OXFORD

Miscellaneous

Investigating the transparency of reporting in two-sample summary data Mendelian randomization studies using the MR-Base platform

Benjamin Woolf [iD] ,[1,2]* Nina Di Cara,[2,3] Christopher Moreno-Stokoe,[1,2]
Veronika Skrivankova,[4] Katie Drax,[1,2] Julian PT Higgins [iD] ,[2,3,5]
Gibran Hemani,[2,3] Marcus R Munafò,[1,2] George Davey Smith [iD] ,[2,3,5]
James Yarmolinsky, [2,3†] and Rebecca C Richmond [iD] [2,3†]

[1]School of Psychological Science, University of Bristol, Bristol, UK, [2]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK, [3]Population Health Sciences, University of Bristol, Bristol, UK, [4]Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland and [5]NIHR Bristol Biomedical Research Centre, Bristol, UK

*Corresponding author. School of Psychological Science, University of Bristol, 12a, Priory Road, Bristol BS8 1TU, UK.
E-mail: benjamin.woolf@bristol.ac.uk
†J.Y. and R.R. contributed equally to this work.

Abstract

Background: Two-sample Mendelian randomization (2SMR) is an increasingly popular epidemiological method that uses genetic variants as instruments for making causal inferences. Clear reporting of methods employed in such studies is important for evaluating their underlying quality. However, the quality of methodological reporting of 2SMR studies is currently unclear. We aimed to assess the reporting quality of studies that used MR-Base, one of the most popular platforms for implementing 2SMR analysis.

Methods: We created a bespoke reporting checklist to evaluate reporting quality of 2SMR studies. We then searched Web of Science Core Collection, PsycInfo, MEDLINE, EMBASE and Google Scholar citations of the MR-Base descriptor paper to identify published MR studies that used MR-Base for any component of the MR analysis. Study screening and data extraction were performed by at least two independent reviewers.

Results: In the primary analysis, 87 studies were included. Reporting quality was generally poor across studies, with a mean of 53% (SD = 14%) of items reported in each study. Many items required for evaluating the validity of key assumptions made in MR were poorly reported: only 44% of studies provided sufficient details for assessing if the genetic variant associates with the exposure ('relevance' assumption), 31% for assessing if there are any variant-outcome confounders ('independence' assumption), 89% for the

assessing if the variant causes the outcome independently of the exposure ('exclusion restriction' assumption) and 32% for assumptions of falsification tests. We did not find evidence of a change in reporting quality over time or a difference in reporting quality between studies that used MR-Base and a random sample of MR studies that did not use this platform.

Conclusions: The quality of reporting of two-sample Mendelian randomization studies in our sample was generally poor. Journals and researchers should consider using the STROBE-MR guidelines to improve reporting quality.

Key words: Mendelian randomization, meta-epidemiology, reproducibility

---

Key Messages

- Good quality reporting is important for evaluating the underlying quality of a study.
- This review found that two-sample Mendelian randomization studies were poorly reported.
- Authors of new Mendelian randomization studies should consider using the STROBE-MR guidelines when writing up their research.

Introduction

Mendelian randomization (MR) is an epidemiological approach to causal inference which uses germline genetic variants strongly associated with exposures of interest to appraise the effect of those exposures on one or more outcomes.[1] In a traditional 'one-sample' MR design, information on exposure, outcom, and genetic instruments are collected in a single sample, and the effect of the exposure on the outcome estimated using only these data. Two-sample Mendelian randomization (2SMR) instead uses two studies with information on the association between the genetic instrument(s) and the exposure (Sample 1) and the genetic instrument(s) and the outcome (Sample 2), under the assumption that both samples are representative of the same underlying population, to calculate an effect estimate.[2] Because 2SMR only requires information on each genotype-phenotype association, summary statistics from two genome-wide association studies (GWASs) often provide sufficient information to implement the analysis. The increase in publicly accessible summary statistics from GWAS has vastly facilitated the application of 2SMR. Since the summary statistics generally come from previously published data, there is no requirement to apply for and access individual-level data or to perform data cleaning. This in turn makes performing MR analysis more rapid. The emergence of platforms automating the curation of GWAS summary statistics, and statistical packages for performing 2SMR (e.g. MR-Base/TwoSampleMR,[3] SMR[4] and MendelianRandomization[5]) has also facilitated the growth in studies employing this method, by reducing the complexity of implementing 2SMR analysis.[6]

IEU OpenGWAS [https://gwas.mrcieu.ac.uk/] is a curated, open source, GWAS summary statistic repository managed by the University of Bristol, with an accompanying R package (TwoSampleMR) and web platform (MR-Base) for performing 2SMR analysis.[3] At the start of 2021, the repository contained nearly 40 000 GWASs across many categories of traits including biomarkers, clinical conditions and behavioural traits. The R package and web platform, hereafter collectively referred to as MR-Base, can be used to perform many stages of a 2SMR analysis, including harmonization of data across datasets and commonly used sensitivity analyses, using data available from a linked data repository or uploaded by the user. The integration of large GWAS repositories and an easy-to-implement analysis package makes MR-Base one of the most popular tools for performing two-sample MR analysis. For example, at the time of writing in June 2021, the descriptor paper for MR-Base[3] and associated R-package had between two and three times more citations in Google Scholar than the citations for an alternative package's description paper published at a similar time,[7] and for the 12 months prior to June 2021 MR-Base received around 95 million API (Application Programming Interface) requests (personal communication with Tom Gaunt, June 2021).

MR-Base increases the breadth and speed at which 2SMR analyses can be performed, but it might also lead to discrepancies in their design, conduct and reporting. The accessibility of such resources may permit analyses to be performed without careful consideration of the analytical choices that are being made or the assumptions inherent in

the approach. Further, the abundance of genetic instruments, datasets and methods available for use might influence the robustness of study results.[6,8] For example, easily accessible GWAS data of uncertain quality may be used in an analysis without the ability to examine or correct for bias in the underlying dataset. Even if the methodological design of the underlying GWAS used in a 2SMR analysis is robust, the ability to rapidly extract summary statistics directly from a data repository, rather than from the original GWAS itself, may encourage insufficient assessment and reporting of the methods used in the GWAS (e.g. with regard to selection or phenotyping procedures). These factors may encourage the generation of spurious and/or non-reproducible results which could present a possible threat to the robustness and reliability of the literature. Automation of MR analysis permits rapid assessment of numerous potentially causal relationships which may also encourage 'data fishing' (e.g. examining various hypotheses and reporting only 'positive findings') or selectively 'cherry-picking' results from sensitivity analyses.

Given concern that such platforms may facilitate poor quality or poorly reported research,[9] there is therefore a need to systematically appraise the quality of reporting in 2SMR studies in order to make an assessment of the potential quality and rigour of the analysis performed. Assessing the transparency of reported studies is also a necessary part of assessing their risk of bias and is a requirement for good attempts at replication. Unless the methods and assumptions of a study are explicitly stated, there will likely be a large number of 'researcher degrees of freedom' leading to variability and inconsistency in findings.[10] A final benefit of systematically assessing transparency is that doing so requires the creation of an explicit checklist for the appraisal of the reporting quality in MR studies.

This project aimed to appraise the transparency in reporting of two-sample Mendelian randomization analyses that used MR-Base. We therefore examined the reporting quality of 2SMR analyses that have been performed using either the MR-Base statistical package (TwoSampleMR in R) or the online web platform, using a checklist developed specifically for this study.

Methods

Development of the reporting quality checklist

We developed a bespoke checklist to examine the quality of reporting of 2SMR studies because none of the existing MR-reporting reviews or guidelines[1,11–14] were specifically tailored to 2SMR, and therefore did not include items on 2SMR specific assumptions and requirements, such as data harmonization. To inform the checklist, B.W. reviewed papers describing assumptions and sources of bias in MR analyses in addition to papers presenting reporting checklists for MR and/or instrumental variable studies. B.W. produced a first draft of the checklist which was then amended iteratively based on feedback from J.Y. and R.R., and piloted on three 2SMR studies. The development of the checklist was also informed by discussions about the STROBE-MR checklist, which was being developed at the same time by many of this study's authors.

Eligibility criteria

The study aimed to assess the quality of reporting of 2SMR studies published in peer-reviewed academic journals. Therefore, any published study that conducted MR and used the MR-Base R package (TwoSampleMR) or online platform [http://www.mrbase.org/] during any component of the MR analysis was eligible for inclusion. Studies were included irrespective of the type of study participants, setting, exposure(s) or outcome(s) being investigated.

Identification and selection of 2SMR studies

We searched Web of Science Core Collection, PsycInfo, MEDLINE and EMBASE from January 2016, because MR-Base was developed in 2016. The last search date was April 2019 and no language or other constraints were applied. The search terms are provided in the Supplementary Methods (available as Supplementary data at *IJE* online).

Studies were also identified by performing a citation search using Google Scholar for citations of the R package: the eLife article,[3] the correctly cited bioRxiv preprint,[15] incorrectly cited bioRxiv preprint version 1,[16] incorrectly cited bioRxiv preprint version 2,[16] the LD-hub and MR-Base presentation at the 2016 Annual Meeting of the Behaviour-Genetics-Association[17] and references of 'www.mrbase.org' in Google Scholar.

Citations retrieved by the search were uploaded onto Rayyan [https://rayyan.qcri.org],[18] a website specifically designed for paper screening in systematic reviews. Rayyan automatically identifies duplicates of citation/abstracts, which were then manually checked for errors. Two reviewers (B.W., N.D.) screened the abstracts and titles for relevance using the eligibility criteria. Studies identified as potentially relevant had their full text screened. When both reviewers agreed that a study met the eligibility criteria, it was included in the review. Initial disagreements were discussed between N.D. and B.W., with unresolved items arbitrated by a third researcher (J.Y. or R.R.).

Data collection process

We developed a standardized data extraction form in advance of data collection. The form required information on each paper's author(s), date of publication, title and quality of reporting across all items included in the reporting quality checklist. Each item in the checklist was graded as having been reported (1), not having been reported (0) or not being applicable to the study (NA), based on the reviewer's opinion of the detail of reporting. Information to support each of these assessments was also collected (e.g. as quotations from the papers). The form was pilot-tested on three 2SMR papers that did not use MR-Base, to ensure that the same studies would not be included in the final review.

Data extraction and grading for each study were performed by two independent reviewers (from among five reviewers: B.W., N.D., C.M.S., J.Y. and R.R.) to minimize transcription errors. The data extraction forms were then combined and checked for errors in extraction and disagreements. Disagreements were arbitrated by the reviewers rereading the paper and coming to a joint conclusion.

To check that unique studies were only included once, studies that shared at least one author were compared based on similarity of study population, date and methodology. Duplicate studies were treated as a single study in the analysis. Because this review aimed to assess the quality of reporting of 2SMR studies in the published literature, no attempt was made to contact study authors for further information.

Analysis

Defining a 'study' as an individual publication, we calculated (i) the percentage of studies reporting each individual item in the checklist; (ii) the percentages of studies reporting at least 25%, 33%, 50%, 67%, 75% and 100% of all the items in the checklist; and (iii) the overall mean percentage of items reported across all studies. Percentages excluded studies rated as not applicable for any specific item.

Since studies analysing multiple phenotypes (e.g. phenome-wide scans) may not report all exposure-outcome associations with equal detail, we also evaluated studies based on whether they were a 'multi-phenotype study' or not, defining 'multi-phenotype' as a study with ≥ 1 exposure and ≥ 10 outcomes or a study with ≥ 10 exposures and ≥ 1 outcome. For brevity, 'multi-phenotype' studies are referred to as '≥ 10 phenotype' studies and non-multi-phenotype studies are referred to as '<10 phenotype studies'. For multi-phenotype studies, the items in the GWAS section of the checklist were calculated as the mean

reporting quality across all included GWASs used in the MR analysis.

We undertook five supplementary analyses:

i. We compared changes in the level of reporting (mean percentage of all items reported) across studies by the year of publication.

ii. B.W. and G.H. classified each item as having information that would be available to all users (i.e. the information required to report the item is provided by the output of both the R package and web platform), information that would be available to some users (i.e. the information required to report the item is available to users of the R package or the web platform, but not both), information that is not given by MR-Base (i.e. the information required to report the item is not available to users of either the R package or the web platform), or as not applicable to the MR-Base analysis (e.g. for items defining the study and/or articulating the question) (Table 1). We then compared the level of reporting according this classification.

iii. We compared the level of reporting according to method of citation, classified as explicitly citing the R-package or not citing the R package.

iv. While drafting the STROBE-MR extension,[19] V.S., R.R., B.W. and J.Y. examined a random sample of MR papers using the original STROBE checklist: A description of the methods used in this review can be found in the Supplementary Methods (available as Supplementary data at *IJE* online). To explore the generalizability of our finds to non-MR-Base papers, we compared the level of reporting between MR-Base and non-MR Base papers for those items that could be harmonized across the reviews. Because this review is a random sample of MR-studies, 95% CIs were calculated to quantify uncertainty due to random sampling error for these estimates.

v. We attempted to estimate the popularity of MR-Base by comparing the number of citations included in our search with the number of research articles included in a search for two-sample MR studies in PubMed between 1 January 2016 and 1 January 2019. A more detailed description of this search can be found in the Supplementary Methods.

Registration

This study, including the first draft of the checklist, was pre-registered on the Open Science Framework: doi 10.17605/OSF.IO/NFM27.We made several modifications to the protocol, with each change made prospectively in light of piloting. The changes to the protocol are described in the Supplementary Methods.

Table 1 Reporting checklist

Question	Criteria	Available to all users (Y), partial coverage (P), not provided by MR-Base (N), item not analysis related (NA)
Clear articulation of research question		
1) Clearly define exposure(s)	State clearly what the exposure is	NA
2) Clearly define outcome(s)	Clearly state what the outcome is	NA
3) State how many exposure/outcome relationships were tested in the main analysis	State how many exposure/outcome relationships were tested	NA
4) Clearly state hypothesis under investigation	State the hypothesis under investigation	NA
Data sources		
5) Provide an evidence trail to a description of the source GWAS[a]	e.g. a citation or link to relevant study or methods paper.	Y
6) Describe the methods used to recruit participants into GWAS[a]	As in question	N
7) State the number of participants included in GWAS[a]	As in question	Y
8) Describe how exposure and outcome were measured[a]	As in question	N
9) State the units that the exposure and outcome were measured in and whether the data for these measures were transformed (or if the study is a case/control)[a]	State the units of the GWAS or that it was a case/control	P
10) List covariates included in GWAS	As in question	N
11) Describe quality control procedures adopted in GWAS or explicitly provide a reference to it[a]	Provide some information on the GWAS QC.	N
12) If applicable, comment on any attempts made to address inadequate quality control or GWAS design in MR analysis[a]	If they describe the QC, do they describe any attempts to improve it	N
13) For MR studies examining where exposure and/or outcome is a disease/binary trait, describe if cases were incident or prevalent or both	As in question	P
2SMR specific assumptions		
14) Provide adequate descriptive information on GWAS samples to assess whether they represent the same underlying population	e.g. at least two of age, sex and ancestry being similar, or dissimilarity on one of the domains	P
15) Provide information on sample overlap		N
Data harmonization		
16) Explain how reference alleles were harmonized across exposure and outcome datasets		Y
17) Describe how palindromic SNPs were addressed		Y
Instrument construction		
18) Provide clear criteria or a clear description for the inclusion of genetic variants as instruments	e.g. a biological justification or GWAS *P*-value threshold	Y
19) Describe whether genetic variant-exposure and genetic variant-outcome association estimates were obtained from a discovery GWAS sample, replication GWAS sample, pooled discovery-replication GWAS samples or a meta-analysis of two or more samples, or another source		N
20) Describe whether the instrument was restricted to independent variants or whether the instrument consisted of correlated variants		Y
21) If the instrument consisted of independent variants, how was independence defined/what were the clumping parameters used?	e.g. the LD r^2, base distance	Y
22) If the instrument consisted of correlated variants, explain how this was accounted for in the model employed		P

(Continued)

Table 1 Continued

Question	Criteria	Available to all users (Y), partial coverage (P), not provided by MR-Base (N), item not analysis related (NA)
23) State the number of primary instruments constructed	e.g. the number of PGRs or distinct SNP aggregation.	Y
24) State the number of included genetic variants.	e.g. the number of SNPs	Y
Instrumental variable (IV) assumptions and considerations		
25) Describe how IV assumption 1 (relevance) was assessed	i.e. provide an r^2 or F statistic for each SNP or the PGRS	P
26) Describe how IV assumption 2 (independence) was assessed	i.e. do they describe methods for ensuring that instruments are not associated with confounders, like using BOLT LMM, adjusting for ancestry/PCs, within family GWASs, etc	N
27) Describe how IV assumption 3 (exclusion restriction) was assessed	e.g. stating the use of a falsification test	Y
28) Describe how homogeneity/monotonicity/constant effect (IV assumption 4) was assessed		N
29) If applicable, acknowledge all major assumptions introduced by falsification tests	e.g. the INSIDE assumption for MR-Egger	N
30) Explain why the used sensitivity analyses were used	e.g. because they allow the relaxation of an assumption	P
Analytic methods		
31) List the primary model(s) employed to examine the exposure-outcome association	e.g. inverse variance weighted model or a Wald ratio	Y
32) Describe whether any proxy genetic variants were used in the analysis, and if so which reference panels and LD threshold criteria were used for selecting these proxy genetic variants, and how these were derived (e.g. SNP SNAP, LD Link, MR-Base, manual look up, etc)	NB: if no description was provided, then papers were marked as providing insufficient information. This is because the default in MR-Base web platform is to try to use proxies, but not state if they were used. Without any information provided it is impossible for a reader to know if the paper used the default settings or chose not to use proxies	Y
33) Describe whether a power calculation has been performed.		N
34) If there was >1 exposure/outcome relationships, either state whether this was corrected for multiple testing or provide justification for/against correcting for multiple testing (or discuss in discussion)		N
35) Describe any attempts to examine directionality of genetic variant-exposure and genetic variant-outcome associations	e.g. the use of a bidirectional design or Steiger filtering.	P
36) If analyses were performed using a single genetic variant as an instrumental variable, was co-localisation performed		P
37) If the exposure GWAS used a binary variable, is the causal effect described in terms of liability (or susceptibility) to the exposure in the discussion		NA
38) Describe the causal effect in terms of the units of measurement, or relate it to a clinically understandable scale of the exposure	NB: for measures like psychometric instruments which do not have 'units', a description of the effect in terms of a standard deviation is sufficient. Other standardized measures require the standard deviation to be defined in terms of units	P

(Continued)

Table 1 Continued

Question	Criteria	Available to all users (Y), partial coverage (P), not provided by MR-Base (N), item not analysis related (NA)
39) Describe whether any plots are presented to visualize results	NB: only plots used to visualize MR results are eligible	Y
40) Describe whether other forms of MR specific bias could be present in analyses	e.g. dynastic effects, assortative mating, canalization. NB:. if pleiotropy or another bias is covered elsewhere in the text, only residual pleiotropy/bias would be eligible	NA
Reproducibility and open science		
41) Present all data used to perform all analyses or describe where data can be accessed	e.g. if using data from another study, provide citations; if using primary data provide link to data access	Y
42) Provide R code for performing all analyses		Y

MR, Mendelian randomization; LD, linkage disequilibrium; GWAS, genome-wide association study; QC, quality control; PGRs, polygenic risk score; SNP, single nucleotide polymorphism; PCs, principal components (of the genetic relationship matrix); IV, instrumental variable.

ªGive information separately for exposure and outcome GWAS.

Results

Checklist

The items in the final checklist are presented in Table 1. The checklist contains 42 items across eight domains. A glossary of MR technical terms in the checklist, adapted from the MR Dictionary,[20] is presented in Supplementary Table S1 (available as Supplementary data at *IJE* online).

The 'Clear articulation of research question' domain has questions assessing whether authors clearly define the exposure(s), outcome(s) and number of hypothesises tested. 'Data sources' examines reporting quality of the underlying GWASs used by each study, including items relating to GWAS methods, such as quality control (QC), covariates included in models, measurement of traits and recruitment of participants into studies. '2SMR specific assumptions' asks about reporting if the GWASs are representative of the same underlying population, and any quantification of the amount of sample overlap between GWASs.

'Data harmonization' consists of two questions pertaining to data harmonization, including how alleles were harmonized and how palindromic single nucleotide polymorphisms (SNPs) were addressed. 'Instrument construction' has various items related to how the instrument was constructed, which includes questions on whether there was a clear description of how genetic variants were chosen, if the genetic variants were independent, how weights were estimated, the number of variants included in instruments and the number of instruments used for each exposure. 'Instrumental Variable (IV) assumptions and considerations' has questions about the instrumental variables assumptions, and asks how the three core assumptions were assessed, if additional sensitivity analyses were used and whether any additional assumptions had been acknowledged.

The 'Analytic methods' domain has miscellaneous questions about the reporting of other analytical methods, such as a definition of the primary model, power calculations and the use of plots to visualize findings. 'Reproducibility and open science' has two questions about 'open science': whether the data analysed have been provided or an explanation of where they can be accessed, and whether the R code used in the analysis has been presented.

Included studies

The search resulted in 876 citations, including 505 unique studies. After full-text screening, 87 studies were identified for inclusion in the review. Study selection is illustrated in Figure 1. Studies excluded in the full-text screen and their reason for exclusion are presented in Supplementary Table S3 (available as Supplementary data at *IJE* online). A list of studies included and excluded can be found Supplementary Table S4 and Supplementary Table S5 (available as Supplementary data at *IJE* online), respectively.

Details of the data extracted and reviewers' judgements on reporting quality are provided in Supplementary Table S6 (available as Supplementary data at *IJE* online); 14 (16%) of the 87 studies, listed in Supplementary Table S7 (available as Supplementary data at *IJE* online), were classified as ≥10 phenotype studies.
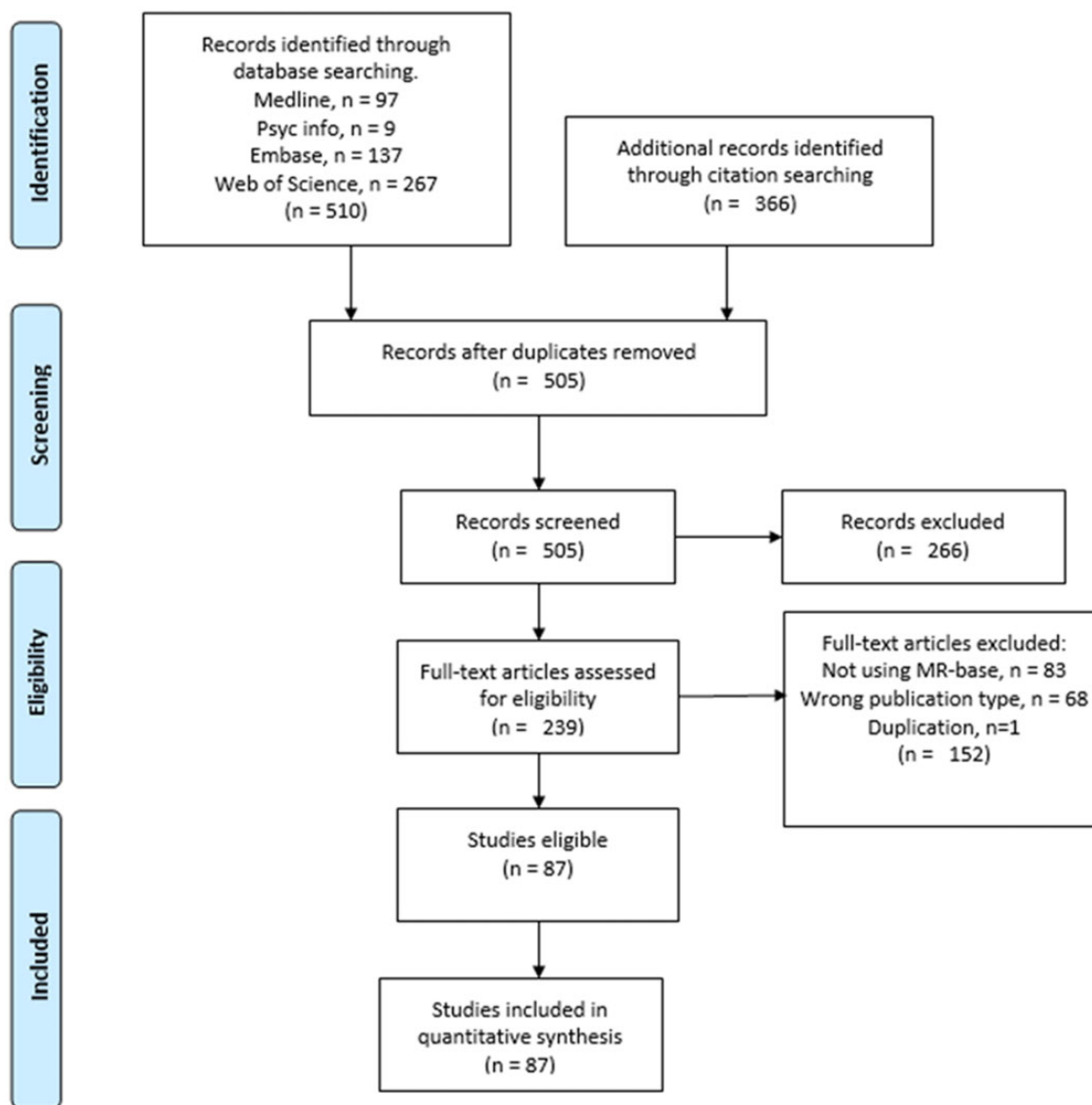
Figure 1 PRISMA Flow chart for study inclusion

Quality of reporting

Table 2 presents the percentage of items reported per question, which ranged from 1% (item 28: homogeneity/monotonicity/constant effect assumption) to 99% (item 31: list primary model[s]). Supplementary Figure S1 (available as Supplementary data at *IJE* online) shows the number of studies that reported certain minimum thresholds for the percentage of items reported; 53 studies (61%) reported at least 50% of items, five (6%) reported at least 75% of items and none reported 100% of items. The overall mean percentage of items reported in each study was 52% (SD = 14%).

All items in the 'Clear articulation of research question' domain were well reported, with each item being reported in at least 90% of studies. In the 'Data sources' domain,

the items requiring an evidence trail, sample size, and units were well reported (98%, 85% and 71%, respectively). In addition, around 81% of studies that stated a potential issue with the GWAS QC described a correction to address this issue. However, the other questions on study design details were less well reported, with 15–38% of studies reporting these items.

Neither question in the '2SMR specific assumption' domain (on sample overlap and consistency of populations) was answered adequately, both being reported in fewer than 40% of studies. Likewise, the two questions in the 'Data harmonization' domain were reported poorly, with fewer than 30% of studies providing information. All items in the 'Instrument construction' domain were reported in most studies, with these items being reported at

Table 2 Percentage of studies reporting each item, by item

Question	Total % items reported ($n = 87^{a}$)	% in studies with ≥10 phenotypes ($n = 14^{a}$)	% in studies with <10 phenotypes ($n = 73^{a}$)
Clear articulation of research question			
1) Clearly define exposure(s)	91.95	64.29	97.26
2) Clearly define outcome(s)	90.80	71.43	94.52
3) State how many exposure/outcome relationships were tested in the main analysis	91.95	85.71	93.15
4) Clearly state hypothesis under investigation	89.66	78.57	91.78
Data sources			
5) Provide an evidence trail to a description of the source GWAS[b]	97.92	100.00	97.52
6) Describe the methods used to recruit participants into GWAS[b]	16.15	13.88	16.56
7) State the number of participants included in GWAS[b]	84.58	92.67	83.02
8) Describe how exposure and outcome were measured[b]	37.62	40.47	37.07
9) State the units of measurement[b]	70.53	63.68	71.85
10) List covariates included in GWAS[b]	27.55	26.38	27.77
11) Describe quality control procedures adopted in GWAS or explicitly provide a reference to them[b]	24.67	29.42	23.76
12) If applicable, comment on any attempts made to address inadequate quality control or GWAS design in MR analysis[b]	81.25 (of GWASs in 4 studies[b])	62.50 (of GWASs in 2 studies[b])	100.00 (of GWASs in 2 studies[b])
13) For MR studies examining where exposure and/or outcome is a disease/binary trait, describe if cases were incident or prevalent or both	15.38 (10/65)	9.09 (1/11)	16.67 (9/54)
2SMR specific assumptions			
14) Provide adequate descriptive information on GWAS samples to assess whether they represent the same underlying population	37.93	35.71	38.36
15) Provide information on sample overlap.	33.33	35.71	32.88
Data harmonization			
16) Explain how reference alleles were harmonized across exposure and outcome datasets	27.59	28.57	27.40
17) Describe how palindromic SNPs were addressed	14.94	21.43	13.70
Instrument construction			
18) Provide clear criteria or a clear description for the inclusion of genetic variants	72.41	85.71	69.86
19) Describe whether genetic variant-exposure and genetic variant-outcome association estimates were obtained from a discovery GWAS sample, replication GWAS sample, pooled discovery-replication GWAS samples, a meta-analysis of two or more samples or another source	66.67	57.14	68.49
20) Describe whether the instrument was restricted to independent variants or whether the instrument consisted of correlated variants	79.31	92.86	76.71
21) If the instrument consisted of independent variants, how was independence defined/what were the clumping parameters used	67.65 (46/68)	78.57 (11/14)	64.81 (35/54)
22) If the instrument consisted of correlated variants, explain how this was accounted for in the model employed	66.67 (4/6)	NA ($n = 0$)	66.67 (4/6)
23) State the number of primary instruments constructed	97.70	92.86	98.63
24) State the number of included genetic variants	93.10	78.57	95.89
Instrumental variable (IV) assumptions and considerations			
25) Describe how IV assumption 1 (relevance) was assessed	43.68	57.14	41.10
26) Describe how IV assumption 2 (independence) was assessed	31.03	14.29	34.25
27) Describe how IV assumption 3 (exclusion restriction) was assessed	88.51	78.57	90.41

(Continued)

Table 2 Continued

| Question	Total % items reported ($n = 87^a$)	% in studies with ≥10 phenotypes ($n = 14^a$)	% in studies with <10 phenotypes ($n = 73^a$)
28) Homogeneity/monotonicity/constant effect (IV assumption 4) was assessed	1.15	0	1.37
29) If applicable, acknowledge all major assumptions introduced by falsification tests	31.51 (23/73)	16.67 (2/12)	34.43 (21/61)
30) Do they explain why the used sensitivity analyses were used	90.90 (70/77)	90.90 (10/11)	90.91 (60/66)
Analytic methods			
31) List the primary model(s) employed to examine the exposure-outcome association	98.85	92.86	100.00
32) Describe whether any proxy genetic variants were used in the analysis.	24.14	21.43	24.66
33) Describe whether a power calculation has been performed	34.48	35.71	34.25
34) If there was >1 exposure/outcome relationships, either state whether this was corrected for multiple testing or provide justification for/against correcting for multiple testing (or discuss in discussion)	50.94 (27/53)	57.14 (8/14)	48.72 (19/39)
35) Describe any attempts to examine directionality of genetic variant-exposure and genetic variant-outcome associations	39.08	42.86	38.36
36) If analyses were performed using a single genetic variant as an instrumental variable, was co-localisation performed?	16.67 (1/6)	50.00 (1/2)	0 (0/4)
37) If the exposure GWAS used a binary variable, is the causal effect described in terms of liability (or susceptibility) to the exposure in the discussion	29.89	21.43	31.51
38) Do they describe the causal effect in terms of the units of measurement, or relate it to a clinically understandable scale of the exposure	42.31 (11/26)	0 (0/5)	52.38 (11/21)
39) Describe whether any plots are presented to visualize results	72.41	71.43	72.60
40) Describe whether other forms of MR specific bias could be present in analyses	56.32	50.00	57.53
Reproducibility and open science			
41) Present all data used to perform all analyses or describe where data can be accessed	66.67	50.00	69.86
42) Provide R code for performing all analyses	9.20	14.29	8.19

MR, Mendelian randomization; LD, linkage disequilibrium; GWAS, genome-wide association study; QC, quality control; PGRs, polygenic risk score; SNP, single nucleotide polymorphism; PCs, principal components (of the genetic relationship matrix); NA, not available.

[a]For items that were conditional, the percentages were calculated with respect to the number of eligible questions. The numbers in parentheses represent the numerator and denominator of the percentage.

[b]Give information separately for exposure and outcome GWAS. Percentages represent the mean percentage of GWASs that reported the item in each study.

least 67% of the time. The worst-reported items in this domain were the two conditional questions on details about the independence or non-independence of SNPs (both ~67%) and the best-reported items were the two on the number of instruments and genetic variants, both >93%. Items in the 'Instrumental variable (IV) assumptions and considerations' domain were generally not well reported: 89% of studies stated the exclusion restriction assumption and why sensitivity analyses were performed. However, most studies did not describe the other assumptions (44% for relevance and 31% for independence), or any assumptions of the falsification tests (32%).

Most studies described the primary model (99%), presented plots (72%) and described at least one MR specific limitation (56%). However, the other items in the 'Analytic methods' domain were not reported by most studies. For example, just under 50% of studies that tested more than one hypothesis described the use of a multiple testing correction, and fewer than 25% described the conditions for using proxy variants. In the 'Reproducibility and open science' domain, 67% of studies presented all the data used or information on how to access it, but fewer than 10% provided the analysis code.
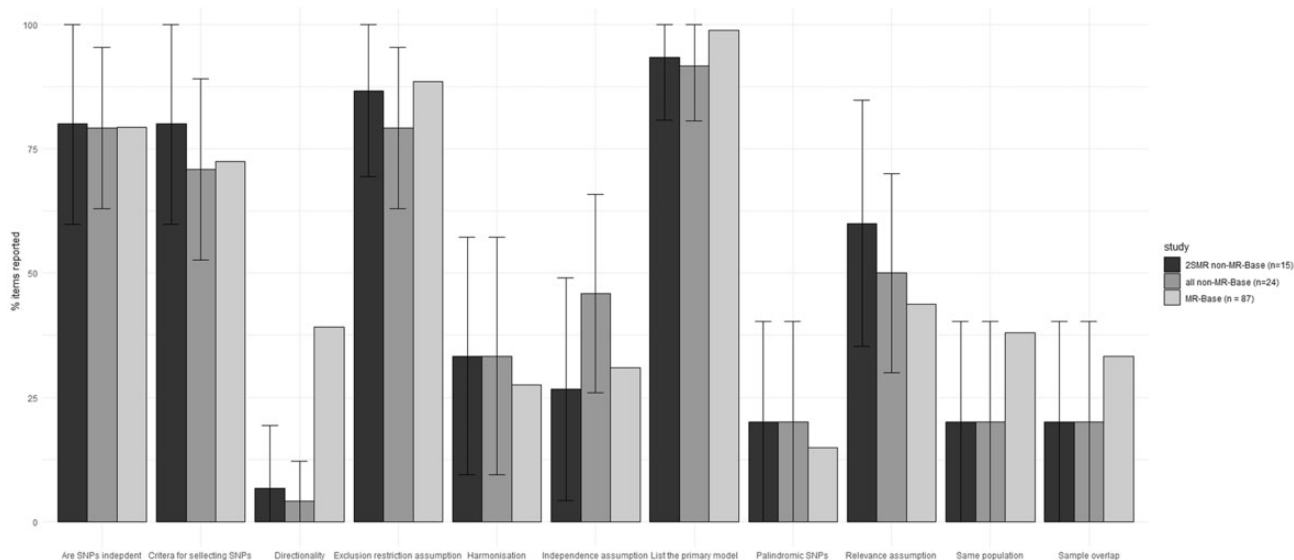
Figure 2 Comparison of MR-Base with non-MR-Base studies. SNP, single nucleotide polymorphism; 2SMR, two-sample Mendelian randomization

Multi-phenotype studies

Broadly, having <10 or ≥10 phenotypes did not seem to affect reporting (overall mean percentage of items reported was 53% and 51%, respectively). However, studies with ≥10 phenotypes were less likely to clearly articulate the research question compared with studies with <10 phenotypes. For example, 97% of studies with <10 phenotypes clearly defined the exposure of interest as compared with only 64% of studies with ≥10 phenotypes. Likewise, 95% and 71% of these studies clearly defined the outcome(s) of interest, respectively. On the other hand, studies with ≥10 phenotypes were more likely to provide sufficient information to assess the relevance assumption (57% vs 41%) and to describe a multiple testing correction when testing multiple hypotheses (57% vs 49%) (Table 2).

Additional analyses

The mean reporting of checklist items did not differ markedly across the 3 years for which there were published studies, with a mean of 47% of items reported in 2017, 55% in 2018 and 51% in 2019 (Supplementary Figure S2, available as Supplementary data at *IJE* online). No studies were included from 2016. Additionally, items for which information is given to the user by MR-Base did appear to increase the probability that they were reported, with a mean reporting of 66% for items covered by both the R-package and web platform, 47% for items with partial coverage and 36% for items not covered by either the R-package or web platform (Supplementary Figure S3, available as Supplementary data at *IJE* online). Studies that used the R Package reported more items (56%) than

studies that did not (46%) (Supplementary Figure S4, available as Supplementary data at *IJE* online).

The results of the additional review we conducted in a random sample of non-MR-Base studies can be found in the Supplementary Results and Supplementary Tables S8, S9 and S10 (available as Supplementary data at *IJE* online). Broadly, studies using MR-Base had a similar level of mean reporting (52%) to those that did not (47%, 95% CI: 34–60, Figure 2). Of the 11 items examined, only one, testing for directionality, implied a discrepancy between the MR-Base and non-MR-Base studies; with 39% of MR-Base studies exploring directionality but only 4% (95% CI: 0–12) of non-MR-Base studies and 6.7% (95% CI: 0–19) of non-MR-Base 2SMR studies. Our search for two-sample MR studies in PubMed yielded 191 publications, of which seven were classed as reviews or systematic reviews by PubMed. Our initial search of two-sample MR studies that used MR-Base yielded 87 studies, suggesting that around half (47%) of all two-sample MR studies used MR-Base.

Discussion

Summary of evidence

We sought to collate published studies that used the MR-Base platform to perform two-sample Mendelian randomization (2SMR) analysis and to assess the quality of reporting of these studies. Our results reveal that reporting quality is generally poor, with 48% of items included in our reporting checklist not reported in an average study.

Many studies omitted information that is important for evaluating the methodological quality of a 2SMR study.

This included, for example, information on the core instrumental variables assumptions (around 40% of studies had adequate reporting for the relevance assumption and around 30% had adequate reporting for the independence assumption), the assumptions of sensitivity analyses for the exclusion restriction assumption (around 30% of studies) and the 2SMR specific assumptions (<40% of studies). Like any epidemiological design, MR can be affected by biases due to procedures employed during data collection.[21] However, papers included in this review did not tend to describe core aspects of the design and methods employed in the underlying GWAS used in MR analyses, including methods used to recruit participants into the study (16%), phenotypic measurements (38%) and quality control measures (25%). Importantly, these issues in reporting do not seem to be specific to MR-Base, with similar reporting quality being found in a random sample of MR papers which did not report using MR-Base (Figure 2).

Previous reviews of MR studies have found similarly poor levels of reporting. For example, Boef and colleagues found that only 33% of studies verified the F statistic of the instruments, 70% provided any discussion of instrument-confounder associations and 33% discussed the exclusion restriction assumption.[12] The increase in the number of studies in this review exploring the exclusion restriction assumption (88.5%) may be due to the development of several commonly used sensitivity analyses for appraising this assumption since the study by Boef and colleagues was published. MR-Base was developed around the time when a number of methods were starting to be developed to examine potential violations of the exclusion restriction assumption, and part of the motivation for developing this platform was to improve the quality of causal inference by automating the application of exclusion restriction sensitivity analyses.[3] By contrast, the higher rate of studies assessing the independence assumption in Boef *et al.* than in this review (70% vs 31%) may be due to the higher prevalence of one-sample MR studies in Boef *et al.*, since it is more straightforward to assess the confounding structure from individual-level data and because assessing this assumption did not feature in the reporting design of MR-Base. Similar to the poor levels of reporting of MR studies identified in Boef *et al*, in their review of 77 oncology MR studies, Lor *et al.* found that only 48% stated the core MR assumptions, 31% estimated statistical power for analyses, 49% described the sample characteristics and 62% assessed the relevance assumption (i.e. instrument-exposure correlation).[14]

Where MR-Base provides information pertaining to the relevant items of the checklist, this was linked to improvements in quality of reporting. However, our results do not necessarily imply that provision of information for item

reporting by MR-Base improves reporting. The similarity of reporting with non-MR-Base studies implies that the apparent trend may be due to selection effects, with MR-Base providing information on items that are easier to report. However, this does not discount the possibility that integrating reporting-nudges into the MR-Base platform, or other modifications, may improve reporting quality. For example, directionality testing is relatively unique to MR-Base compared with other MR software, and there was substantially higher reporting of this in MR-Base studies compared with non-MR-Base studies.

Strengths and limitations

There are several limitations to the current review. The included studies may not be a complete evaluation of all 2SMR papers that used MR-Base, since our inclusion criteria required that authors explicitly mention MR-Base or the MR-Base R package, or cite one of the platform's methods/description papers or their iterations. It is likely that some studies will have been conducted using the platform or R package without providing any citation. If failure to report the statistical software used is indicative of poor reporting in general, then our study will overestimate the quality of reporting in studies using MR-Base. However, such studies would have been eligible for the non-MR-Base sample, and the consistency between these results implies that the difference in reporting may not be large.

Studies included in the review were published between 2016 and 2019. Based on the proportion of studies included, we estimate that approximately 155 relevant studies will have been published subsequently (Supplementary Table S11, available as Supplementary data at *IJE* online). However, no time trend in reporting quality was apparent in our data, so we consider it unlikely that more recent studies would differ markedly in their reporting quality as compared with those studies included in this review. However, if there were changes in reporting quality of studies between 2019 and 2021, this change could partially reflect an effect of the posting of the pre-print for the STROBE-MR checklist,[19] made available in July 2019. Our final search date of April 2019 ensures that the quality of reporting presented is uncontaminated by the availability of the STROBE-MR guidelines. However, it will be important to evaluate the transparency of reporting in 2SMR before and after the publication of the STROBE-MR studies in a future study.

We excluded non-published preprints and other papers not published in peer-reviewed journals from this review to prevent bias that could arise from differences in methodological reporting across studies that underwent peer review

(which could improve reporting) as compared with studies that did not. However, this also means the results of the study may not reflect the quality of reporting in preprints that used MR-Base. Additionally, this means that we were not able to evaluate whether peer review improves reporting quality of 2SMR studies that used MR-Base.

Conclusions

Our review of the reporting quality of 87 two-sample Mendelian randomization studies conducted using MR-Base found that most studies were not well reported. We make two sets of suggestions in light of this. First, MR-Base itself could be adapted to improve study reporting, for example by testing if the introduction of 'nudges', encouraging explicit thought about the analytical decisions being made, enhances reporting quality. Future development of user interfaces according to agreed specifications of appropriate reporting may be effective in improving the quality of published papers. Second, we suggest that authors of Mendelian randomization studies consider using, and that journals endorse, guidelines for reporting MR studies. Indeed, the development of this study's checklist was used to inform the STROBE-MR reporting guidelines.[19] We hope that these guidelines, as well as other MR guidelines which have been published since we started this review, such as the guidelines produced by Burgess *et al.*,[22] will help to improve the quality of reporting of Mendelian randomization studies more generally, and would encourage its use by journals and researchers alike.

Ethics approval

Not applicable for this paper.

Data availability

All materials used in this study are available in the Supplementary material or main text.

Supplementary data

Supplementary data are available at *IJE* online.

Author contributions

B.W. is the guarantor of the review. B.W. drafted the protocol and manuscript, all other authors commented on and edited these. R.R., J.Y., K.D. contributed to the development of the selection criteria, the data extraction criteria and the search strategy. The search was implemented by B.W. and N.C. B.W., N.D., R.R., J.Y. and C.M.S. conducted the data extraction. B.W. conducted the analysis, and B.W., R.R. and J.Y. wrote the report. V.S., R.R., J.Y. and B.W.

implemented the search of non-MR-Base papers. All authors provided feedback on the final draft.

Conflict of interest

None declared.

References

1. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* 2018;**362**:k601.
2. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014;**23**:R89–98.
3. Hemani G, Zheng J, Elsworth B *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *elife* 2018;**7**:e34408.
4. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 2016;**48**:481–87.
5. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* 2017;**46**:1734–9.
6. Pires Hartwig F, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol* 2016;**45**:1717–26.
7. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* 2017;**46**:1734–39.
8. Pires Hartwig F, Tilling K, Davey Smith G, Lawlor D, Borges MC. Bias in two-sample Mendelian randomization when using heritable covariable-adjusted summary associations. *Int J Epidemiol* 2021;**50**:1639–50.
9. Holmes MV, Ala-Korpela M, Davey Smith G. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat Rev Cardiol* 2017;**14**:577–90.
10. Gelman A, Loken E. *The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "p-Hacking" and the Research*

Hypothesis Was Posited Ahead of Time. Department of Statistics, Columbia University, 2013:p. 348. https://doi.org/10.1037/gpr0000135.

11. Swanson SA, Hernán MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology* 2013;**24**:370–74.

12. Boef AG, Dekkers OM, Le Cessie S. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int J Epidemiol* 2015;**44**:496–511.

13. Davies NM, Davey Smith G, Windmeijer F, Martin RM. Issues in the reporting and conduct of instrumental variable studies: a systematic review. *Epidemiology* 2013;**24**:363–69.

14. Lor GC, Risch HA, Fung WT *et al.* Reporting and guidelines for Mendelian randomization analysis: a systematic review of oncological studies. *Cancer Epidemiol* 2019;**62**:101577.

15. Hemani G, Zheng J, Wade KH *et al.* MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv* 2016;078972. doi:10.1101/078972. 16 December 2016, preprint: not peer reviewed.

16. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv* 2016;078972.

17. Zheng J, Haycock P, Hemani G *et al.* LD hub and MR-base: online platforms for preforming LD score regression and Mendelian randomization analysis using GWAS summary data. *Behav Genet* 2016;**46**:815.

18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016;**5**:210.

19. Skrivankova VW, Richmond RC, Woolf BA *et al.* Strengthening the reporting of observational studies in epidemiology using mendelian randomization: the STROBE-MR statement. *JAMA* 2021;**326**:1614–21.

20. Lawlor DA, Wade K, Borges MC *et al.* A Mendelian randomization dictionary: Useful definitions and descriptions for undertaking, understanding and interpreting Mendelian Randomization studies. *Open Science Framework* 2019. doi:10.31219/osf.io/6yzs7, 05 February preprint: not peer reviewed.

21. Swanson SA. A practical guide to selection bias in instrumental variable analyses. *Epidemiology* 2019;**30**:345–49.

22. Burgess S, Davey Smith G, Davies NM *et al.* Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res* 2020;**4**:186.