Guidelines for performing Mendelian randomization investigations [version 1; peer review: awaiting peer review]

Stephen Burgess¹,², George Davey Smith³,⁴, Neil M. Davies³,⁴, Frank Dudbridge⁵, Dipender Gill⁶, M. Maria Glymour⁷, Fernando P. Hartwig³,⁸, Michael V. Holmes⁹,¹⁰, Cosetta Minelli¹¹, Caroline L. Relton³,⁴, Evropi Theodoratou¹²,¹³

¹MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
²Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK
³MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
⁴Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
⁵Department of Health Sciences, University of Leicester, Leicester, UK
⁶Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
⁷Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA
⁸Postgraduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil
⁹MRC Population Health Research Unit at the University of Oxford, Nuffield Department of Population Health, University of Oxford, Oxford, UK
¹⁰Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
¹¹National Heart and Lung Institute, Imperial College London, London, UK
¹²Centre for Global Health, Usher Institute, University of Edinburgh, Edinburgh, UK
¹³Edinburgh Cancer Research Centre, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

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 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
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: © 2019 Burgess S et al. This is an open access article 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.

How to cite this article: Burgess S, Davey Smith G, Davies NM et al. Guidelines for performing Mendelian randomization investigations [version 1; peer review: awaiting peer review] Wellcome Open Research 2019, 4:186 (https://doi.org/10.12688/wellcomeopenres.15555.1)

First published: 26 Nov 2019, 4:186 (https://doi.org/10.12688/wellcomeopenres.15555.1)
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 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 confounding5–7. 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 effect5.

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 unnecessary7,8. For example, Mendelian randomization has demonstrated a causal effect of time spent in formative education on Alzheimer’s disease9. 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 exposure9, 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 population10. 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 regression11 or the latent causal variable method12) 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 considered13,14. Such analyses are generally regarded as exploratory or “hypothesis-generating”.

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 data15. 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.
Figure 1. Flowchart highlighting some of the key analytic choices in performing a Mendelian randomization (MR) analysis.

1. If there are genetic variants having biological relevance to the exposure, then consider performing the MR analysis using these variants only, and perform appropriate sensitivity analyses.

2. If such variants are not available, consider initially performing a “liberal” MR analysis using a less stringent choice of variants. If the estimate is null, then there is little evidence for a causal effect.

3. If the estimate from the initial analysis is non-null, then assess the robustness of the finding using different approaches: stricter criteria for variant selection, leave-one-out analyses, robust methods, positive/negative controls, subgroup analyses, colocalization (for analyses based on single gene region).

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? Summary 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, R²/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.

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

Each of the methods in the table can be implemented using summarized data. False positive rates refer to the simulation study by Slob and Burgess\(^2^2\). 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](https://cran.r-project.org/web/packages/MendelianRandomization/index.html))

† Implemented in mrobust package for Stata ([https://github.com/qingyuanzhao/mr.raps](https://github.com/qingyuanzhao/mr.raps))

‡ Implemented for R in its own software package:
- MR-PRESSO in mrpresso package ([https://github.com/remlapmot/mrrobus](https://github.com/remlapmot/mrrobus))
- MR-RAPS in mr.raps package ([https://github.com/qingyuanzhao/mr.raps](https://github.com/qingyuanzhao/mr.raps))
- MR-Mix in MRMix package ([https://github.com/gqi/MRMix](https://github.com/gqi/MRMix)).

### 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 dataset\(^2^3\). 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\(^2^4,2^5\). Such differences can affect not only the interpretation of causal estimates, but also the validity of causal inferences\(^2^6\). 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\textsuperscript{32}. Several large consortia have made such estimates publicly available for hundreds of thousands of variants\textsuperscript{33,34}. 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. Individual data in a one-sample setting are required to investigate non-linear effects\textsuperscript{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 instruments\textsuperscript{36}. 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 1 error) rates, particularly when more than one variant is included in the analysis\textsuperscript{37}. 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 outcome\textsuperscript{38}. 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 setting\textsuperscript{39}.

### 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 analysis\textsuperscript{40}. 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 protein\textsuperscript{41}. 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 exposure\textsuperscript{42} – 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 variants\textsuperscript{43}. 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, Mendelian randomization investigations for vitamin D typically use variants from four gene regions that are biologically implicated in the synthesis or metabolism of vitamin D\textsuperscript{44}. However, caution is required as biological understanding is rarely infallible. As an example, although genetic variants in the 1L6R 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 measurements\textsuperscript{45}.

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, this is done on the basis of the dataset in which genetic associations with the exposure are estimated. However, this could be a different dataset entirely: in a “three-sample” analysis, variants are identified in one dataset, and the genetic associations with the exposure and outcome are estimated in separate datasets\textsuperscript{46}. This avoids bias from “winner’s curse” – genetic associations are typically overestimated in the dataset in which they were first discovered.

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 outcome. 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 exposure (assuming the other instrumental variable assumptions are satisfied), even if it is associated with another variable. 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.

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

![Figure 4](Image)

**Figure 4.** Directed acyclic graphs illustrating validity and invalidity of instrumental variable assumptions in different scenarios.

- **a.** Genetic variants influence the exposure, which has downstream effects on a related variable which affects the outcome.
- **b.** Mediator is on causal pathway from exposure to outcome.
- **c.** Genetic variants influence the exposure, which has downstream effects 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 exposure, which has effects on the outcome directly and indirectly via a related variable.
- **f.** Genetic variants influence a related variable, which influences the outcome directly and indirectly via the exposure.

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. In scenario e, the instrumental variable assumptions are satisfied as all causal pathways from the genetic variants to the outcome pass via the exposure. The standard causal estimate represents the total effect of the exposure, including the direct and indirect pathways. In scenario f, the instrumental variable assumptions are not satisfied, as there is a causal pathway from the genetic variants to the outcome that does not pass via the exposure.
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 variants. 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 allele. 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 manuscripts.

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 covariates. 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 bias.

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 regression, although in this case estimates are sensitive to correct specification of the first-stage regression model and other approaches that make weaker distributional assumptions, such as structural mean models, may be preferred.

The 2SLS method can be applied to the two-sample setting if individual-level data are available for both samples. 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) method. With uncorrelated variants, the IVW estimate can be obtained from an IVW meta-analysis of the ratio estimates for the individual variants. 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 sample. A modification of this method has been proposed to allow for correlation (linkage disequilibrium) between variants.

For continuous outcomes, the IVW estimate is asymptotically equivalent to the 2SLS estimate obtained from individual level data. The 2SLS method (and thus also the IVW method) is the most efficient estimate of the causal effect when all genetic variants are valid instruments.

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 alone. 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 mechanisms. 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).

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 suggest 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 estimates\(^2\). 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 negative\(^7\).

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 strategy\(^8\). 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 exposure\(^9\). 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)\(^10\), and body composition measures (such as fat mass and fat-free mass)\(^11\). 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 estimated\(^12\). Both the 2SLS and IVW methods can be adapted to the multivariable setting\(^12\). 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 to address pleiotropy

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\(^10\); investigators should consider using multiple methods that make different assumptions about the nature of the underlying pleiotropy\(^27\). 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. 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. 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)\(^29\). A multivariable version of the MR-Egger method is available\(^30\). Estimates from the MR-Egger method are particularly affected by outlying and influential datapoints\(^31\), and are prone to be imprecise, particularly when the variant—exposure associations are all similar in magnitude. A heterogeneity measure has been proposed to quantify the similarity between variant—exposure associations and the potential impact on MR-Egger analyses\(^32\). 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 data\(^33\).

The median- and mode-based methods\(^30\) 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. Other methods have been proposed that make the same plurality valid assumption as the mode-based method, including the contamination mixture method\(^32\) and MR-Mix\(^35\).
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 variants\textsuperscript{27}. 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 method\textsuperscript{27} and MR-Mix\textsuperscript{40} 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 I\textsuperscript{2} statistic, is reported as a part of a polygenic Mendelian randomization investigation\textsuperscript{39,50,71}. 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 variant\textsuperscript{71}. 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 time)\textsuperscript{71} 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 exposure\textsuperscript{71}. 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 identified\textsuperscript{91}.

7. Other approaches for sensitivity analysis to address pleiotropy

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 acid\textsuperscript{96}. 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 asthma\textsuperscript{77}. 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\textsuperscript{74,79}. 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 disequilibrium\textsuperscript{86}. 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 disease\textsuperscript{81}. 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 region\textsuperscript{82}. 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\textsuperscript{83}. 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 drinkers\textsuperscript{84}. 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\textsuperscript{85}. However, if the subgroup is defined by a collider, then stratification can introduce bias to the analysis\textsuperscript{86}. A collider is a common effect of two variables – for example the exposure is influenced by the genetic variants and the exposure—outcome confounders. Stratifying on a collider 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\textsuperscript{87}.

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 model\textsuperscript{88}. This analysis will estimate the direct effect of the exposure on the outcome keeping these variables constant.

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 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 not\textsuperscript{89}. 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 estimates\textsuperscript{90}. Plots allow the investigators

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{scatter_plot.png}
\caption{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 \textit{et al}., 2018\textsuperscript{68}.}
\end{figure}
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 R\(^2\) 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 bias\(^{38}\). 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://csgenomics.com/shiny/mRnd/](http://csgenomics.com/shiny/mRnd/) and [https://sb452.shinyapps.io/power/](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\(^ {34}\). In line with the STROBE-MR guidelines\(^ 1\), 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.

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 is often reasonable\(^ {37}\). 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\(^ {39,34}\). 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 exposure\(^ {35}\). 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 trials\(^ {31}\). 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 practice\(^ {36}\). 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 robustness\(^ {37}\). 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 straightforward\(^ {17}\). The difficulty is not in performing a Mendelian randomization analysis, but rather in performing a credible analysis\(^ {38}\) and providing a reasoned interpretation\(^ {39}\). 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.