## COMPILING MODEL 'MRHevo.summarystats' NOW.

52

N Prop\_Invalid F\_stat R2\_stat WME\_Av WME\_SE WME\_Pos\_Rate Hevo\_Av Hevo\_SE Hevo\_Lower\_CI Hevo\_Upper\_CI Hevo\_Pos\_Rate

10000 0.1 14.739 0.036 0.091 0.09 0.5 0.121 0.001 -0.052 0.302 0

##

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

##

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

##

## COMPILING MODEL 'MRHevo.summarystats' NOW.

##

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

test\_tib\_summ\_MR\_row **%>%**

**kable**() **%>%**

**kable\_styling**(latex\_options="scale\_down")

53

**D Appendix: R Packages Used**

**D.1 Package Citations**

This work was completed using R version 4.4.321 with the following R packages: acronymsdown v. 0.11.145,

bookdown v. 0.4346,47, car v. 3.1.348, cowplot v. 1.1.349, crayon v. 1.5.350, devtools v. 2.4.551, flextable v.

0.9.952, ftExtra v. 0.6.453, ggdag v. 0.2.1354, gghighlight v. 0.4.155, gluedown v. 1.0.956, grateful v. 0.2.1257,

grid v. 4.4.358, here v. 1.0.159, infer v. 1.0.860, kableExtra v. 1.4.061, knitr v. 1.5062–64, matrixStats v.

1.5.065, medicaldata v. 0.2.044, officer v. 0.6.1066, parallel v. 4.4.367, rmarkdown v. 2.2968–70, rstan

v. 2.32.771, tables v. 0.9.3172, tidyverse v. 2.0.023, TwoSampleMR v. 0.6.1673,74, wordcountaddin v.

0.3.0.900075.

**D.2 Session Information**

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

RAM: 16.9 GB

## setting value

## version R version 4.4.3 (2025-02-28 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 2025-06-08

## rstudio 2025.05.0+496 Mariposa Orchid (desktop)

## pandoc 3.4 @ C:/Program Files/RStudio/resources/app/bin/quarto/bin/tools/ (via rmarkdown)

## quarto ERROR: Unknown command "TMPDIR=C:/Users/timol/AppData/Local/Temp/RtmpoL0HFL/file4d9c24bb4b5d". ## # A tibble: 33 x 5

## package ondiskversion loadedversion date source

## <chr> <chr> <chr> <chr> <chr>

## 1 acronymsdo~ 0.11.1 0.11.1 2025~ Githu~

## 2 bookdown 0.43 0.43 2025~ CRAN ~

## 3 cowplot 1.1.3 1.1.3 2024~ CRAN ~

## 4 dplyr 1.1.4 1.1.4 2023~ CRAN ~

## 5 flextable 0.9.9 0.9.9 2025~ CRAN ~

## 6 forcats 1.0.0 1.0.0 2023~ CRAN ~

## 7 ftExtra 0.6.4 0.6.4 2024~ CRAN ~

## 8 ggdag 0.2.13 0.2.13 2024~ CRAN ~

## 9 gghighlight 0.4.1 0.4.1 2023~ CRAN ~

## 10 ggplot2 3.5.2 3.5.2 2025~ CRAN ~

## 11 gluedown 1.0.9 1.0.9 2024~ CRAN ~

## 12 grateful 0.2.12 0.2.12 2025~ CRAN ~

## 13 here 1.0.1 1.0.1 2020~ CRAN ~

## 14 infer 1.0.8 1.0.8 2025~ CRAN ~

## 15 kableExtra 1.4.0 1.4.0 2024~ CRAN ~

## 16 knitr 1.50 1.50 2025~ CRAN ~

## 17 koRpus 0.13.8 0.13-8 2021~ CRAN ~

## 18 koRpus.lan~ 0.1.4 0.1-4 2020~ CRAN ~

54

## 19 lubridate 1.9.4 1.9.4 2024~ CRAN ~

## 20 magrittr 2.0.3 2.0.3 2022~ CRAN ~

## 21 medicaldata 0.2.0 0.2.0 2021~ CRAN ~

## 22 officer 0.6.10 0.6.10 2025~ CRAN ~

## 23 purrr 1.0.4 1.0.4 2025~ CRAN ~

## 24 readr 2.1.5 2.1.5 2024~ CRAN ~

## 25 rstan 2.32.7 2.32.7 2025~ CRAN ~

## 26 StanHeaders 2.32.10 2.32.10 2024~ CRAN ~

## 27 stringr 1.5.1 1.5.1 2023~ CRAN ~

## 28 sylly 0.1.6 0.1-6 2020~ CRAN ~

## 29 tables 0.9.31 0.9.31 2024~ CRAN ~

## 30 tibble 3.2.1 3.2.1 2023~ CRAN ~

## 31 tidyr 1.3.1 1.3.1 2024~ CRAN ~

## 32 tidyverse 2.0.0 2.0.0 2023~ CRAN ~

## 33 TwoSampleMR 0.6.16 0.6.16 2025~ https~

55