



METHOD ARTICLE

REVISED **Guidelines for performing Mendelian randomization investigations [version 2; peer review: 2 approved]**

Stephen Burgess ^{1,2}, George Davey Smith ^{3,4}, Neil M. Davies ^{3,4},
Frank Dudbridge⁵, Dipender Gill ⁶, M. Maria Glymour⁷, Fernando P. Hartwig ^{3,8},
Michael V. Holmes ^{9,10}, Cosetta Minelli ¹¹, Caroline L. Relton^{3,4},
Evropi Theodoratou ^{12,13}





¹MRC Biostatistics Unit, University of Cambridge, Cambridge, UK²Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK³MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK⁴Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK⁵Department of Health Sciences, University of Leicester, Leicester, UK⁶Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK⁷Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA⁸Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil⁹MRC Population Health Research Unit at the University of Oxford, Nuffield Department of Population Health, University of Oxford, Oxford, UK¹⁰Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK¹¹National Heart and Lung Institute, Imperial College London, London, UK¹²Centre for Global Health, Usher Institute, University of Edinburgh, Edinburgh, UK¹³Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK**v2** **First published:** 26 Nov 2019, 4:186
<https://doi.org/10.12688/wellcomeopenres.15555.1>**Latest published:** 28 Apr 2020, 4:186
<https://doi.org/10.12688/wellcomeopenres.15555.2>**Abstract**

This paper provides guidelines for performing Mendelian randomization investigations. It is aimed at practitioners seeking to undertake analyses and write up their findings, and at journal editors and reviewers seeking to assess Mendelian randomization manuscripts. The guidelines are divided into nine sections: motivation and scope, data sources, choice of genetic variants, variant harmonization, primary analysis, supplementary and sensitivity analyses (one section on robust statistical methods and one on other approaches), data presentation, and interpretation. These guidelines will be updated based on feedback from the community and advances in the field. Updates will be made periodically as needed, and at least every 18 months.

Keywords

Mendelian randomization, guidelines, genetic epidemiology, causal inference

Open Peer Review**Reviewer Status**  

	Invited Reviewers	
	1	2
version 2 (revision) 28 Apr 2020		 report
version 1 26 Nov 2019	 report	  report

1 **Jean Morrison** , University of Chicago, Chicago, USA

2 **Sonja A. Swanson**, Erasmus University Medical Center, Rotterdam, The Netherlands

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Stephen Burgess (sb452@medschl.cam.ac.uk)

Author roles: **Burgess S:** Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Davey Smith G:** Conceptualization, Writing – Review & Editing; **Davies NM:** Conceptualization, Writing – Review & Editing; **Dudbridge F:** Conceptualization, Writing – Review & Editing; **Gill D:** Conceptualization, Writing – Review & Editing; **Glymour MM:** Conceptualization, Writing – Review & Editing; **Hartwig FP:** Conceptualization, Writing – Review & Editing; **Holmes MV:** Conceptualization, Writing – Review & Editing; **Minelli C:** Conceptualization, Writing – Review & Editing; **Relton CL:** Conceptualization, Writing – Review & Editing; **Theodoratou E:** Conceptualization, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Stephen Burgess is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (204623). Dipender Gill is funded by the Wellcome Trust. Evropi Theodoratou is supported by a CRUK Career Development Fellowship (C31250/A22804). George Davey Smith, Neil M Davies, and Caroline L Relton work within a Unit funded by the Medical Research Council (MC_UU_12013/1, MC_UU_12013/9, MC_UU_00011/1, MC_UU_00011/5). Neil Davies is supported by a Future Research Leaders grant funded by the Economics and Social Research Council (ES/N000757/1). Michael Holmes works in a unit that receives funding from the UK Medical Research Council and is supported by a British Heart Foundation Intermediate Clinical Research Fellowship (FS/18/23/33512) and the National Institute for Health Research Oxford Biomedical Research Centre.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Burgess S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Burgess S, Davey Smith G, Davies NM *et al.* **Guidelines for performing Mendelian randomization investigations [version 2; peer review: 2 approved]** Wellcome Open Research 2020, 4:186 <https://doi.org/10.12688/wellcomeopenres.15555.2>

First published: 26 Nov 2019, 4:186 <https://doi.org/10.12688/wellcomeopenres.15555.1>

REVISED Amendments from Version 1

We have updated the manuscript to address the reviewers' concerns. Updates include relating to reverse causation, pleiotropy and instrument invalidity, time, and the interpretation of estimates.

Any further responses from the reviewers can be found at the end of the article

The aim of this paper is to provide guidelines for performing Mendelian randomization investigations. It is written both for practitioners seeking to undertake analyses and write up their findings, and for journal editors and reviewers seeking to assess Mendelian randomization manuscripts. These guidelines are deliberately written as suggestions and recommendations rather than as prescriptive rules, as we believe that there is no recipe or single “right way” to perform a Mendelian randomization investigation. Best practice will depend on the aim of the investigation and the specific exposure and outcome variables. However, we believe these guidelines will help investigators to consider the key issues in designing, undertaking and presenting Mendelian randomization analyses. These guidelines will be updated based on feedback from the community and advances in the field. Updates will be made periodically as needed, and at least every 18 months.

These guidelines are complementary to the STROBE-MR recommendations on reporting Mendelian randomization investigations¹. Here, we provide advice on which analyses to perform in a Mendelian randomization investigation, whereas the STROBE-MR guidelines focus on reporting the analyses chosen by the investigators. We assume a familiarity with the basic concepts of Mendelian randomization and genetic epidemiology, such as pleiotropy and linkage disequilibrium²⁻⁴. We use the term “exposure” to refer to the proposed causal factor, and “outcome” to refer to the factor or condition that the exposure is hypothesized to influence.

Flowcharts highlighting some of the key analytic steps and choices for investigators are provided as [Figure 1](#) and [Figure 2](#), and a one-page checklist summarizing these guidelines written for reviewers of Mendelian randomization analyses is provided as [Figure 3](#). The guidelines are divided into nine sections: motivation and scope, data sources, choice of genetic variants, variant harmonization, primary analysis, supplementary and sensitivity analyses (one section on robust statistical methods and one on other approaches), data presentation, and interpretation. Software to implement the statistical methods is referenced in [Table 1](#).

1. Motivation and scope

Mendelian randomization uses genetic variants to assess causal relationships in observational data. A genetic variant can be considered as an instrumental variable for a given exposure if it satisfies the instrumental variable assumptions: it is associated with the exposure in a specific way, meaning that it does not

affect the outcome except via the exposure, and it is not associated with the outcome due to confounding^{5,6}. Before embarking on a Mendelian randomization analysis, investigators should consider the aims of their investigation and the primary hypothesis of interest. There are many potential motivations for using Mendelian randomization, and the motivation should influence decisions on how to perform the analysis, and how to arrange and present its results. The objective of a Mendelian randomization analysis is a test of a causal hypothesis, and often additionally an estimate of a causal effect⁷. The straightforward statement of the causal hypothesis is that interventions in the exposure variable will affect the outcome. If the genetic associations with the exposure vary with time, then there are some nuances in terms of what causal hypotheses can be tested⁸; we discuss the impact of time-varying relationships between variables in Section 9.

If a Mendelian randomization investigation is performed primarily to assess whether an exposure has a causal effect on an outcome, then estimating the size of the causal effect of the exposure on the outcome is less important and may even be unnecessary^{7,9}. For example, Mendelian randomization has demonstrated a causal effect of time spent in formative education on Alzheimer's disease¹⁰. However, those at risk of Alzheimer's disease are unable to extend their time in formative education. The analysis tests a meaningful causal hypothesis, but the size of the causal estimate has limited utility. Priorities in such an analysis are to find genetic variants that satisfy the instrumental variable assumptions and to test their associations with the outcome in the largest available dataset that is relevant to the causal question of interest.

In contrast, if investigators seek to estimate the quantitative impact on the outcome of a proposed intervention in the exposure¹¹, then further questions become more important, such as how well the genetic variant proxies the specific intervention, whether genetic associations with the exposure are estimated in a relevant population, and whether the relationships between variables are linear and homogeneous in the population¹². However, as we discuss in Section 9, causal estimates from Mendelian randomization should always be interpreted with caution. Alternatively, if investigators simply want to assess whether traits share common genetic predictors (potentially implying shared aetiological mechanisms), then an analytic approach that assesses shared heritability (such as LD-score regression¹³ or the latent causal variable method¹⁴) may be preferable to conducting a Mendelian randomization investigation.

Investigators should also give thought to the scope of their analysis. If the aim of the investigation is to understand disease aetiology, then consideration of a limited set of exposures/outcomes as main analyses may be justified, whereas if the question relates to public health, then consideration of a broad range of outcomes influenced by an exposure may be worthwhile. At the extreme end of the spectrum is a phenome-wide Mendelian randomization investigation, in which very large numbers of exposure/outcome pairs are considered¹⁵⁻¹⁷.

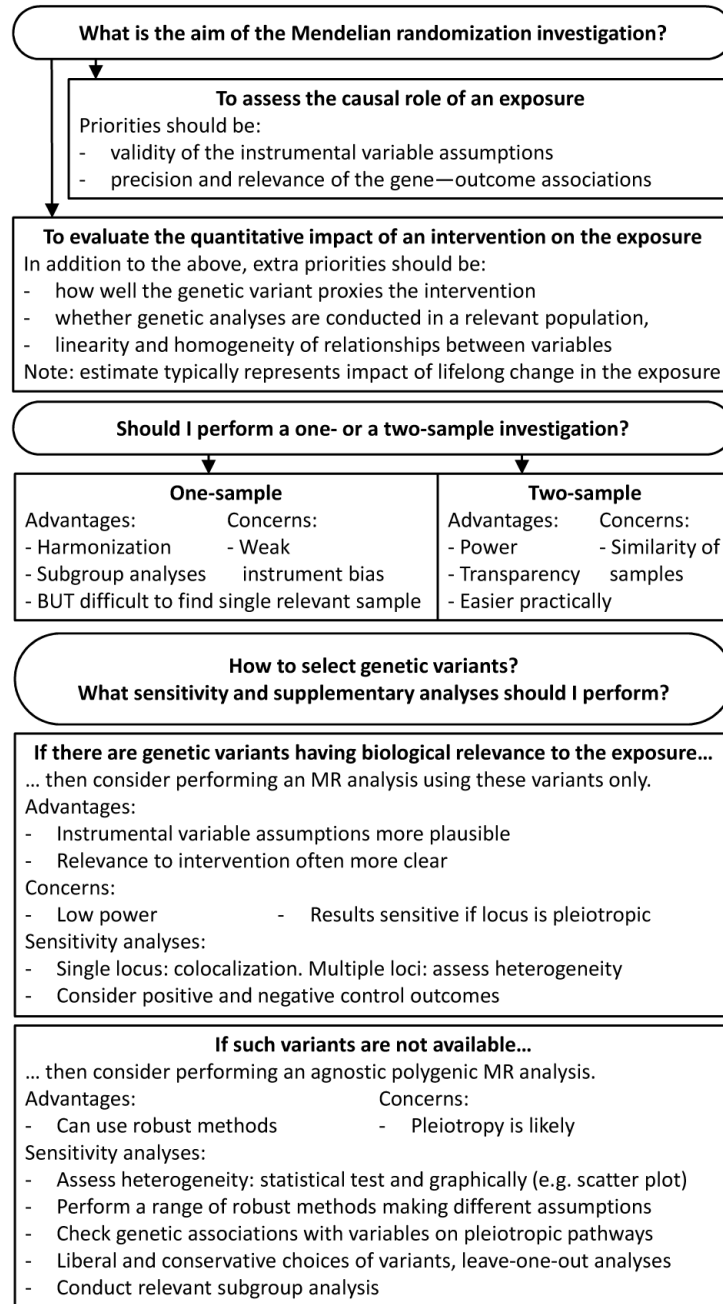


Figure 1. Flowchart highlighting some of the key analytic choices in performing a Mendelian randomization (MR) analysis.

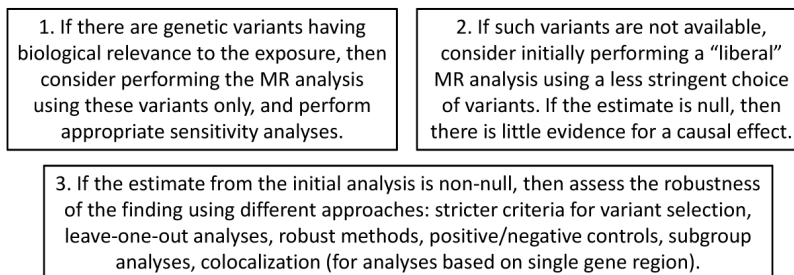


Figure 2. Generic analytic pipeline for Mendelian randomization (MR).

Checklist for reviewing Mendelian randomization investigations

1. What is the primary hypothesis of interest? What is the motivation for using Mendelian randomization? What is the scope of the investigation? What and how many primary analyses?
-

2. Data sources

What type of Mendelian randomization investigation is this? One-sample or two-sample? Sample overlap? Summarized data or individual-level data? Drawn from same population? Relevance to applied research?

3. Selection of genetic variants – how were the genetic variants chosen? Single or multiple gene regions?

- a. Biological rationale?
 - b. GWAS analysis? If so, what dataset? What was the p-value threshold? Clumping?
 - c. Were genetic variants excluded from the analysis? Associations with pleiotropic pathways?
 - d. How else was the validity of genetic variants as instrumental variables assessed?
-

4. Variant harmonization (for two-sample analyses)

Was it checked that genetic variants were appropriately orientated across the datasets?

5. Primary analysis

What was the primary analysis? What was the statistical method? How implemented? Multiple testing?

- 6 and 7. Supplementary and sensitivity analyses

What analyses were performed to support and assess the validity of the primary analysis?
For example: stricter criteria for variant selection, assess heterogeneity, robust methods, subgroup analysis, positive/negative controls, 'leave-one-out' analyses, colocalization (single gene region)

8. Data presentation

How are the data and results presented to allow readers to assess the analysis and assumptions?
For example: scatter plot, forest/funnel/radial plot, R^2 /F statistics, comparison of methods, power

9. Interpretation

How have results been interpreted, particularly any numerical estimates?

Figure 3. Checklist of questions to consider when reviewing a Mendelian randomization investigation.

Table 1. Summary of some methods proposed for Mendelian randomization: inverse-variance weighted method and robust methods.

Method	Consistency assumption	Strengths and weaknesses	Reference	Software
Inverse-variance weighted	All variants valid or balanced pleiotropy	Most efficient (greatest statistical power), biased if average pleiotropic effect differs from zero	18	*†
MR-Egger	InSIDE	Sensitive to outliers, sensitive to violations of InSIDE assumption, InSIDE assumption often not plausible, often less efficient	19	*†
Weighted median	Majority valid	Robust to outliers, sensitive to addition/removal of genetic variants	20	*†
Mode-based estimation	Plurality valid	Robust to outliers, sensitive to bandwidth parameter and addition/removal of genetic variants, generally conservative	21	*†
MR-PRESSO	Outlier-robust	Removes outliers, efficient with valid IVs, very high false positive rate with several invalid IVs	22	‡
MR-Robust	Outlier-robust	Downweights outliers, efficient with valid IVs, high false positive rate with several invalid IVs	23	*
MR-Lasso	Outlier-robust	Removes outliers, efficient with valid IVs, high false positive rate with several invalid IVs	23	
MR-RAPS	Balanced pleiotropy (except outliers)	Downweights outliers, sensitive to violations of balanced pleiotropy assumption	24	‡
Contamination Mixture	Plurality valid	Robust to outliers, sensitive to variance parameter and addition/removal of genetic variants	25	*
MR-Mix	Plurality valid	Robust to outliers, requires large numbers of genetic variants, very high false positive rate in several scenarios	26	‡

Each of the methods in the table can be implemented using summarized data. False positive rates refer to the simulation study by Slob and Burgess²⁷. InSIDE is the Instrument Strength Independent of Direct Effect assumption.

* Implemented in MendelianRandomization package for R (<https://cran.r-project.org/web/packages/MendelianRandomization/index.html>)

† Implemented in mrrobust package for Stata (<https://github.com/remlapmot/mrrobust>)

‡ Implemented for R in its own software package:

- MR-PRESSO in mrpresso package (<https://github.com/rondolab/MR-PRESSO>),

- MR-RAPS in mr.raps package (<https://github.com/qingyuanzhao/mr.raps>),

- MR-Mix in MRMix package (<https://github.com/gqi/MRMix>).

Such analyses are generally regarded as exploratory or “hypothesis-generating”, and results are typically treated as provisional until replicated in an independent dataset.

Specifying the primary analyses in a Mendelian randomization investigation is important to address problems of multiple testing, particularly given the large number of analyses that could be performed using available genetic data²⁸. Additional analyses, including subgroup analyses and analyses on related outcomes may be presented as supplementary, exploratory, or sensitivity analyses. An overly conservative approach to multiple testing is often excessive, given the typically low power of Mendelian randomization studies and the fact that Mendelian randomization typically investigates exposure/outcome relationships with prior epidemiological or biological support. As with all epidemiological analyses, it is good practice to avoid selective reporting of “significant” results (leading to reporting bias) and to describe transparently all analyses performed.

2. Data sources

The next fundamental question is which data sources will be used: how many datasets are included in the analysis and

whether the analysis is performed using individual-level data or summarized data.

Mendelian randomization investigations can be performed using data from a single sample (known as one-sample Mendelian randomization), in which genetic variants, exposure, and outcome are measured in the same individuals, or from two samples (known as two-sample Mendelian randomization), in which variant—exposure associations are estimated in one dataset, and variant—outcome associations are estimated in a second dataset²⁹. Two-sample investigations often occur when genetic associations with the exposure are estimated in a cross-sectional sample of healthy individuals, to reflect genetic associations with usual levels of the exposure in the population, and genetic associations with a binary disease outcome are estimated in a case-control study.

There are benefits and limitations of both one- and two-sample settings. A one-sample setting allows the investigation to be conducted in a single population sample, meaning that Mendelian randomization and conventional epidemiological findings (for example from multivariable-adjusted regression) can be compared in the same individuals. In a two-sample setting,

the populations from which the two samples were extracted may differ. This is problematic if associations of the genetic variant with the exposure or with variables on pleiotropic pathways differ between the two samples, as this could affect the validity of the instrumental variable assumptions. A particular concern arises if the two samples represent different ethnic groups, as patterns of linkage disequilibrium can differ between populations, meaning that a genetic variant may not be as strongly (or even not at all) associated with the exposure in the outcome dataset. Alternatively, the two samples could differ substantially according to population characteristics such as age, sex, socio-economic background, and so on^{8,30}. Such differences can affect not only the interpretation of causal estimates, but also the validity of causal inferences³¹. For example, genetic variants associated with smoking intensity may be strongly associated with disease outcomes in populations where smoking is common, but not in populations where smoking is rare. One-sample analyses do not suffer from these concerns, nor do they require harmonization of the genetic variants across the datasets (see Section 4).

Another related issue is whether the analysis is performed using individual-level data or summarized data. Summarized data are genetic association estimates from regression of the exposure or outcome on a genetic variant^{18,32}. Several large consortia have made such estimates publicly available for hundreds of thousands of variants^{28,33}. Although the use of summarized data is often synonymous with the two-sample setting, the benefits and limitations for the analysis of the two choices (i.e. one- vs two-sample and individual-level vs summarized data) are distinct. Summarized data are often available for larger sample sizes, meaning that power to detect a causal effect is increased. However, access to only summarized data limits the range of analyses that can be performed. Individual-level data are required to conduct analyses in specific subgroups or strata of the population, or to choose which variables to adjust for when generating the summarized data. If published summarized association estimates have already been adjusted for a variable causally downstream of the exposure or outcome, collider bias (see Section 7) may be unavoidable. Individual data in a one-sample setting are required to investigate non-linear effects^{34,35}. An advantage of publicly available summarized data is transparency, as the analysis can be reproduced by a third party with access to the same data.

One- and two-sample investigations also differ in terms of bias with weak instruments³⁶. In a one-sample setting, if the genetic variant–exposure associations are weak, then chance variation means that genetic associations with the exposure and outcome are correlated in the direction of the confounded association between the two. This results in instrumental variable estimates that are biased in the direction of the confounded association, and inflated false positive (type I error) rates, particularly when more than one variant is included in the analysis³⁷. In a two-sample setting without sample overlap, bias due to weak instruments is in the direction of the null, and does not lead to false positive findings. However, as several large consortia have overlapping studies, participants may overlap between the datasets used to estimate the genetic associations with the exposure and outcome³⁸.

In this case, the direction and size of the bias varies linearly depending on the degree of overlap (formally, depending on the degree of correlation between the genetic association estimates). For the special case of a one-sample analysis with a binary disease outcome, if the genetic associations with the exposure are estimated in the controls only, then genetic associations with the exposure and outcome will not be correlated, and bias will follow the pattern of the two-sample setting³⁸.

3. Selection of genetic variants

The most important decision to be made in designing a Mendelian randomization investigation is which genetic variants to include in the analysis³⁹. First, it is necessary to decide whether the analysis is performed using variants from a single gene region, or using variants from multiple regions of the genome (a polygenic analysis). For example, a Mendelian randomization analysis for C-reactive protein may be conducted using variants in the neighbourhood of the *CRP* gene region (which encodes C-reactive protein), or it may be conducted using all independent genome-wide significant predictors of C-reactive protein⁴⁰. The former has advantages of specificity – if a gene region has a specific biological link with the exposure, then the Mendelian randomization investigation is more plausible as an assessment of the causal role of that particular exposure. However, if only one gene region is included in the analysis, then several robust statistical analysis methods (see Section 6) are not possible, as they assume independence in whether variants violate the instrumental variable assumptions. Variants in the same gene region are likely to either all be valid instruments or all invalid. Additionally, when genetic variants are all valid instruments, power is maximized when genetic variants explain the greatest proportion of variance in the exposure⁴¹ – hence a polygenic Mendelian randomization investigation will typically have greater power than one including variants only from a single gene region.

When the analysis is based on a single gene region, it may be that a single variant is included in the analysis. However, if there are multiple variants that explain independent variance in the exposure, then their inclusion will increase power to detect a causal effect, even if the variants are partially correlated. With summarized data, appropriate methods should be used to account for correlated variants³⁰. If variants in a gene region can be thought of as proxies for an intervention that targets the exposure (such as variants in the *HMGCR* gene region for statin drugs), then the analysis has particular relevance for predicting the effect of that intervention.

For a polygenic analysis, there are two main strategies for selecting variants: either a biologically driven approach or a statistically driven approach. The two approaches are not mutually exclusive, and the overall decision of which variants to include may comprise elements from both approaches.

A biological approach to the selection of genetic variants would be to include variants from regions that have a biological link to the exposure of interest. For example, several Mendelian randomization investigations for vitamin D have used variants from four gene regions that are biologically implicated in the synthesis

or metabolism of vitamin D⁴². However, caution is required as biological understanding is rarely infallible. As an example, although genetic variants in the *IL6R* gene region are associated with increased circulating levels of interleukin-6, they in fact decrease interleukin-6 signalling, leading to opposite directions of association with disease outcomes to those expected based on serum interleukin-6 measurements⁴³.

A common statistical approach when selecting genetic variants is to include all variants that are associated with the exposure of interest at a given level of statistical significance (typically, a genome-wide significance threshold, such as $p < 5 \times 10^{-8}$). Often, selection is based on the dataset in which genetic associations with the exposure are estimated. However, this leads to “winner’s curse” –genetic associations tend to be overestimated in the dataset in which they were first discovered. If genetic variants are selected based on their associations with the exposure in the dataset under analysis, weak instrument bias is exacerbated (in the direction of the observational association in a one-sample setting, and in the direction of the null in a two-sample setting)³⁷. Bias can be avoided by selecting genetic variants based on a different dataset entirely. This can lead to a “three-sample” analysis, in which variants are identified in one dataset, and the genetic associations with the exposure

and outcome are estimated in separate datasets⁴⁴. If genetic variants are chosen solely based on their association with the exposure without reference to the function of the variants, then researchers should be especially careful about the possibility of variants being pleiotropic.

A more nuanced approach to variant selection would be to start off with a statistical rationale for choosing genetic variants, but then to exclude variants that are known to be pleiotropic or that are associated with variables that represent pleiotropic pathways to the outcome. However, a genetic association with a variable does not necessarily reflect that the instrumental variable assumptions are violated.

We use the term “horizontal pleiotropy” (sometimes referred to as simply “pleiotropy”) to refer to the scenario where a genetic variant is associated with variables on different causal pathways to the outcome, and “vertical pleiotropy” (sometimes referred to as “indirect pleiotropy” or “mediated pleiotropy”) to refer to the scenario where a genetic variant is associated with variables that are on the same causal pathway to the outcome⁴⁵. Provided that the causal pathway from the genetic variant to the outcome is mediated entirely via the exposure (see [Figure 4](#)), a genetic variant is a valid instrument for assessing the causal role of the

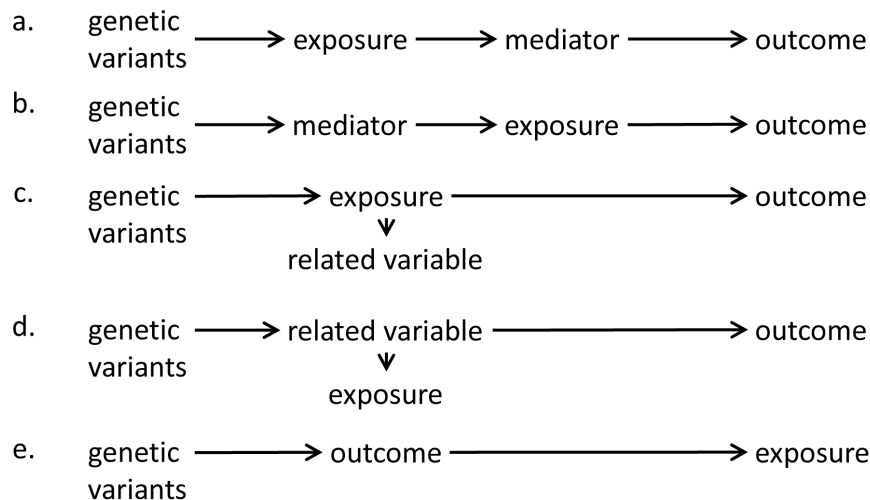


Figure 4. Directed acyclic graphs illustrating validity and invalidity of instrumental variable assumptions in different scenarios. **a)** Mediator is on causal pathway from exposure to outcome. **b)** Mediator is on causal pathway from genetic variants to exposure. **c)** Genetic variants influence the exposure, which has downstream effect on a related variable which does not affect the outcome. **d)** Genetic variants influence a related variable, and the related variable affects the outcome and exposure of interest. **e)** Genetic variants influence the outcome primarily, and only influence the exposure via the outcome. We note that the related variable may be known or unknown.

In scenarios a, b, and c, as there is no alternative pathway from the genetic variants to the outcome, the instrumental variable assumptions are satisfied. In scenario d, the pathway from the genetic variants to the outcome does not pass via the exposure, and so the instrumental variable assumptions are not satisfied for the exposure (although they are satisfied for the related variable). Scenarios a, b, and c are examples of “vertical pleiotropy” that do not invalidate the instrumental variable assumptions. Scenario d reflects a situation where the causal risk factor has been incorrectly identified – it is not the exposure, but the related variable. Scenario e reflects a reverse causation situation where the genetic variant has been incorrectly identified as primarily affecting the exposure.

exposure (assuming the other instrumental variable assumptions are satisfied), even if it is associated with another variable²⁹. In practice, distinguishing between horizontal pleiotropy and vertical pleiotropy requires knowledge of the relationships between the variables in the analysis. When there are multiple genetic variants, horizontal pleiotropy is more likely if a genetic association with a specific variable is only observed for a small number of variants. In contrast, vertical pleiotropy (in particular corresponding to the scenarios in Figure 4a–c) is likely to lead to genetic associations with that variable for all variants that associate with the exposure. While removing horizontally pleiotropic variants from a Mendelian randomization analysis should lead to more reliable results, care must be exercised, as removing vertically pleiotropic variants could lead to distorted causal estimates. Another possible scenario that would lead to instrument invalidity is if genetic variants influence the outcome primarily rather than the exposure (Figure 4e, see also discussion on reverse causation in Section 7). If there is a reverse causal effect of the outcome on the exposure, then genetic predictors of the outcome would be identified as hits in a genome-wide association study for the exposure. However, such variants would not be valid instrumental variables.

In conclusion, there is no one correct way to choose which genetic variants to include in an analysis. Causal conclusions will be more reliable when the instrumental variable assumptions are more plausible. Generally speaking, this means that analyses of exposures such as proteins conducted using variants in a coding gene region for the protein (referred to as “*cis*-variants”), or otherwise where variants having biological relevance to the exposure can be found, are likely to be more credible. Analyses based on *cis*-variants only are also likely to be more reliable for assessing the causal role of molecular phenotypes such as gene expression and DNA methylation. However, in many cases (and particularly for multifactorial exposures such as body mass index or blood pressure), it is not possible to find a *cis*-variant, and so a more agnostic polygenic analysis may be necessary. This allows investigators to test for consistency of the causal finding across multiple variants that influence the exposure via different biological pathways. A balance needs to be struck between including fewer variants (and potentially having insufficient power) and including more variants (and potentially including pleiotropic variants).

A practical suggestion for performing a polygenic analysis is to consider both a liberal analysis, including more genetic variants, and a conservative analysis, including fewer variants²⁹. While it is theoretically possible for pleiotropy to lead to a false negative finding, it is generally more likely that pleiotropy will bias estimates away from the null. Hence a null finding in a liberal analysis is more convincing evidence of a true null relationship – there is little evidence for a causal relationship even when potentially pleiotropic genetic variants are included in the analysis. Section 6 describes sensitivity analyses for assessing the instrumental variable assumptions and the robustness of non-null findings.

4. Variant harmonization

Genetic associations with exposures and outcomes are typically reported per additional copy of a particular allele. Hence, when combining summarized data on genetic associations, it is important to ensure that genetic associations are expressed per additional copy of the same allele⁴⁶. This is particularly important as not all publicly-available data resources are consistent about reporting strand information correctly. For example, if a genetic variant is a biallelic single nucleotide polymorphism (SNP) with alleles A and G on the positive strand, then the corresponding base pairs on the negative strand will be T and C. In this case, one dataset may report the association per additional copy of the A allele, and another per additional copy of the T allele – but the same comparison is being made. Allele and strand information can be double-checked by comparing allele frequency information – if the allele frequencies are similar for the A and T alleles, then the researcher can be more confident that this is a strand mismatch. Additional care should be taken for palindromic variants – if the alleles were A and T (or C and G), then the same alleles would appear on both the positive and negative strands. In such a case, if the allele frequency is close to 50%, it may be necessary to drop the variant from the analysis if it is not possible to verify that the alleles have been correctly orientated. While this is a conservative recommendation, allele alignment problems have led to incorrect results in Mendelian randomization analyses, and retractions and corrections of manuscripts⁴⁷.

5. Primary analysis

Different statistical methods have been proposed for Mendelian randomization with individual-level data and with summarized data. In a one-sample setting with individual-level data, a causal effect estimate can be obtained using the two-stage least-squares (2SLS) method. In the first stage, the exposure is regressed on the genetic variants and any relevant covariates. In the second stage the outcome is then regressed on the predicted values of the exposure from the first regression and the same covariates⁴⁸. In general, we recommend only including as covariates age, sex, genomic principal components of ancestry, and technical covariates (such as recruitment centre), as further adjustment may bias estimates either if adjustment is for a variable on the causal pathway from the genetic variants to the outcome (a mediator), or if adjustment induces collider bias⁴⁹. Strictly speaking, the 2SLS method refers to a two-stage analysis using linear regression for continuous outcomes and exposures. Similar two-stage analyses can be performed with binary variables using logistic regression⁵⁰, although in this case estimates are sensitive to correct specification of the first-stage regression model⁵¹ and other approaches that make weaker distributional assumptions, such as structural mean models, may be preferred⁵².

The 2SLS method can be applied to the two-sample setting if individual-level data are available for both samples⁵³. However, it is typical for two-sample investigations to use summarized data. With summarized data, if only one genetic variant is used as an instrument, the causal effect estimate is simply the ratio of the variant–outcome association to the variant–exposure

association. With multiple variants as instruments, the most commonly used method is the inverse-variance weighted (IVW) method¹⁸. With uncorrelated variants, the IVW estimate can be obtained from an IVW meta-analysis of the ratio estimates for the individual variants⁵⁴. The same estimate can equivalently be calculated as the ratio estimate using a weighted genetic risk score as a single instrument, with the weights equal to the associations of each variant with the exposure estimated in the first sample³⁰. A modification of this method has been proposed to allow for correlation (linkage disequilibrium) between variants³⁰. For continuous outcomes, the IVW estimate is asymptotically equivalent to the 2SLS estimate obtained from individual level data¹⁸. The 2SLS method (and thus also the IVW method) is the most efficient estimate of the causal effect when all genetic variants are valid instruments³⁰.

If all genetic variants are valid instruments and the relationships between all variables (genetic variants, exposure and outcome) are linear and homogeneous for all individuals in the population, then we would expect the variant-specific estimates (that is, the ratio estimates based on each variant in turn) to all target the same causal parameter, and for there to be no more heterogeneity between the variant-specific estimates than would be expected by chance alone¹². However, there are many reasons why excess heterogeneity may occur in practice. These include statistical reasons (such as departure from linearity and homogeneity across individuals) and biological reasons. For instance, variants associated with body mass index (BMI) influence BMI via different biological mechanisms⁵⁵. Additionally, some variants are associated with BMI from early childhood and others from adolescence or later. Variants that influence BMI for longer may be expected to have stronger proportional associations with chronic disease outcomes for which BMI is a cause. Hence if there is a true causal effect of the exposure on the outcome, some heterogeneity may be expected in the variant-specific causal estimates. However, heterogeneity would also arise if some genetic variants are not valid instrumental variables (see Section 6)⁵⁶.

The IVW method can be performed using a fixed-effects or a random-effects meta-analysis model. Unless there are very few variants (meaning that heterogeneity between the variant-specific estimates cannot be estimated reliably) or all variants are taken from the same gene region, we recommend using a multiplicative random-effects model as the default option for the IVW method. If there is no more heterogeneity between the ratio estimates for the individual variants than would be expected by chance alone, then the random-effect analysis is equivalent to the fixed-effect analysis, and there is no loss of precision in making the weaker random-effects assumption. However, if there is excess heterogeneity, then the fixed-effect analysis is inappropriate, as its confidence intervals are misleadingly narrow. A multiplicative random-effects model is preferred to the additive random-effects model that is more common in the meta-analysis literature as it does not change the relative weighting of the variant-specific estimates³². In contrast, an additive random-effects model upweights outlying estimates, which are more likely to represent pleiotropic variants. The multiplicative random-effects IVW

method provides valid causal estimates under the assumption of balanced pleiotropy; that is, pleiotropic effects on the outcome are equally likely to be positive as negative³².

We recommend the IVW method with multiplicative random-effects as the primary analysis method for use with summarized data, because it is the most efficient analysis method with valid instrumental variables, and it accounts for heterogeneity in the variant-specific causal estimates. If a causal effect is detected using this method, then investigators should proceed to perform sensitivity analyses to assess the robustness of their finding to the assumption of balanced pleiotropy.

A scenario that requires a different approach to the primary analysis occurs when there are several related exposures that have shared genetic predictors, meaning that it is difficult to find specific predictors of the individual exposures. In this case, a multivariable Mendelian randomization approach may be the primary analysis strategy⁵⁷. Multivariable Mendelian randomization is an extension to standard (univariable) Mendelian randomization that allows genetic variants to be associated with more than one exposure, and estimates the direct causal effects of each exposure in a single analysis model. The instrumental variable assumptions in multivariable Mendelian randomization require each variant to be associated with at least one of the exposures, not associated with the outcome via confounding, and not to affect the outcome except potentially via its association with one or more of the exposures included in the analysis model. For identification, it is also required that there is no perfect collinearity between the genetic associations; that is, there are variants that explain independent variation in each exposure⁵⁸. Examples of exposure sets where multivariable Mendelian randomization has been used include lipid fractions (such as high-density lipoprotein cholesterol, low-density lipoprotein (LDL) cholesterol, and triglycerides)⁵⁹, and body composition measures (such as fat mass and fat-free mass)⁶⁰. Provided that genetic variants act as instrumental variables for the set of exposures, the direct causal effects of the individual exposures on the outcome can be estimated⁶¹. Both the 2SLS and IVW methods can be adapted to the multivariable setting⁵⁸. A multivariable analysis strategy may also be worthwhile if genetic variants are associated with measured exposures that represent potentially pleiotropic pathways from the genetic variants to the outcome, as the effects of these exposures on the outcome will be accounted for in the multivariable analysis model (Section 7).

6. Robust methods for sensitivity analysis

A robust analysis method is defined here as a method that can provide valid causal inferences under weaker assumptions than the standard IVW method. Many robust analysis methods are available to detect and adjust for pleiotropy when using multiple genetic variants. Any polygenic Mendelian randomization investigation that does not perform one or more robust methods may be viewed as somewhat incomplete^{40,62}; investigators should consider using multiple methods that make different assumptions about the nature of the underlying pleiotropy²⁷. Although robust methods typically use the term 'pleiotropy', there is a mathematical correspondence between instrument invalidity and

pleiotropy⁶³, and so these methods can help assess sensitivity of findings to instrument invalidity more generally, and not simply invalidity that arises from horizontal pleiotropy. However, the robust methods are more likely to be effective for addressing instrument invalidity that arises due to issues such as pleiotropy or linkage disequilibrium with a variant influencing a confounder, which affect specific variants in a sporadic way, and less effective for instrument invalidity that arises due to issues such as population stratification or dynastic effects, which affect all variants in a systematic way. We here use the language of pleiotropy to make mathematically precise statements about the assumptions needed for methods to provide consistent estimates.

While a full comparison of all the robust methods that have been proposed is beyond the scope of this paper, a summary of several methods is provided as [Table 1](#). This table is based on a broader review and comparison of methods²⁷. We proceed to provide a brief description of some commonly used methods.

The most commonly used robust methods are MR-Egger, median- and mode-based methods, and MR-PRESSO. We focus on these methods here as they can be implemented using summarized data alone, and they rely on different assumptions to provide consistent causal estimates. The MR-Egger method estimates the causal effect as the slope from the weighted regression of the variant—outcome associations on the variant—exposure associations, and the average pleiotropic effect as the intercept. The method allows all genetic variants to have pleiotropic effects; however, it requires that the pleiotropic effects are independent of the variant—exposure associations (referred to as the Instrument Strength Independent of Direct Effect (InSIDE) assumption)¹⁹. A multivariable version of the MR-Egger method is available⁶⁴. Estimates from the MR-Egger method are particularly affected by outlying and influential datapoints⁶⁵, and are prone to be imprecise, particularly when the variant—exposure associations are all similar in magnitude. This can lead to the method having low power to detect a causal effect. A heterogeneity measure has been proposed to quantify the similarity between variant—exposure associations and the potential impact on MR-Egger analyses⁶⁶. A Bayesian model averaging method has also been proposed, which averages over the IVW and MR-Egger results using weights based on the degree of pleiotropy observed in the data⁶⁷.

The median- and mode-based methods^{20,21,68} rely on some genetic variants being valid instruments, but make weaker assumptions about the invalid instruments and are more robust to outliers. Specifically, the median-based method assumes that at least half of the variants are valid instruments (majority valid assumption), and the mode-based method assumes more variants estimate the true causal effect than estimate any other quantity (plurality valid assumption). Intuitively speaking, both methods take the variant-specific causal estimates (i.e. the ratio estimates based on the individual variants), and calculate a measure of central tendency of these estimates. These methods have a natural robustness to variants with outlying ratio

estimates, and so are not as affected by the presence of a small number of pleiotropic variants as the IVW and MR-Egger methods. The mode-based method has been shown to have low precision in some simulated and real datasets²⁷. Other methods have been proposed that make the same plurality valid assumption as the mode-based method, including the contamination mixture method²⁵ and MR-Mix²⁶.

The MR-PRESSO method is a variation on the IVW method that first removes genetic variants from the analysis whose variant-specific causal estimate differs substantially from those of other variants²². The IVW method is then performed for all variants that are not judged to be heterogeneous.

While it would be excessive to perform every robust method for Mendelian randomization that has been proposed, or even all the methods mentioned here, investigators should pick a sensible range of methods to assess the sensitivity of their findings. A recommendation is to perform the MR-Egger, median-based method, and mode-based method, as these methods require different assumptions to be satisfied for asymptotically consistent estimates. If estimates from all methods are similar, then any causal claim is more credible. However, differences between estimates does not necessarily imply the absence of a causal effect. Different methods will perform better and worse in different scenarios, so critical thought and judgement is required. Two recent simulation studies that compared different methods recommended the contamination mixture method²⁷ and MR-Mix⁶⁹ as having the lowest mean squared error across a range of different methods – these methods both make the same consistency assumption as the mode-based method, and so either could be used in preference to it.

We also recommend that a measure of the heterogeneity between variant-specific causal estimates, such as Cochran's Q statistic or the I² statistic, is reported as a part of a polygenic Mendelian randomization investigation^{56,70,71}. Conclusions are more reliable when multiple genetic variants provide concordant evidence for a causal effect, and particularly when there is no more heterogeneity between the variant-specific causal estimates than expected by chance. As discussed in Section 5, some heterogeneity may be expected even when all genetic variants are valid instruments. However, causal conclusions are less reliable when there is substantial heterogeneity, especially when there are distinct outliers (which may represent pleiotropic variants) or when evidence for a causal effect depends on one or a small number of variants.

Leave-one-out analyses (i.e. remove one variant from the analysis and re-estimate the causal effect) can be valuable in assessing the reliance of a Mendelian randomization analysis on a particular variant⁷². If there is one genetic variant that is particularly strongly associated with the exposure, then it may dominate the estimate of the causal effect. Investigators should assess the robustness of findings to the removal of such variants. If a causal effect is only evidenced by one variant, then the validity of inference depends only on that variant. If there are many variants in an analysis, leaving one variant out at a time is unlikely

to change the estimate substantially, and leaving out subsets of the variants (say, a randomly chosen 30% at a time⁷³) may be more appropriate. A further approach for identifying variants to remove from the analysis is Steiger filtering, which removes variants from the analysis if their association with the outcome is stronger than that with the exposure⁷⁴. It is unlikely that variants could have a stronger association with the outcome than the exposure if the instrumental variable assumptions are satisfied and the genetic association with the outcome is entirely mediated via the exposure.

While removing horizontally pleiotropic variants from a Mendelian randomization analysis will improve the validity of causal inferences, there is some danger in a *post hoc* or data-driven selection of genetic variants. This is particularly true if many genetic variants are judged to be heterogeneous: the removal of too many variants from the analysis could provide a false impression of agreement amongst the remaining variants, and over-precision in the causal estimate. Removing a variant from the analysis is more justified when a pleiotropic association of the variant has been identified⁷⁵.

A further class of robust methods uses latent modelling to distinguish to what extent genetic associations with the outcome arise due to a causal effect of the exposure, as opposed to via direct (pleiotropic) or confounder-driven effects of particular variants. A causal model is evidenced if the predominance of variants that associate with the exposure also associate with the outcome in a proportional way. If the genetic associations with the outcome do not follow this pattern, then a non-causal explanation would be preferred. Emerging methods that take this approach include the Causal Analyses Using Summary Effect Estimates (CAUSE)⁷⁶ and Latent Heritable Confounder Mendelian randomization (LHC-MR)⁷⁷ methods.

7. Other approaches for sensitivity analysis

Sensitivity analysis should not be limited to the application of different statistical methods. This is particularly important for investigations based on a single gene region, as several of the methods discussed above are not applicable in this case. Other approaches for assessing robustness include varying the dataset and choice of genetic variants in the analysis (including the suggestion of liberal and conservative variant sets in Section 3), the use of positive and negative control outcomes and/or samples, colocalization, subgroup analyses, and examining associations with potentially pleiotropic variables. We continue to describe each of these in turn.

A positive control outcome is an outcome for which it is already established that the exposure is causal. For example, the outcome of gout may be used as a positive control in a Mendelian randomization investigation for serum uric acid as an exposure, as raised uric acid levels are known to increase risk of gout. Provided that there is sufficient statistical power, if genetic variants that are associated with serum uric acid are not also associated with risk of gout, then we may question whether the genetic variants are truly able to assess the effects of varying serum uric acid⁷⁸. Conversely, a negative control outcome is

an outcome for which it is believed that the exposure cannot be causal. For example, pre-pubertal asthma was used as a negative control outcome in a Mendelian randomization study on the effects of age at puberty on asthma⁷⁹. If a Mendelian randomization investigation suggests that the negative control is caused by the exposure, then violation of the instrumental variable assumptions (such as through pleiotropy) may be suspected.

Colocalization assesses whether the same genetic variant (or variants) influences two traits^{80,81}. Even if genetic variants in a given gene region are associated with both an exposure and an outcome, this does not imply that the same genetic variants influence both exposure and outcome (implying the likely presence of a causal pathway including the exposure and outcome). It may be that the two associations are driven by different causal variants, and these variants are correlated due to linkage disequilibrium⁸². An example of this is the *APOE* gene region, in which genetic variants are associated with LDL-cholesterol and Alzheimer's disease, but LDL-cholesterol does not appear to be a cause of Alzheimer's disease⁸³. Colocalization can be useful for assessing exposures such as proteins and gene expression, particularly when the Mendelian randomization analysis is based on a single gene region⁸⁴. However, there are several limitations to such an analysis, including whether gene expression is estimated in a relevant tissue. Although colocalization differs from Mendelian randomization in a number of ways, the approach can provide complementary evidence supporting or questioning the presence of a biological mechanism linking the exposure and outcome via a common genetic predictor.

A subgroup analysis compares Mendelian randomization estimates (or equivalently genetic associations) estimated on different subgroups of the population in which the genetic variants have different degrees of association with the exposure. An example of subgroup analysis is the comparison of genetic associations with blood pressure in men and women in an East Asian population for variants implicated in the metabolism of alcohol^{85,86}. As women in East Asia tend not to drink alcohol, genetic associations with blood pressure are observed in men but not in women. Also, genetic associations are stronger in heavier drinkers⁸⁵. This provides confidence that the genetic associations are driven by alcohol consumption and not by a pleiotropic mechanism. Such an analysis can be performed if there is a subgroup of the population that has reduced or increased levels of the exposure^{87,88}. However, if the subgroup is defined by a collider (see below), then stratification can introduce bias to the analysis⁸⁹.

A further possible sensitivity analysis is to check the genetic associations with other variables associated with the outcome, and which are thought not to lie on the causal pathway through the exposure (i.e. are not mediators). Such variables may lie on alternative pleiotropic pathways to the outcome. If the genetic variants are not associated with such variables, then some reassurance can be drawn that the Mendelian randomization assumptions are satisfied. A further possibility in this case is to perform a multivariable Mendelian randomization, including the putative pleiotropic variables as additional exposures in the analysis model⁹⁰. This analysis will estimate the direct effect of the exposure on the outcome keeping these variables constant.

Methods have been proposed based on a multivariable approach in the context of gene expression data, including the MR-link⁹¹, and transcriptome-wide summary statistics-based Mendelian Randomization (TWMR)⁹² methods.

There are several other potential sources of bias in a Mendelian randomization analysis other than invalid instruments. We consider here collider bias, selection bias, and reverse causation as three potential sources of bias, and direct readers to reviews that list further potential sources of bias^{45,93}.

A collider is a common effect of two variables – for example, the any variable causally downstream of the exposure is influenced by the genetic variants and the exposure—outcome confounders, and so is a collider. Even if two variables are unrelated (they are marginally independent), they will typically be related when conditioning on the collider (conditionally dependent)⁸⁹. Stratifying on or adjusting for a collider therefore leads to an association between variables that influence the collider. An association between the genetic variants and the exposure—outcome confounders would lead to biased causal estimates⁹⁴. Collider bias is not unique to Mendelian randomization, but it is particularly relevant as some published genetic association estimates have been adjusted for potential colliders. Methods to account for collider bias have recently been proposed⁹⁵.

Selection bias is a specific example of collider bias which occurs when selection into a study sample depends on a collider. Simulation studies have shown that selection bias can have a severe impact on Mendelian randomization estimates, but only when the associations of variables with the collider are quite strong^{49,94}. Selection bias can potentially be addressed using inverse-probability weighting, although this requires estimation of the probability of selection into the study sample for all individuals.

While the genetic code is fixed at conception and so cannot be influenced by reverse causation, if the outcome influences the risk factor, this can result in gene—outcome associations becoming distorted and lead to misleading inferences. As discussed above and shown in Figure 4e, if genetic variants used as instrumental variables for the exposure in fact influence the outcome primarily, then genetic associations with the outcome could be present without the exposure influencing the outcome. The MR-Steiger method has been developed to detect such variants and remove them from the analysis⁷⁴. Bidirectional Mendelian randomization analyses have been proposed that use separate sets of instrumental variables for the exposure and outcome to assess the direction of causal effect⁹⁶.

8. Data presentation

An attractive feature of Mendelian randomization is that the analysis can be summarized graphically in a transparent way. For example, in a polygenic analysis, a scatter plot of the genetic associations with the outcome against the genetic associations with the exposure reveals much about the analysis – whether different genetic variants provide similar estimates of the causal effect or if there is considerable heterogeneity, and whether the analysis

is dominated by a single genetic variant or not²⁹. The scatter plot is appealing as it presents the data with no manipulation. Examples of scatter plots illustrating heterogeneity and no heterogeneity in the causal estimates from different variants are shown in Figure 5. Alternatives are forest plots, funnel plots, and radial plots – each of these assesses heterogeneity in the variant-specific causal estimates⁹⁷. Plots allow the investigators and readers to assess the reliability of the analysis method and its underlying assumptions, and we strongly recommend their inclusion in a manuscript.

Other important information to report include the R² statistic (when the exposure is continuous), which is a measure of the variance in the exposure explained by the genetic variants, and (particularly in a one-sample setting) the related F statistic, which is a measure of instrument strength and can be used to judge the extent of weak instrument bias⁹⁸. Investigators can also make some statement about the power of their analysis. Power to detect a causal effect depends on the proportion of variance in the exposure explained by the genetic variants, proposed size of causal effect, sample size (for the genetic associations with the outcome), and (with a binary outcome) proportion of individuals with an outcome event. Power calculators can be found at <http://cnsgenomics.com/shiny/mRnd/> and <https://sb452.shinyapps.io/power/>. Power calculations are often performed *post hoc*, as sample sizes are rarely determined based on a proposed Mendelian randomization analysis. Power calculations are more meaningful when performed prior to the analysis, and can guide investigators which exposure/outcome pairs to consider, and so focus on analyses that have a better chance of giving meaningful results.

9. Interpretation

Finally, we discuss the interpretation of findings from Mendelian randomization investigations. In the first instance, a Mendelian randomization investigation assesses the association of genetic predictors of an exposure with an outcome, or equivalently, the association of genetically-predicted levels of an exposure with an outcome. Making causal inferences from observational data always relies on untestable assumptions. In Mendelian randomization, the assumption is that observed differences in the outcome associated with genetically predicted levels of the exposure would also be seen if the exposure were intervened on^{7,48}. In line with the STROBE-MR guidelines¹, we recommend that a cautious interpretation should be taken when describing the extent to which a causal effect has been demonstrated by a Mendelian randomization investigation. The appropriate degree of caution will depend on the plausibility of the instrumental variable assumptions, the concordance of estimates from different methods and different analytical approaches, the results from sensitivity and supplementary analyses, and so on.

Mendelian randomization estimates relate specifically to changes in the exposure induced by the genetic variants used as instrumental variables. The genetic code is fixed at conception, and so Mendelian randomization investigations typically compare groups of the population having different trajectories in their distribution of the exposure over time⁹⁹. Analyses

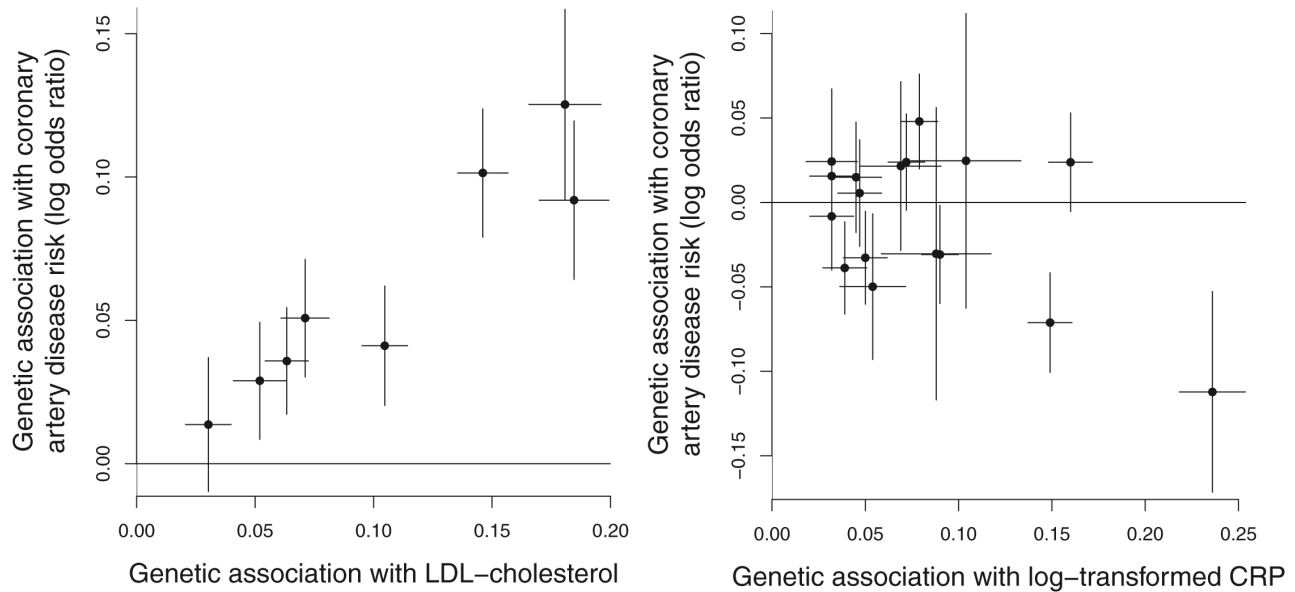


Figure 5. Scatter plot of genetic associations with the outcome (vertical axis) against genetic associations with the exposure (horizontal axis). Examples illustrated are: (left) no heterogeneity in the variant-specific causal estimates (effect of LDL-cholesterol on coronary heart disease risk using 8 variants associated with LDL-cholesterol); and (right) heterogeneity in the variant-specific causal estimates (effect of C-reactive protein on coronary heart disease risk using 17 genome-wide significant predictors of C-reactive protein). As indicated by differences in estimates, not all genetic variants are valid instrumental variables for C-reactive protein, and so a causal interpretation is not appropriate. Taken from Burgess *et al.*, 2018⁶⁸.

therefore typically can be interpreted as assessing the impact of long-term elevated levels of an exposure. However, in most cases, we have incomplete information about how the genetic variant changes the distribution of the exposure across the life course. If the genetic associations with the exposure vary over time, then Mendelian randomization estimates based on genetic associations with the exposure measured at a single timepoint can be unreliable³¹. Similar difficulties of interpretation arise if the impact on the outcome relates to levels of the exposure at a specific time period in life. A plausible example of this is the effect of vitamin D on multiple sclerosis; multiple sclerosis risk is hypothesized to be influenced by vitamin D levels during early childhood, but not vitamin D levels in adulthood¹⁰⁰. If measurements of genetic associations with the exposure are available at different timepoints, then multivariable Mendelian randomization analyses can be performed to distinguish between the effect of the exposure at each time point¹⁰¹.

That said, results from Mendelian randomization investigations have often been shown to qualitatively agree with the results from randomized trials, suggesting that a causal interpretation for Mendelian randomization findings is often reasonable⁹³. Mendelian randomization investigations are worthwhile in providing an alternative line of aetiological evidence even if though the instrumental variable assumptions can never be proved beyond all doubt^{102,103}. However, quantitative differences between estimates from Mendelian randomization and from trials are likely, particularly as there are differences between how genetic variants influence the exposure and how clinical and pharmaceutical interventions influence the exposure¹⁰⁴. As genetic

variants typically affect usual levels of exposures on a long-term basis, Mendelian randomization estimates are often larger than those from conventional observational studies or randomized trials³¹. Hence, the causal estimate from a Mendelian randomization investigation should not generally be interpreted directly as the expected impact of intervening on the exposure in applied practice¹⁰⁵. The estimate from a Mendelian randomization investigation is therefore better interpreted as a test statistic for a causal hypothesis rather than the estimated impact of a well-defined intervention at a specific point in time. But even when a Mendelian randomization investigation is performed primarily to assess the causal role of an exposure, causal estimates can still be useful, for example to assess heterogeneity in estimates from different variants as a test of instrument validity, or to compare results from different analysis methods as an assessment of robustness⁵⁶. A logical consequence of the 2SLS/IVW method providing the most efficient causal estimate when combining evidence across multiple valid instrumental variables is that, under the same assumptions, the method provides the most powerful test of the presence of a causal effect.

Summary

Overall, the key elements of a Mendelian randomization investigation that we would expect to be present in any manuscript are: i) motivation for why a Mendelian randomization analysis should be performed and for the scope of the analysis, ii) a clear description and justification of the choice of dataset(s) for the analysis, including why a one- or two-sample approach was chosen for the primary analysis, iii) a clear

description and justification of the choice of genetic variants used in the analysis, iv) a discussion, whether statistically or biologically led, of whether the genetic variants are likely to satisfy the instrumental variable assumptions, v) a clear graphical presentation of the data, such as a scatter plot of the genetic associations, and vi) some attempt to test the robustness of the main findings, whether by use of robust methods (for a polygenic analysis) or another approach – whatever is most appropriate to the analysis under consideration. Without these elements, the reader is not fully able to judge the reliability of a Mendelian randomization investigation.

Particularly with the advent of summarized data and the two-sample setting, performing a Mendelian randomization analysis has become more straightforward²⁸. The difficulty is not in performing a Mendelian randomization analysis, but rather

in performing a credible analysis¹⁰⁶ and providing a reasoned interpretation¹⁰⁰. We hope that these guidelines, summarized in the accompanying flowcharts (Figure 1 and Figure 2) and checklist (Figure 3), will aid practitioners in performing reliable analyses, and editors and reviewers in judging the reliability of analyses, and that their use will help improve the overall quality of Mendelian randomization investigations.

Data availability

No data are associated with this article.

Disclaimer

The views expressed in this article are those of the authors. Publication in Wellcome Open Research does not imply endorsement by Wellcome.

References

- STROBE-MR Steering Group: **STROBE-MR: Guidelines for strengthening the reporting of Mendelian randomization studies.** *PeerJ*. 2019. [PubMed Abstract](#) | [Publisher Full Text](#)
- Smith GD, Ebrahim S: 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. 2003; **32**(1): 1–22. [PubMed Abstract](#) | [Publisher Full Text](#)
- Burgess S, Thompson SG: **Mendelian Randomization: Methods for using Genetic Variants in Causal Estimation.** Taylor & Francis Group; 2015. [Reference Source](#)
- Davies NM, Holmes MV, Davey Smith G: **Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians.** *BMJ*. 2018; **362**: k601. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Greenland S: **An introduction to instrumental variables for epidemiologists.** *Int J Epidemiol*. 2000; **29**(4): 722–729. [PubMed Abstract](#) | [Publisher Full Text](#)
- Martens EP, Pestman WR, de Boer A, et al.: **Instrumental variables: application and limitations.** *Epidemiology*. 2006; **17**(3): 260–267. [PubMed Abstract](#) | [Publisher Full Text](#)
- Didelez V, Sheehan N: **Mendelian randomization as an instrumental variable approach to causal inference.** *Stat Methods Med Res*. 2007; **16**(4): 309–330. [PubMed Abstract](#) | [Publisher Full Text](#)
- Swanson SA, Labrecque J, Hernán MA: **Causal null hypotheses of sustained treatment strategies: What can be tested with an instrumental variable?** *Eur J Epidemiol*. 2018; **33**(8): 723–728. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, et al.: **Methodological challenges in Mendelian randomization.** *Epidemiology*. 2014; **25**(3): 427–435. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Larsson SC, Traylor M, Malik R, et al.: **Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis.** *BMJ*. 2017; **359**: j5375. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hernán MA, Taubman SL: **Does obesity shorten life? The importance of well-defined interventions to answer causal questions.** *Int J Obes (Lond)*. 2008; **32** Suppl 3: S8–S14. [PubMed Abstract](#) | [Publisher Full Text](#)
- Hernán MA, Robins JM: **Instruments for causal inference: an epidemiologist's dream?** *Epidemiology*. 2006; **17**(4): 360–372. [PubMed Abstract](#) | [Publisher Full Text](#)
- Bulik-Sullivan BK, Loh PR, Finucane HK, et al.: **LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.** *Nat Genet*. 2015; **47**(3): 291–5. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- O'Connor LJ, Price AL: **Distinguishing genetic correlation from causation across 52 diseases and complex traits.** *Nat Genet*. 2018; **50**(12): 1728–1734. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Millard LA, Davies NM, Timpson NJ, et al.: **MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization.** *Sci Rep*. 2015; **5**: 16645. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Li X, Meng X, Spiliopoulou A, et al.: **MR-PheWAS: exploring the causal effect of SUA level on multiple disease outcomes by using genetic instruments in UK Biobank.** *Ann Rheum Dis*. 2018; **77**(7): 1039–1047. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gill D, Benyamin B, Moore LSP, et al.: **Associations of genetically determined iron status across the phenome: A mendelian randomization study.** *PLoS Med*. 2019; **16**(6): e1002833. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Burgess S, Butterworth A, Thompson SG: **Mendelian randomization analysis with multiple genetic variants using summarized data.** *Genet Epidemiol*. 2013; **37**(7): 658–665. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bowden J, Davey Smith G, Burgess S: **Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.** *Int J Epidemiol*. 2015; **44**(2): 512–525. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bowden J, Davey Smith G, Haycock PC, et al.: **Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator.** *Genet Epidemiol*. 2016; **40**(4): 304–314. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hartwig FP, Davey Smith G, Bowden J: **Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption.** *Int J Epidemiol*. 2017; **46**(6): 1985–1998. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Verbanck M, Chen CY, Neale B, et al.: **Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases.** *Nat Genet*. 2018; **50**(5): 693–698. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rees JMB, Wood AM, Dudbridge F, et al.: **Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates.** *PLoS One*. 2019; **14**(9): e0222362. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhao Q, Wang J, Hemani G, et al.: **Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score.** *arXiv*. 2018; 1801.09652. [Reference Source](#)
- Burgess S, Foley CN, Allara E, et al.: **A robust and efficient method for Mendelian randomization with hundreds of genetic variants: unravelling mechanisms linking HDL-cholesterol and coronary heart disease.** *Nat Commun*. 2020; **11**(1): 376. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Qi G, Chatterjee N: **Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects.** *Nat Commun*. 2019; **10**(1): 1941. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

27. Slob E, Burgess S: **A comparison of robust Mendelian randomization methods using summary data.** *Genet Epidemiol.* 2020.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Hemani G, Zheng J, Elsworth B, *et al.*: **The MR-Base platform supports systematic causal inference across the human phenome.** *eLife.* 2018; 7: pii: e34408.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Burgess S, Scott RA, Timpson NJ, *et al.*: **Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors.** *Eur J Epidemiol.* 2015; 30(7): 543–552.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Burgess S, Dudbridge F, Thompson SG: **Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods.** *Stat Med.* 2016; 35(11): 1880–1906.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Labrecque JA, Swanson SA: **Interpretation and Potential Biases of Mendelian Randomization Estimates With Time-Varying Exposures.** *Am J Epidemiol.* 2019; 188(1): 231–238.
[PubMed Abstract](#) | [Publisher Full Text](#)
32. Bowden J, Del Greco MF, Minelli C, *et al.*: **A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization.** *Stat Med.* 2017; 36(11): 1783–1802.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Staley JR, Blackshaw J, Kamat MA, *et al.*: **PhenoScanner: a database of human genotype-phenotype associations.** *Bioinformatics.* 2016; 32(20): 3207–3209.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Burgess S, Davies NM, Thompson SG, *et al.*: **Instrumental variable analysis with a nonlinear exposure-outcome relationship.** *Epidemiology.* 2014; 25(6): 877–885.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Staley JR, Burgess S: **Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization.** *Genet Epidemiol.* 2017; 41(4): 341–352.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
36. Pierce BL, Burgess S: **Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators.** *Am J Epidemiol.* 2013; 178(7): 1177–1184.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
37. Burgess S, Thompson SG, CRP CHD Genetics Collaboration: **Avoiding bias from weak instruments in Mendelian randomization studies.** *Int J Epidemiol.* 2011; 40(3): 755–764.
[PubMed Abstract](#) | [Publisher Full Text](#)
38. Burgess S, Davies NM, Thompson SG: **Bias due to participant overlap in two-sample Mendelian randomization.** *Genet Epidemiol.* 2016; 40(7): 597–608.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Swerdlow DI, Kuchenbaecker KB, Shah S, *et al.*: **Selecting instruments for Mendelian randomization in the wake of genome-wide association studies.** *Int J Epidemiol.* 2016; 45(5): 1600–1616.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
40. Burgess S, Bowden J, Fall T, *et al.*: **Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants.** *Epidemiology.* 2017; 28(1): 30–42.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Brion MJ, Shakhbuzov K, Visscher PM: **Calculating statistical power in Mendelian randomization studies.** *Int J Epidemiol.* 2013; 42(5): 1497–1501.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Mokry LE, Ross S, Ahmad OS, *et al.*: **Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study.** *PLoS Med.* 2015; 12(8): e1001866.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
43. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium Swerdlow DI, Holmes MV, *et al.*: **The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis.** *Lancet.* 2012; 379(9822): 1214–1224.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
44. Zhao Q, Chen Y, Wang J, *et al.*: **Powerful three-sample genome-wide design and robust statistical inference in summary-data Mendelian randomization.** *Int J Epidemiol.* 2019; 48(5): 1478–1492.
[PubMed Abstract](#) | [Publisher Full Text](#)
45. Davey Smith G, Hemani G: **Mendelian randomization: genetic anchors for causal inference in epidemiological studies.** *Hum Mol Genet.* 2014; 23(R1): R89–R98.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
46. Hartwig FP, Davies NM, Hemani G, *et al.*: **Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique.** *Int J Epidemiol.* 2016; 45(6): 1717–1726.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
47. Inoshita M, Numata S, Tajima A, *et al.*: **Retraction: A significant causal association between C-reactive protein levels and schizophrenia.** *Sci Rep.* 2018; 8: 46947.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
48. Lawlor DA, Harbord RM, Sterne JA, *et al.*: **Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.** *Stat Med.* 2008; 27(8): 1133–1163.
[PubMed Abstract](#) | [Publisher Full Text](#)
49. Hughes RA, Davies NM, Davey Smith G, *et al.*: **Selection Bias When Estimating Average Treatment Effects Using One-sample Instrumental Variable Analysis.** *Epidemiology.* 2019; 30(3): 350–357.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
50. Didelez V, Meng S, Sheehan NA: **Assumptions of IV Methods for Observational Epidemiology.** *Stat Sci.* 2010; 25(1): 22–40.
[Publisher Full Text](#)
51. Vansteelandt S, Bekaert M, Claeskens G: **On model selection and model misspecification in causal inference.** *Stat Methods Med Res.* 2012; 21(1): 7–30.
[PubMed Abstract](#) | [Publisher Full Text](#)
52. Palmer TM, Sterne JA, Harbord RM, *et al.*: **Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses.** *Am J Epidemiol.* 2011; 173(12): 1392–1403.
[PubMed Abstract](#) | [Publisher Full Text](#)
53. Inoue A, Solon G: **Two-Sample Instrumental Variables Estimators.** *Rev Econ Stat.* 2010; 92(3): 557–561.
[Publisher Full Text](#)
54. Thompson JR, Minelli C, Del Greco MF: **Mendelian Randomization using Public Data from Genetic Consortia.** *Int J Biostat.* 2016; 12(2): pii: /ijb.2016.12.issue-2/ijb-2015-0074/ijb-2015-0074.xml.
[PubMed Abstract](#) | [Publisher Full Text](#)
55. Walter S, Kubzansky LD, Koenen KC, *et al.*: **Revisiting Mendelian randomization studies of the effect of body mass index on depression.** *Am J Med Genet B Neuropsychiatr Genet.* 2015; 168B(2): 108–115.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
56. Greco MFD, Minelli C, Sheehan NA, *et al.*: **Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome.** *Stat Med.* 2015; 34(21): 2926–2940.
[PubMed Abstract](#) | [Publisher Full Text](#)
57. Burgess S, Thompson SG: **Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects.** *Am J Epidemiol.* 2015; 181(4): 251–260.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
58. Sanderson E, Davey Smith G, Windmeijer F, *et al.*: **An examination of multivariable Mendelian randomization in the single-sample and two-sample data settings.** *Int J Epidemiol.* 2019; 48(3): 713–727.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
59. Burgess S, Freitag DF, Khan H, *et al.*: **Using multivariable Mendelian randomization to disentangle the causal effects of lipid fractions.** *PLoS One.* 2014; 9(10): e108891.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
60. Larsson SC, Bäck M, Rees JMB, *et al.*: **Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study.** *Euro Heart J.* 2019; pii: ehz388.
[PubMed Abstract](#) | [Publisher Full Text](#)
61. Burgess S, Thompson DJ, Rees JMB, *et al.*: **Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer.** *Genetics.* 2017; 207(2): 481–487.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
62. Hemani G, Bowden J, Davey Smith G: **Evaluating the potential role of pleiotropy in Mendelian randomization studies.** *Hum Mol Genet.* 2018; 27(R2): R195–R208.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
63. Kang H, Zhang A, Cai TT, *et al.*: **Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization.** *J Am Stat Assoc.* 2016; 111(513): 132–144.
[Publisher Full Text](#)
64. Rees JM, Wood AM, Burgess S: **Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy.** *Stat Med.* 2017; 36(29): 4705–4718.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
65. Burgess S, Thompson SG: **Interpreting findings from Mendelian randomization using the MR-Egger method.** *Eur J Epidemiol.* 2017; 32(5): 377–389.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
66. Bowden J, Del Greco MF, Minelli C, *et al.*: **Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic.** *Int J Epidemiol.* 2016; 45(6): 1961–1974.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
67. Thompson JR, Minelli C, Bowden J, *et al.*: **Mendelian randomization incorporating uncertainty about pleiotropy.** *Stat Med.* 2017; 36(29): 4627–4645.
[PubMed Abstract](#) | [Publisher Full Text](#)
68. Burgess S, Zuber V, Gkatzionis A, *et al.*: **Modal-based estimation via heterogeneity-penalized weighting: model averaging for consistent and efficient estimation in Mendelian randomization when a plurality of candidate instruments are valid.** *Int J Epidemiol.* 2018; 47(4): 1242–1254.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
69. Qi G, Chatterjee N: **A Comprehensive Evaluation of Methods for Mendelian Randomization Using Realistic Simulations of Genome-wide Association Studies.** *bioRxiv.* 2019; 702787.
[Publisher Full Text](#)
70. Bowden J, Del Greco MF, Minelli C, *et al.*: **Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption.** *Int J Epidemiol.* 2019; 48(3): 728–742.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

71. Bowden J, Hemani G, Davey Smith G: **Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization-A Job for the Humble Heterogeneity Statistic?** *Am J Epidemiol.* 2018; **187**(12): 2681–2685. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
72. Corbin LJ, Richmond RC, Wade KH, *et al.*: **BMI as a Modifiable Risk Factor for Type 2 Diabetes: Refining and Understanding Causal Estimates Using Mendelian Randomization.** *Diabetes.* 2016; **65**(10): 3002–3007. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
73. Smith JG, Luk K, Schulz CA, *et al.*: **Association of low-density lipoprotein cholesterol-related genetic variants with aortic valve calcium and incident aortic stenosis.** *JAMA.* 2014; **312**(17): 1764–1771. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
74. Hemani G, Tilling K, Davey Smith G: **Orienting the causal relationship between imprecisely measured traits using GWAS summary data.** *PLoS Genet.* 2017; **13**(11): e1007081. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
75. Cho Y, Haycock PC, Sanderson E, *et al.*: **Exploiting horizontal pleiotropy to search for causal pathways within a Mendelian randomization framework.** *Nat Commun.* 2020; **11**(1): 1010. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
76. Morrison J, Knoblauch N, Marcus J, *et al.*: **Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics.** *bioRxiv.* 2019; 682237. [Publisher Full Text](#)
77. Darrous L, Mounier N, Kutalik Z: **Simultaneous estimation of bi-directional causal effects and heritable confounding from GWAS summary statistics.** *medRxiv.* 2020; 2020.2001.2027.20018929. [Publisher Full Text](#)
78. Palmer TM, Nordestgaard BG, Benn M, *et al.*: **Association of plasma uric acid with ischaemic heart disease and blood pressure: mendelian randomisation analysis of two large cohorts.** *BMJ.* 2013; **347**: f4262. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
79. Minelli C, van der Plaat DA, Leynaert B, *et al.*: **Age at puberty and risk of asthma: A Mendelian randomisation study.** *PLoS Med.* 2018; **15**(8): e1002634. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
80. Giambartolomei C, Vukcevic D, Schadt EE, *et al.*: **Bayesian test for colocalisation between pairs of genetic association studies using summary statistics.** *PLoS Genet.* 2014; **10**(5): e1004383. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
81. Hormozdizari F, Kostem E, Kang EY, *et al.*: **Identifying causal variants at loci with multiple signals of association.** *Genetics.* 2014; **198**(2): 497–508. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
82. Solovieff N, Cotsapas C, Lee PH, *et al.*: **Pleiotropy in complex traits: challenges and strategies.** *Nat Rev Genet.* 2013; **14**(7): 483. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
83. Benn M, Nordestgaard BG, Frikke-Schmidt R, *et al.*: **Low LDL cholesterol, PCSK9 and HMGCR genetic variation, and risk of Alzheimer's disease and Parkinson's disease: Mendelian randomisation study.** *BMJ.* 2017; **357**: j1648. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
84. Zheng J, Haberland V, Baird D, *et al.*: **Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases.** *bioRxiv.* 2019; 627398. [Publisher Full Text](#)
85. Chen L, Smith GD, Harbord RM, *et al.*: **Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach.** *PLoS Med.* 2008; **5**(3): e52. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
86. Millwood IY, Walters RG, Mei XW, *et al.*: **Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China.** *Lancet.* 2019; **393**(10183): 1831–1842. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
87. van Kippersluis H, Rietveld CA: **Pleiotropy-robust Mendelian randomization.** *Int J Epidemiol.* 2018; **47**(4): 1279–1288. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
88. Spiller W, Slichter D, Bowden J, *et al.*: **Detecting and correcting for bias in Mendelian randomization analyses using Gene-by-Environment interactions.** *Int J Epidemiol.* 2019; **48**(3): 702–712. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
89. Munafò MR, Tilling K, Taylor AE, *et al.*: **Collider scope: when selection bias can substantially influence observed associations.** *Int J Epidemiol.* 2018; **47**(1): 226–235. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
90. White J, Sofat R, Hemani G, *et al.*: **Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis.** *Lancet Diabetes Endocrinol.* 2016; **4**(4): 327–336. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
91. van der Graaf A, Claringbould A, Rimbart A, *et al.*: **A novel Mendelian randomization method identifies causal relationships between gene expression and low-density lipoprotein cholesterol levels.** *bioRxiv.* 2019; 671537. [Publisher Full Text](#)
92. Porcu E, Rieger S, Lepik K, *et al.*: **Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits.** *Nat Commun.* 2019; **10**(1): 3300. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
93. Haycock PC, Burgess S, Wade KH, *et al.*: **Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies.** *Am J Clin Nutr.* 2016; **103**(4): 965–978. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
94. Gkatzionis A, Burgess S: **Contextualizing selection bias in Mendelian randomization: how bad is it likely to be?** *Int J Epidemiol.* 2019; **48**(3): 691–701. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
95. Dudbridge F, Allen RJ, Sheehan NA, *et al.*: **Adjustment for index event bias in genome-wide association studies of subsequent events.** *Nat Commun.* 2019; **10**(1): 1561. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
96. Burgess S, Daniel RM, Butterworth AS, *et al.*: **Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways.** *Int J Epidemiol.* 2015; **44**(2): 484–495. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
97. Bowden J, Spiller W, Del Greco F, *et al.*: **Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the Radial plot and Radial regression.** *Int J Epidemiol.* 2018; **47**(4): 1264–1278. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
98. Burgess S, Thompson SG: **Bias in causal estimates from Mendelian randomization studies with weak instruments.** *Stat Med.* 2011; **30**(11): 1312–1323. [PubMed Abstract](#) | [Publisher Full Text](#)
99. Swanson SA, Tiemeier H, Ikram MA, *et al.*: **Nature as a Trialist?: Deconstructing the Analogy Between Mendelian Randomization and Randomized Trials.** *Epidemiology.* 2017; **28**(5): 653–659. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
100. Holmes MV, Ala-Korpela M, Smith GD: **Mendelian randomization in cardiometabolic disease: challenges in evaluating causality.** *Nat Rev Cardiol.* 2017; **14**(10): 577–590. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
101. Richardson TG, Sanderson E, Elsworth B, *et al.*: **Can the impact of childhood adiposity on disease risk be reversed? A Mendelian randomization study.** *medRxiv.* 2019; 19008011. [Publisher Full Text](#)
102. Munafò MR, Davey Smith G: **Robust research needs many lines of evidence.** *Nature.* 2018; **553**(7689): 399–401. [PubMed Abstract](#) | [Publisher Full Text](#)
103. Lawlor DA, Tilling K, Davey Smith G: **Triangulation in aetiological epidemiology.** *Int J Epidemiol.* 2016; **45**(6): 1866–1886. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
104. Burgess S, Butterworth A, Malarstig A, *et al.*: **Use of Mendelian randomisation to assess potential benefit of clinical intervention.** *BMJ.* 2012; **345**: e7325. [PubMed Abstract](#) | [Publisher Full Text](#)
105. Burgess S, Ference BA, Staley JR, *et al.*: **Association of LPA Variants With Risk of Coronary Disease and the Implications for Lipoprotein(a)-Lowering Therapies: A Mendelian Randomization Analysis.** *JAMA Cardiol.* 2018; **3**(7): 619–627. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
106. Burgess S, Davey Smith G: **How humans can contribute to Mendelian randomization analyses.** *Int J Epidemiol.* 2019; **48**(3): 661–664. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  

Version 2

Reviewer Report 24 July 2020

<https://doi.org/10.21956/wellcomeopenres.17423.r38569>

© 2020 Swanson S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Sonja A. Swanson

Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands

I appreciate the authors' willingness to engage with my suggestions in the prior review. I look forward to seeing this living document continue to improve as the methods, settings, and therefore guidelines relevant to Mendelian randomization investigators continue to evolve.

Competing Interests: I have published work with some of the authors before but I do not believe this affected my ability to review impartially.

Reviewer Expertise: Epidemiology; causal inference.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 17 March 2020

<https://doi.org/10.21956/wellcomeopenres.17034.r37985>

© 2020 Swanson S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Sonja A. Swanson

Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands

These guidelines are a welcomed addition to the literature and reflect the experience and common practices performed by some thoughtful and influential researchers in the MR community. I am especially

encouraged to see that this is a living document that the researchers expect to update periodically. Generally, the guidelines reflect consensus on analytic practices with some appropriate caveats provided for the targeted audience. In this review, I highlight areas that are not discussed in great detail that I think warrant attention in the development of further guidelines.

- 1. Well-defined causal effect estimands and causal null hypotheses:** Part 1 carefully separates the goal of testing for a non-null effect and estimating a causal effect. Its consideration of testing, however, does not make clear what types of null hypotheses can really be assessed (see Swanson *et al.* 2018)¹, and the description on effect estimation only suggests qualitatively weighing whether the genetic variants act like the intervention of interest. More guidance could be provided on question framing, especially by giving more specifics on the types of causal questions MR is able to answer. See also Hernán & Robins (2006)² and Swanson *et al.* (2017)³.
- 2. The importance of time:** Related to the framing of the research question is also how time is a part of this framing. MR estimates are often described as “lifetime effects” but the standard IV methods used in MR are developed for time-fixed exposures. (Note that this type of concern also applies to causal null hypothesis testing, as described in Swanson *et al.* 2018)¹. At very least, some qualitative guidance on how to interpret results in light of the time-varying nature of exposures and outcomes over the life-course is needed. See again Swanson *et al.* (2017)³ and also Labrecque & Swanson (2019)⁴.
- 3. Available robust methods, sensitivity analyses, and falsification strategies:** These guidelines state that Sections 6 and 7 are not exhaustive considerations of these tools, but it would be helpful to the MR community if guidance on these points was expanded in future versions of these guidelines. It is not very clear why the authors chose to present the tools they did have room for, and not others. It also is not very clear why these sections focus so much on the risk of bias due to pleiotropy when other types of biases can also threaten MR estimates.

References

1. Swanson SA, Labrecque J, Hernán MA: Causal null hypotheses of sustained treatment strategies: What can be tested with an instrumental variable?. *Eur J Epidemiol.* 2018; **33** (8): 723-728 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Hernán MA, Robins JM: Instruments for causal inference: an epidemiologist's dream?. *Epidemiology.* 2006; **17** (4): 360-72 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Swanson SA, Tiemeier H, Ikram MA, Hernán MA: Nature as a Trialist?: Deconstructing the Analogy Between Mendelian Randomization and Randomized Trials. *Epidemiology.* **28** (5): 653-659 [PubMed Abstract](#) | [Publisher Full Text](#)
4. Labrecque J, Swanson S: Interpretation and Potential Biases of Mendelian Randomization Estimates With Time-Varying Exposures. *American Journal of Epidemiology.* 2019; **188** (1): 231-238 [Publisher Full Text](#)

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: I have published work with some of the authors before but I do not believe this affected my ability to review impartially.

Reviewer Expertise: Epidemiology; causal inference.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 20 Apr 2020

Stephen Burgess, University of Cambridge, Cambridge, UK

> We thank the reviewers for their comments, which we have numbered for reference. We provide point-by-point responses to their suggestions below. <

Reviewer 2 (Sonja Swanson)

B0. These guidelines are a welcomed addition to the literature and reflect the experience and common practices performed by some thoughtful and influential researchers in the MR community. I am especially encouraged to see that this is a living document that the researchers expect to update periodically. Generally, the guidelines reflect consensus on analytic practices with some appropriate caveats provided for the targeted audience. In this review, I highlight areas that are not discussed in great detail that I think warrant attention in the development of further guidelines.

> We thank the reviewer for her comments and positive view of these guidelines. <

B1. Well-defined causal effect estimands and causal null hypotheses: Part 1 carefully separates the goal of testing for a non-null effect and estimating a causal effect. Its consideration of testing, however, does not make clear what types of null hypotheses can really be assessed (see Swanson et al. 2018)¹, and the description on effect estimation only suggests qualitatively weighing whether the genetic variants act like the intervention of interest. More guidance could be provided on question framing, especially by giving more specifics on the types of causal questions MR is able to answer. See also Hernán & Robins (2006)² and Swanson et al. (2017)³.

> We appreciate this point, which we now introduce at the beginning of the manuscript, referencing the Swanson 2018 paper. We have added to the manuscript: "The straightforward statement of the causal hypothesis is that interventions in the exposure variable will affect the outcome. If the genetic associations with the exposure vary with time, then there are some nuances in terms of

what causal hypotheses can be tested [Swanson et al, 2018]; we discuss the impact of time-varying relationships between variables in Section 9.” (Section 1). <

> We also refer to this debate in Section 9 (Interpretation) when we discuss numerical estimates from Mendelian randomization and how these relate to clinically-meaningful causal parameters. In particular, we have added to the manuscript: “The estimate from a Mendelian randomization investigation is therefore better interpreted as a test statistic for a causal hypothesis rather than the estimate of a well-defined intervention.” (Section 9).

B2. The importance of time: Related to the framing of the research question is also how time is a part of this framing. MR estimates are often described as “lifetime effects” but the standard IV methods used in MR are developed for time-fixed exposures. (Note that this type of concern also applies to causal null hypothesis testing, as described in Swanson et al. 2018)¹. At very least, some qualitative guidance on how to interpret results in light of the time-varying nature of exposures and outcomes over the life-course is needed. See again Swanson et al. (2017)³ and also Labrecque & Swanson (2019)⁴.

> As per the response to point B1, we now reference the discussion about time-varying relationships between variables in Section 1 in relation to tests of the causal null hypothesis.

> We have edited the discussion in Section 9 to include issues relating to time-varying relationships between variables, referencing the papers cited above: “Mendelian randomization estimates relate specifically to changes in the exposure induced by the genetic variants used as instrumental variables. Genetic variants are present from before birth, and so Mendelian randomization investigations typically compare groups of the population having different trajectories in their distribution of the exposure over time. Analyses therefore typically can be interpreted as assessing the impact of long-term elevated levels of an exposure. However, in most cases, we have incomplete information about how the genetic variant changes the distribution of the exposure across the life course. If the genetic associations with the exposure vary over time, then Mendelian randomization estimates based on genetic associations with the exposure measured at a single timepoint can be unreliable. Similar difficulties of interpretation arise if the impact on the outcome relates to levels of the exposure at a specific time period in life. A plausible example of this is the effect of vitamin D on multiple sclerosis; multiple sclerosis risk is hypothesized to be influenced by vitamin D levels during early childhood, but not vitamin D levels in adulthood.” (Section 9). <

B3. Available robust methods, sensitivity analyses, and falsification strategies: These guidelines state that Sections 6 and 7 are not exhaustive considerations of these tools, but it would be helpful to the MR community if guidance on these points was expanded in future versions of these guidelines. It is not very clear why the authors chose to present the tools they did have room for, and not others. It also is not very clear why these sections focus so much on the risk of bias due to pleiotropy when other types of biases can also threaten MR estimates.

> As we have stated, we will revise these guidelines over time. However, balance is needed in this manuscript between comprehensiveness and comprehensibility. It is perhaps better for a comprehensive discussion of methods to be provided separately. Also, while we aim to update these guidelines regularly, we are unable to update advice with the regularity needed to offer up-to-the-minute recommendations on methods as they are developed and updated. We therefore focus on methods whose performance is generally understood. We have now added reference to a

more extensive review of methods: “This table is based on a broader review and comparison of methods [Slob, 2020].” (Section 6).

> We have now more clearly stated the motivation behind our choice of presentation: “We focus on these methods here as they can be implemented using summarized data alone, and they rely on different assumptions to provide consistent causal estimates.” (Section 6).

> In terms of focusing on pleiotropy, we were unclear in the initial submission. As described in Kang et al, JASA 2016 “Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization”, there is a statistical correspondence between pleiotropy and instrument validity, meaning that any instrument invalidity can be expressed algebraically in terms of pleiotropy. Hence, while we use the language of pleiotropy to provide mathematically precise statements of the assumptions needed for consistent estimates, this section does not only cover pleiotropy, but invalid instruments more generally.

> We have changed the section headings for Sections 6 and 7 to: “Robust methods for sensitivity analysis” and “Other approaches for sensitivity analysis” to make clear that these sections are not narrowly focused on pleiotropy, but cover other sources of instrument invalidity. We have also added the sentence: “Although robust methods typically use the term ‘pleiotropy’, there is a mathematical correspondence between instrument invalidity and pleiotropy [Kang et al, 2016], and so these methods can help assess sensitivity of findings to instrument invalidity more generally, and not simply invalidity that arises from horizontal pleiotropy. We here use the language of pleiotropy to make mathematically precise statements about the assumptions needed for methods to provide consistent estimates.” (Section 6).

> As per the response to point A1, we have added new paragraphs in Section 7 on alternative sources of bias. We hope this addresses the reviewer’s concerns. <

Competing Interests: None

Reviewer Report 24 December 2019

<https://doi.org/10.21956/wellcomeopenres.17034.r37168>

© 2019 Morrison J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Jean Morrison 

Department of Human Genetics, University of Chicago, Chicago, IL, USA

This article is a practical review of Mendelian randomization (MR) and set of guidelines intended for practitioners. The writing is clear and well organized and fairly thorough. I think this is an invaluable resource to investigators who wish to carry out an MR analysis and provides a good survey of relevant literature. My comments are mostly minor, I feel this article is much needed contribution to the field.

Major comments

- Reverse direction effects: To me it seemed to one major omission was discussion of reverse direction effects. These can easily cause false positives for methods like IVW and are especially an issue if one is using the "agnostic" variable selection method. I think it could be good to add a graph or two graphs to Figure 4 displaying a reverse effect (of outcome on exposure) and possibly also a feedback loop. This should be accompanied by a discussion of when reverse effects are something the investigator should think about and when they aren't. The sensitivity testing section should include a discussion of testing in the opposite direction if agnostic variable selection is used and if the pair of traits warrants that consideration. It would also be good to include a discussion of how to interpret a positive result in both directions.
- Collider bias when using summary statistics: When discussing individual level data approaches the authors give a set of recommended covariates and note that collider bias is a concern. A brief explanation of what a collider is and how it causes bias should be added. A parallel discussion should be added concerning summary statistic based analyses. In particular, it is important that investigators know which covariates were adjusted to compute the summary statistics and how to identify a potential collider.

Minor comments:

- Slightly more attention should be given to analyses described early as "exploratory" in which the investigator scans through many potential causal effects and how these should be treated differently. Relatedly, more attention could be given to analyses that use agnostic variable selection method. These issues are linked because a phenome wide MR analysis is likely to use the agnostic approach. In my view, for studies like these, a robust method (or multiple robust methods) should always be used, the investigator should assume that some variants are pleiotropic.
- There is more danger in using a one sample approach with an agnostic variant set due to weak instrument bias, which should be mentioned. Methods exist that estimate and correct for this bias by using correlation among test statistics for variants that aren't associated with either trait. CAUSE (Morrison *et al.* bioRxiv 2019)¹ is one but there must be other approaches to this issue as well.
- Both Egger regression and the modal estimator have much lower power than other methods. This is worth mentioning when discussing interpreting results from sensitivity analyses.
- It is worth mentioning in figure 4 that the "related variable" may not always be known.
- In the paragraph mentioning the "three sample" approach, it would be interesting to include any results about how much bias in the causal effect might be created by selection bias if one simply selects significant variants from the exposure GWAS.

References

1. Morrison J, Knoblauch N, Marcus J, Stephens M, et al.: Mendelian randomization accounting for horizontal and correlated pleiotropic effects using genome-wide summary statistics[preprint]. *Biorxiv*. 2019. [Publisher Full Text](#) | [Reference Source](#)

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.**Reviewer Expertise:** statistical genetics, Mendelian randomization**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 24 Dec 2019

Stephen Burgess, University of Cambridge, Cambridge, UK

Jean,

Thanks for the rapid and thoughtful response. We will wait for other reviewers to comment, and then update the article to address your points in due course.

Best wishes, Stephen

Competing Interests: No competing interests were disclosed.

Author Response 20 Apr 2020

Stephen Burgess, University of Cambridge, Cambridge, UK

> We thank the reviewers for their comments, which we have numbered for reference. We provide point-by-point responses to their suggestions below. <

Reviewer 1 (Jean Morrison):

A0. This article is a practical review of Mendelian randomization (MR) and set of guidelines intended for practitioners. The writing is clear and well organized and fairly thorough. I think this is an invaluable resource to investigators who wish to carry out an MR analysis and provides a good survey of relevant literature. My comments are mostly minor, I feel this article is much needed contribution to the field.

> We thank the reviewer for her comments and positive view of these guidelines. <

Major comments

A1. Reverse direction effects: To me it seemed to one major omission was discussion of reverse direction effects. These can easily cause false positives for methods like IVW and are especially an issue if one is using the "agnostic" variable selection method. I think it could be good to add a graph or two graphs to Figure 4 displaying a reverse effect (of outcome on exposure) and possibly also a feedback loop. This should be accompanied by a discussion of when reverse effects are something the investigator should think about and when they aren't. The sensitivity testing section should include a discussion of testing in the opposite direction if agnostic variable selection is used and if the pair of traits warrants that consideration. It would also be good to include a discussion of how to interpret a positive result in both directions.

> We have now added new paragraphs to Section 7 covering biases in estimation arising due to issues other than invalid instruments. These covers reverse causation (point A1) and collider bias (point A2).

> Relating to reverse causation, we have added to the manuscript: "While the genetic code is fixed at conception and so cannot be influenced by reverse causation, if the outcome influences the risk factor, this can result in gene—outcome associations becoming distorted and lead to misleading inferences. As shown in Figure 4e, if genetic variants that are supposed to be instrumental variables for the exposure in fact influence the outcome primarily, then genetic associations with the outcome could be present without the exposure influencing the outcome. The MR-Steiger method has been developed to detect such variants and remove them from the analysis. Bidirectional Mendelian randomization analyses have been proposed that use separate sets of instrumental variables for the exposure and outcome to assess the direction of causal effect." (Section 7).

> We have added a reverse causation scenario and a feedback loop scenario to Figure 4 as requested: "Another possible scenario that would lead to instrument invalidity is if genetic variants influence the outcome primarily rather than the exposure (Figure 4e, see also discussion on reverse causation in Section 7)." (Section 3). <

A2. Collider bias when using summary statistics: When discussing individual level data approaches the authors give a set of recommended covariates and note that collider bias is a concern. A brief explanation of what a collider is and how it causes bias should be added. A parallel discussion should be added concerning summary statistic based analyses. In particular, it is important that investigators know which covariates were adjusted to compute the summary statistics and how to identify a potential collider.

> As per the response to point A1, we have expanded the discussion about collider bias and selection bias in new paragraphs in Section 7 on bias due to issues other than invalid instruments.

> We have added to the manuscript: "A collider is a common effect of two variables – for example, any variable causally downstream of the exposure is influenced by the genetic variants and the exposure—outcome confounders, and so is a collider. Even if two variables are unrelated (they are marginally independent), they will typically be related when conditioning on the collider (conditionally dependent). Stratifying on or adjusting for a collider therefore leads to an association between variables that influence the collider. An association between the genetic variants and the exposure—outcome confounders would lead to biased causal estimates. Collider bias is not

unique to Mendelian randomization, but it is particularly relevant as some published genetic association estimates have been adjusted for potential colliders. Methods to account for collider bias have recently been proposed.

“Selection bias is a specific example of collider bias which occurs when selection into a study sample depends on a collider. Simulation studies have shown that selection bias can have a severe impact on Mendelian randomization estimates, but only when the associations of variables with the collider are quite strong. Selection bias can potentially be addressed using inverse-probability weighting, although this requires estimation of the probability of selection into the study sample for all individuals.” (Section 7).

> We also discuss the problem raised by the reviewer about pre-computed summarized data: “If published summarized association estimates have already been adjusted for a variable causally downstream of the exposure, collider bias (see Section 7) may be unavoidable.” (Section 2). <

Minor comments:

A3. Slightly more attention should be given to analyses described early as "exploratory" in which the investigator scans through many potential causal effects and how these should be treated differently. Relatedly, more attention could be given to analyses that use agnostic variable selection method. These issues are linked because a phenome wide MR analysis is likely to use the agnostic approach. In my view, for studies like these, a robust method (or multiple robust methods) should always be used, the investigator should assume that some variants are pleiotropic.

> In reference to phenome scan analyses, we have added: “Such analyses are generally regarded as exploratory or “hypothesis-generating”, and results are typically treated as provisional until replicated in an independent dataset.” (Section 1). In relation to analyses performed using an agnostic set of genetic variants, we have added: “If genetic variants are chosen in a way that is completely agnostic to the function of the variants, then researchers should be especially careful about the possibility of variants being pleiotropic.” (Section 3). <

A4. There is more danger in using a one sample approach with an agnostic variant set due to weak instrument bias, which should be mentioned. Methods exist that estimate and correct for this bias by using correlation among test statistics for variants that aren't associated with either trait. CAUSE (Morrison et al. bioRxiv 2019)¹ is one but there must be other approaches to this issue as well.

> We have added discussion on winner's curse and weak instrument bias when genetic variants are selected in the dataset under analysis (see point A7).

> We have added reference to the CAUSE method and another similar paper: “A further class of robust methods uses latent modelling to distinguish to what extent genetic associations with the outcome arise due to a causal effect of the exposure, as opposed to via direct (pleiotropic) or confounder-driven effects of particular variants. A causal model is evidenced if the predominance of variants that associate with the exposure also associate with the outcome in a proportional way. If the genetic associations with the outcome do not follow this pattern, then a non-causal explanation would be preferred. Emerging methods that take this approach include the Causal Analyses Using Summary Effect Estimates (CAUSE) [Morrison et al] and Latent Heritable Confounder MR (LHC-MR) [Darrous et al] methods.” (Section 6).

> We have also added reference to recently developed methods that are based on multivariable Mendelian randomization: “Methods have been proposed based on a multivariable approach in the context of gene expression data, including the MR-link [van der Graaf et al] and transcriptome-wide summary statistics-based Mendelian Randomization (TWMR) [Porcu et al] methods.” (Section 7). <

A5. Both Egger regression and the modal estimator have much lower power than other methods. This is worth mentioning when discussing interpreting results from sensitivity analyses.

> We have now mentioned this with respect to the MR-Egger method: “Estimates from the MR-Egger method are particularly affected by outlying and influential datapoints, and are prone to be imprecise, particularly when the variant—exposure associations are all similar in magnitude. This can lead to the method having low power to detect a causal effect.” (Section 6), and also with respect to the mode-based method: “The mode-based method has been shown to have low precision in some simulated and real datasets.” (Section 6). <

A6. It is worth mentioning in figure 4 that the “related variable” may not always be known.

> We now mention this in the figure caption: “We note that the related variable may be known or unknown.”. <

A7. In the paragraph mentioning the “three sample” approach, it would be interesting to include any results about how much bias in the causal effect might be created by selection bias if one simply selects significant variants from the exposure GWAS.

> Bias due to selecting variants based on their statistical significance in the dataset under analysis comes under the broader category of winner’s curse. Selecting variants based on their association with the exposure in the dataset under analysis can lead to exacerbation of weak instrument bias. We have expanded the discussion on winner’s curse in Section 3: “Often, selection is based on the dataset in which genetic associations with the exposure are estimated. However, this leads to “winner’s curse” –genetic associations tend to be overestimated in the dataset in which they were first discovered. If genetic variants are selected based on their associations with the exposure in the dataset under analysis, weak instrument bias is exacerbated (in the direction of the observational association in a one-sample setting, and in the direction of the null in a two-sample setting).” (Section 3). <

Competing Interests: None