Journal of Clinical Epidemiology 166 (2024) 111244

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Variation observed in consensus judgments between pairs of reviewers when assessing the risk of bias due to missing evidence in a sample of published meta-analyses of nutrition research

Raju Kanukula[a], Joanne E. McKenzie[a], Aidan G. Cashin[b,c], Elizabeth Korevaar[a], Sally McDonald[d], Arthur T. Mello[e], Phi-Yen Nguyen[a], Ian J. Saldanha[f], Michael A. Wewege[b,c], Matthew J. Page[a,*]

[a]*Methods in Evidence Synthesis Unit, School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia*
[b]*Centre for Pain IMPACT, Neuroscience Research Australia, Sydney, NSW, Australia*
[c]*School of Health Sciences, Faculty of Medicine & Health, University of New South Wales, Sydney, NSW, Australia*
[d]*Charles Perkins Centre, School of Pharmacy, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia*
[e]*Post-Graduate Program in Nutrition, Federal University of Santa Catarina, Florianopolis, Santa Catarina, Brazil*
[f]*Center for Clinical Trials and Evidence Synthesis, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA*

Accepted 19 December 2023; Published online 23 December 2023

Abstract

Objectives: To evaluate the risk of bias due to missing evidence in a sample of published meta-analyses of nutrition research using the Risk Of Bias due to Missing Evidence (ROB-ME) tool and determine inter-rater agreement in assessments.

Study Design and Setting: We assembled a random sample of 42 meta-analyses of nutrition research. Eight assessors were randomly assigned to one of four pairs. Each pair assessed 21 randomly assigned meta-analyses, and each meta-analysis was assessed by two pairs. We calculated raw percentage agreement and chance corrected agreement using Gwet's Agreement Coefficient (AC) in consensus judgments between pairs.

Results: Across the eight signaling questions in the ROB-ME tool, raw percentage agreement ranged from 52% to 100%, and Gwet's AC ranged from 0.39 to 0.76. For the risk-of-bias judgment, the raw percentage agreement was 76% (95% confidence interval 60% to 92%) and Gwet's AC was 0.47 (95% confidence interval 0.14 to 0.80). In seven (17%) meta-analyses, either one or both pairs judged the risk of bias due to missing evidence as ''low risk''.

Conclusion: Our findings indicated substantial variation in assessments in consensus judgments between pairs for the signaling questions and overall risk-of-bias judgments. More tutorials and training are needed to help researchers apply the ROB-ME tool more consistently.

Keywords: Bias; Reporting bias; Meta-analysis; Nutritional sciences; Systematic review; Reliability

Funding: This project was funded by an Australian National Health and Medical Research Council (NHMRC) project grant (APP1139997). RK and P-YN are supported by a Monash Graduate Scholarship and a Monash International Tuition Scholarship. JEM is supported by an Australian NHMRC Investigator Grant (GNT2009612). AGC is supported by an Australian NHMRC Investigator Grant (GNT2010088). SM is supported by the Country Women's Association and Edna Winifred Blackman Postgraduate Research Scholarship. ATM is supported by a Coordination of Improvement of Higher Education Personnel scholarship. MJP was supported by an Australian Research Council Discovery Early Career Researcher Award (DE200101618) during the conduct of this research, and is currently part funded by the Research Support Package of Joanne E McKenzie's NHMRC Investigator Grant (GNT2009612) and a Monash University Future Leader Postdoctoral Fellowship (FLPF23-1069865460). The funders had no role in the study design, data collection and analysis, or preparation of the manuscript.

* Corresponding author. Methods in Evidence Synthesis Unit, School of Public Health and Preventive Medicine, Monash University, 553 St Kilda Road, Melbourne, VIC 3004, Australia. Tel: +61-3-9903-0248.
E-mail address: matthew.page@monash.edu (M.J. Page).

https://doi.org/10.1016/j.jclinepi.2023.111244

What is new?

Key findings

- There was a high raw percentage agreement between pairs for most items in the ROB-ME tool when applied to 42 published meta-analyses of nutrition research, but these estimates had wide confidence intervals. Furthermore, Gwet's ACs, which are corrected for chance, indicated substantial variation in assessments.
- The most common type of judgment observed was one pair judging the risk of bias as "some concerns" while the other judged "high risk". In seven (17%) meta-analyses, either one or both pairs judged the risk of bias due to missing evidence as "low risk".

What this adds to what was known?

- To our knowledge, this is the first study to evaluate the inter-rater agreement in ROB-ME judgments made when users apply the tool to published meta-analyses.

What is the implication and what should change now?

- More tutorials and training are needed to help researchers apply the ROB-ME tool more consistently.
- In future evaluations of inter-rater agreement in ROB-ME assessments, pairing individuals with content and methods expertise would be valuable.

1. Introduction

Systematic reviews (SRs) not only help researchers keep up-to-date with the current evidence and identify research gaps, but they also inform recommendations in clinical practice guidelines, thereby influencing patient care [1]. However, findings of SRs can be compromised if the dissemination of primary study research findings is influenced by factors such as the *P* value, magnitude, or direction of the results. This has been variously referred to as "reporting bias," "dissemination bias," and more recently, "nonreporting bias", which we adopt in this paper [2]. Examples of nonreporting bias include not publishing a study report at all because results were deemed unfavorable ("selective nonpublication of studies" or "publication bias") [3] or not reporting particular results or reporting them incompletely when they were deemed unfavorable ("selective nonreporting of study results") [4]. A consequence of these practices is evidence missing from a meta-analysis, and a potentially biased meta-analysis result.

The Risk of Bias due to Missing Evidence (ROB-ME) tool was developed to assess the risk of bias that arises when entire studies, or particular results within studies, are missing from a meta-analysis because of the *P* value, magnitude or direction of the study results [5]. It guides users to select meta-analyses to evaluate, identify any studies with unavailable results, and consider whether there are unpublished studies before making a judgment about the risk of bias due to missing evidence in a particular meta-analysis result [5]. ROB-ME was designed for use by systematic reviewers seeking to assess the risk of bias due to missing evidence in the meta-analyses they generate as part of their SR. However, ROB-ME could also be used to assess risk of bias in meta-analyses conducted and reported by others; for example, in the context of clinical practice guidelines and overviews of SRs. Given the subjective nature of some steps of the assessment, there is potential for discrepancies between two assessors using the tool. Therefore, investigating the inter-rater agreement in users' ROB-ME assessments of meta-analyses conducted by others can reveal whether there are problems with applying the tool in this manner and inform training needs.

A field for which there is limited research available on the risk of bias due to missing evidence is nutrition. We are aware of one study with similar scope, but the focus was limited to SRs of the association between food/diet and cardiovascular disease or mortality, and the investigators evaluated selective nonpublication of studies only [6]. Therefore, we aimed to (i) evaluate the risk of bias due to missing evidence in a sample of published meta-analyses of the association between food/diet and any health-related outcome using the ROB-ME tool and (ii) determine inter-rater agreement in assessments.

2. Methods

This study was conducted as part of the ROBUST (Risk Of Bias due to Unreported and SelecTively included results in meta-analyses of nutrition research) study [7]. The other aims of the ROBUST study were to explore: (i) whether systematic reviewers selectively included study effect estimates in meta-analyses when multiple effect estimates were available; and (ii) what impact selective inclusion of study effect estimates may have on meta-analytic effects. The results of objectives (i) and (ii) are reported elsewhere [8,9]. We prespecified the methods for the three objectives in a study protocol [7], with deviations from the protocol reported in Supplementary Table S1.

2.1. Creation of a sample of published meta-analyses for assessment

We used the sample of 42 published SRs of nutrition research that we identified for the other components of the ROBUST study. The eligibility criteria, search methods

and selection process we used are described in detail elsewhere [7–9]. Briefly, we searched for SRs indexed in PubMed or Epistemonikos between January 2018 and June 2019 that included a meta-analysis of randomized trials or nonrandomized studies evaluating the effects of at least one type of food or at least one dietary pattern on any health-related outcome. Two investigators independently screened records and potentially relevant full-text reports in random order until 50 SRs meeting the inclusion criteria were identified. One investigator identified the first meta-analysis result reported in the SR (which we call the ''index meta-analysis''). For feasibility reasons, we then restricted inclusion to SRs in which the index meta-analysis included fewer than 20 (and at least two) studies, leaving us with 42 included SRs.

2.2. Preparation of materials for ROB-ME assessments

We used the 22 July 2022 version of the ROB-ME tool in this study (available at https://osf.io/zsb96/). ROB-ME consists of four steps:

1. Complete a table specifying which meta-analyses will be assessed for risk of bias and which study designs and results were eligible for inclusion;
2. Complete a Results Matrix indicating whether each study meeting the inclusion criteria for each meta-analysis has missing results, and if so, whether the missingness is likely related to the *P* value, magnitude or direction of the study result itself;
3. Consider whether scenarios that increase the potential for studies not being identified apply;
4. Assess risk of bias due to missing evidence in each meta-analysis result by answering eight signaling questions (Supplementary Table S2), some of which draw upon the information gathered in Steps 1-3, and others which ask users to consider other factors, such as the pattern of observed study results. The response options for the signaling questions are: Yes; Probably yes; Probably no; No; No information; or Not applicable. ROB-ME includes an algorithm that maps responses to signaling questions onto one of the following proposed risk-of-bias judgments: Low risk of bias; Some concerns; High risk of bias.

To enable an assessment of selective nonreporting of study results (Step 2 of ROB-ME) for the present study, one investigator (RK) sought the reports of all studies not included in the index meta-analyses but reported as meeting the population and intervention eligibility criteria of the index meta-analysis. Reports of studies that were noted as being excluded from the SR for having no useable outcome data were also sought. To enable assessors to compare what was prespecified with what was fully reported, one investigator (RK) searched for a protocol or registration entry for each of the studies that had not been included in the index meta-analyses. This was done by searching PubMed and ClinicalTrials.gov using, if available, the trial registration number specified in the report of the results or the corresponding author's name and key words from the title of the paper.

Table 1. Data items previously extracted for the ROBUST study which informed content of the ROB-ME templates

Category	Data items
SR characteristics	• country and affiliation of corresponding author • source of funding for the SR • number of studies included in the SR
Index meta-analysis characteristics	• type of population • type of interventions/exposures • outcome domain (e.g., weight, cardiovascular function) • which study designs and results were eligible for inclusion • number of studies included in the meta-analysis
Index meta-analysis results	• summary statistics, effect estimates, and measures of precision (e.g., confidence interval) for each study included in the meta-analysis

Abbreviation: SR, systematic review.

Using the data we had previously extracted for the ROBUST study [8,9] (Table 1), one investigator (RK) set up a ROB-ME template for each of the 42 SRs to be assessed (see Supplementary Table S3 for an example) and:

1. Completed Step 1, providing details about the index meta-analysis of the SR;
2. Partially completed Step 2, by listing in the Results Matrix the study identifier and sample size of all studies included and excluded from the index meta-analysis, and recording that results were available for inclusion for each of the studies included in the meta-analysis (leaving the cells for the excluded studies blank).
3. Generated a contour-enhanced funnel plot [10–12] for the meta-analysis, which could be used to inform responses to one of the signaling questions in Step 4.

A second investigator (MJP) reviewed the information recorded in the templates and corrected any errors. We shared the SR report, reports of included and excluded studies, partially completed ROB-ME templates and image files of the funnel plot with each assessor.

2.3. Assessment process

The design of the study was set up to evaluate how ROB-ME is recommended to be applied in practice. The recommendation is that there should be two assessors who independently undertake the assessment, discuss any disagreements in their judgments, and agree on a consensus judgment [5]. Our interest was in examining the agreement

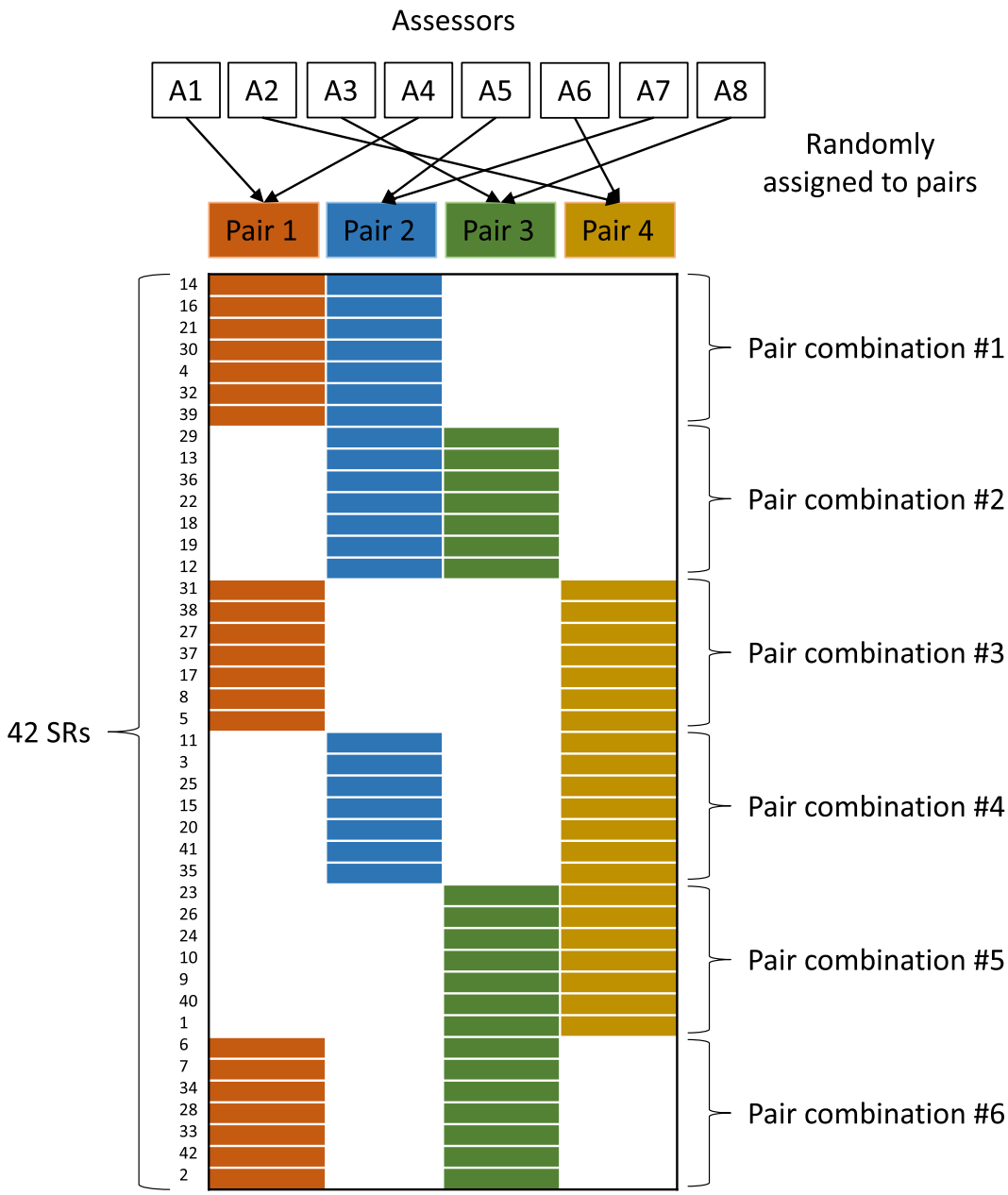

Fig. 1. Representation of study design.

in these consensus judgments, not in the individual judgments. This meant that we needed two pairs (four assessors) assessing each meta-analysis. A constraint that we had to consider in designing the study was that we were only able to recruit a limited number of assessors, and feasibly, we could only ask them to undertake a limited number of assessments.

We planned and were able to recruit eight assessors (RK, AGC, EK, SM, ATM, P-YN, IJS, and MAW) via a personalized invitation email sent from MJP. The assessors were randomly assigned to four pairs. With four pairs, there are six possible pair combinations, and with 42 SRs, this meant that each pair combination was randomly assigned (using computer generated random numbers) to assess the same seven SRs (Fig. 1). This led to each pair assessing the same 21 meta-analyses. To minimize bias, each pair was blinded as to which other pair was assessing a particular meta-analysis, and what their assessments were. MJP, who did not undertake any assessments, communicated with all assessors throughout the project, emailing them all the materials necessary for their assessments along with instructions and prerecorded instructional videos (all materials are available at https://osf.io/zsb96/), and answered questions about the assessment process (without providing information about a particular meta-analysis being assessed).

Assessors were required to complete their ROB-ME assessments using the Microsoft Word templates provided. After completing all 21 assessments, assessors emailed their completed templates to MJP. One of two investigators (MJP or RK) then entered the answers to the eight signaling questions and the risk-of-bias judgment for each meta-analysis into REDCap [13]. To minimize bias, RK did not enter data for any of the SRs which he had been assigned. JEM generated a Microsoft Excel file for each pair, which showed each assessors' answers and judgments, and highlighted those for which a discrepancy existed. Given responses to signaling questions of "Yes" and "Probably yes" have the same implications for risk of bias (as do responses of "No" and "Probably no"), these were not considered discrepancies. Assessors in each pair were given the file and communicated (virtually or in-person) to discuss the discrepancies and to reach consensus. Once finalized, the assessors submitted their consensus responses/judgments to MJP.

2.4. Statistical analysis

We calculated agreement (corrected for chance) in the consensus judgments across the pairs of reviewers for the signaling questions and risk of bias using Gwet's Agreement Coefficient (AC) [14]. Gwet's AC was unweighted for the signaling questions, due to the response options being nominal (see Supplementary Table S2), but was weighted for the risk-of-bias judgment. Ordinal weights were used (as recommended by Gwet for ordinal variables [14]), with categories one-apart weighted as being two-third in agreement ("high risk of bias"-"some concerns", "low risk of bias"-"some concerns"), and two-apart weighted as being 0 in agreement (i.e., complete disagreement) ("high risk of bias"-"low risk of bias"). We undertook a sensitivity analysis to examine the impact of using different weights that attribute more (quadratic) and less (linear) weight, compared with ordinal weights, to the categories one-apart (three-quarter agreement for quadratic, half agreement for linear). In calculating the standard errors (and 95% confidence intervals (CIs)) of Gwet's ACs, we accounted for sampling error arising from having a sample of SRs and of assessors.

To aid in the interpretation of Gwet's ACs, we used a probabilistic method to benchmark the estimated ACs against the following categories: *poor* (< 0.00), *slight* (0.00 to 0.20), *fair* (0.21 to 0.40), *moderate* (0.41 to 0.60), *substantial* (0.61 to 0.80), or *almost perfect* (0.81 to 1.00) [14,15]. We present the *lowest* benchmark category for which there was >95% probability of the true AC falling within the selected, or a higher, category. This approach takes account of the uncertainty in the estimation of the AC. Finally, we calculated raw percentages of agreement (not corrected for chance); for the risk-of-bias judgment, this was calculated with and without the ordinal weights.

Table 2. Characteristics of the included SRs (N = 42)

Characteristic	n (% or median, IQR)
Country of the corresponding author(s)	
China	9 (21)
Iran	7 (17)
United States of America	6 (14)
Other[a]	20 (48)
Affiliation of the corresponding author(s)	
Food industry	2 (5)
Nonindustry	37 (88)
Mixed	2 (5)
Unclear	1 (2)
Source of funding	
Nonprofit	23 (55)
For-profit	3 (7)
Mixed	0
No funding	8 (19)
Not reported	8 (19)
Type of included study	
Only randomized trials	14 (33)
Only nonrandomized studies	26 (62)
Both randomized and nonrandomized studies	2 (5)
Number of included studies in systematic reviews [median (IQR)]	11 (7–19)
Number of studies included in index meta-analyses [median (IQR)]	7 (5–11)
Outcome type	
Continuous	19 (45)
Noncontinuous (e.g., binary, count, time-to-event)	23 (55)

Abbreviations: IQR, interquartile range.
[a] Australia, Austria, Brazil, Canada, Israel, Japan, Malaysia, Spain, Sweden, Thailand, United Kingdom.

Statistical analyses were undertaken in *Stata* version 17 [16] using the *kappaetc* command [17].

3. Results

The results of the search and screening process to identify our sample of SRs are depicted in Supplementary Figure S1. Most of the included SRs (N = 42) [18–59] were authored by systematic reviewers based in China, Iran, or the United States of America (comprising 52% of the sample), with a nonindustry affiliation (88%), and were conducted with funding from a nonprofit source (55%) (Table 2). There were 325 studies included across all index meta-analyses, with a median of seven studies (interquartile range 5-11; range 2-17) per meta-analysis. Designs of studies eligible for inclusion in index meta-analyses were nonrandomized only in 62%, randomized only in 33% and both designs in 5%.

Table 3. Results of percent agreement and Gwet's AC statistic for each item across reviewer pairs ($N = 42$)

Items	Percentage agreement (95% CI)	Gwet's AC (95% CI)	Benchmark descriptor[a]
SQ4.1: Missing or potentially missing results?	81 (55 to 100)	0.70 (0.26 to 1.00)	Fair
SQ4.2: If Yes to 4.1, notable change to synthesis likely?	74 (46 to 100)	0.70 (0.37 to 1.00)	Moderate
SQ4.3: Unclear whether study results were generated?	76 (43 to 100)	0.66 (0.13 to 1.00)	Slight
SQ4.4: If Yes to 4.3, notable change to synthesis likely?	71 (45 to 98)	0.68 (0.36 to 0.99)	Moderate
SQ4.5: Missing or potentially missing studies?	100 (92 to 100)	Ratings do not vary	
SQ4.6: Likely that missing studies had eligible results?	79 (51 to 100)	0.76 (0.46 to 1.00)	Moderate
SQ4.7: Pattern of results suggests missing studies or results?	52 (27 to 77)	0.39 (0.08 to 0.71)	Slight
SQ4.8: Sensitivity analysis suggests synthesis is biased?	69 (41 to 97)	0.66 (0.31 to 1.00)	Fair
Risk-of-bias judgment[b]	76 (60 to 92)	0.47 (0.14 to 0.80)	Slight

Abbreviations: AC, agreement coefficient, CI, confidence interval.

[a] Lowest benchmark category for which there was >95% probability of the true Gwet's Agreement Coefficient (AC) falling within the selected, or a higher, category, which takes account of the uncertainty in the estimation of the AC. Benchmark categories are *poor* (<0.00), *slight* (0.00 to 0.20), *fair* (0.21 to 0.40), *moderate* (0.41 to 0.60), *substantial* (0.61 to 0.80), or *almost perfect* (0.81 to 1.00).

[b] Weighted using ordinal weights.

The eight assessors were Masters/PhD students ($n = 6$), postdoctoral researchers ($n = 1$) or a faculty member ($n = 1$). All assessors had led or contributed to meta-research studies evaluating the conduct and reporting of SRs with meta-analysis. Seven had led or contributed to at least one SR with meta-analysis (range three to 30), had previously appraised a primary study using a tool to assess risk of bias, and had previously assessed nonreporting bias in a study or meta-analysis. Two had conducted research in the field of nutrition.

Across the eight signaling questions in ROB-ME, raw percentage agreement ranged from 52% to 100%, and Gwet's AC ranged from 0.39 to 0.76 (Table 3). These AC estimates corresponded with at least slight to at least moderate agreement. The signaling question for which there was the least agreement between pairs was question 4.7, which asks assessors to consider whether a contour-enhanced funnel plot suggested that the index meta-analysis is likely to be missing results that were systematically different (in terms of *P* value, magnitude or direction) from those observed. For question 4.5, which asks assessors whether circumstances indicate potential for some eligible studies not being identified because of the *P* value, magnitude or direction of the results generated, ratings did not vary between pairs (all pairs responded "Yes").

For the risk-of-bias judgment, the raw weighted percentage agreement was 76% (95% CI 60% to 92%) and raw unweighted percentage agreement was 43% (95% CI 28% to 59%). Gwet's AC was 0.47 (95% CI 0.14 to 0.80), indicating at least *slight* agreement with $\geq$ 95% certainty. Sensitivity analyses examining how Gwet's AC changed depending on the chosen weights showed that the AC was similar using linear (0.38, 95% CI 0.01 to 0.75) or quadratic (0.53, 95% CI 0.21 to 0.85) weights.

The frequencies of different combinations of risk-of-bias judgments for the 42 index meta-analyses are reported in Table 4 (see Supplementary Table S4 for judgments for each meta-analysis); in 18, there was complete agreement, in 21 some disagreement, and in three, complete disagreement. The most common type of judgment observed was one pair judging "some concerns" while the other judged "high risk". In seven (17%) meta-analyses, either one or both pairs judged the risk of bias due to missing evidence as "low risk".

Table 4. Frequency of different types of risk-of-bias judgments by pairs for all index meta-analyses ($N = 42$)

ROB-ME judgment for the meta-analyses	n (%)
Both pairs judged "Low risk"	1 (2)
Both pairs judged "Some concerns"	9 (21)
Both pairs judged "High risk"	8 (19)
One pair judged "Some concerns"; other pair judged "High risk"	18 (43)
One pair judged "Some concerns"; other pair judged "Low risk"	3 (7)
One pair judged "High risk"; other pair judged "Low risk"	3 (7)

Abbreviations: ROB-ME, Risk Of Bias due to Missing Evidence.

4. Discussion

We found there was a high percentage raw agreement (not corrected for chance) between pairs for most items in the ROB-ME tool, but these estimates had wide CIs. Furthermore, Gwet's ACs (corrected for chance) indicated that there was substantial variation in assessments for the signaling questions and for the risk-of-bias judgment. In a minority of cases (individually or as a pair) the risk of bias due to missing evidence was judged as "low risk", which adds to previous concerns that nonreporting biases might frequently influence the results of SRs of nutrition research [6].

There have been several studies investigating agreement in individual judgments when applying tools for assessing the risk of nonreporting bias [60]. For example, studies evaluating the inter-rater agreement in the selective reporting domain for the original Cochrane risk of bias tool found there was slight to fair agreement between two individuals [61−65]. However, our study differed in that we assessed agreement in consensus ratings between pairs of individuals. Given the prevailing advice in the literature that having two assessors is likely to lead to more accurate judgments [66−69], we expected agreement statistics for the consensus pair judgments would be greater than those observed previously for individual assessors. However, this did not arise, and some of the reasons as to why are now explored.

Authors of studies that provide data on inter-rater agreement in risk of bias assessments (for tools other than ROB-ME) have suggested that comprehensive guidance, training, and supporting materials are required to improve the usability and applicability of these tools [70−72]. Support for this claim comes from a study which found that agreement in assessments was higher for assessors who received intensive standardized training compared with assessors who received minimal training [71]. In our study, assessors were given a detailed guidance manual and videos explaining the tool, however, more intensive, formal training appears necessary to enable more consistent application of the tool.

A need for training is exemplified for the assessment involving visual inspection of the funnel plot, which had the least agreement between pairs. Previous research has emphasized the subjective nature of this exercise [73]. e.g., medical researchers shown a sample of funnel plots correctly identified the presence or absence of asymmetry in only half of the plots [74]. In particular, funnel plots can be difficult to interpret when there are few studies, which was the case for the many of the funnel plots included in our sample, with 40% having five or fewer studies. Furthermore, for some meta-analyses in our sample, the disagreement between pairs was due to one making an inference from the plot (e.g., interpreting that it provided evidence of nonreporting bias) while the other pair suggested no such inference was possible given the small number of included studies. More guidance and worked examples on how to interpret funnel plots—particularly those with few studies—should accompany the ROB-ME tool.

ROB-ME assessments of meta-analyses conducted by others might be more challenging (and more prone to inter-rater disagreement) than assessments of meta-analyses conducted by oneself. For example, when conducting one's own meta-analysis, assessments of selective nonreporting of study results might be more consistent because both assessors are likely more familiar with the studies they are evaluating (given their need to consider the studies at multiple stages of the review, e.g. during screening and data collection). Furthermore, assessors are likely to be more familiar with the research field itself when conducting their own meta-analysis, and hence able to more consistently judge whether studies in the field are likely to have been suppressed. Future research is needed to evaluate inter-rater agreement in ROB-ME assessments when the tool is used for the purpose it was originally designed for.

There are several strengths of our study. By calculating agreement between consensus judgments of pairs, not individual judgments, we were able to evaluate how ROB-ME is recommend to be applied in practice (i.e., by two authors, independently). By randomly assigning SRs to pairs, we increased the chance that characteristics of the index meta-analyses that might have influenced assessments (e.g., number of included studies) were balanced across pairs. Blinding of the assessors to what their partner judged for each meta-analysis ensured assessments were done independently and not influenced by one another.

There are also some limitations of our study. None of the assessors attended a training workshop in the use of the ROB-ME tool or received worked examples to consolidate understanding. Also, not all assessors had expertise in nutrition research or experience in applying risk of bias tools. Lack of content expertise might have influenced participants' responses to several signaling questions, such as 4.1 and 4.3 (about selective nonreporting of results in known studies) and 4.6 (about whether potentially missing studies were likely to have had results for the outcome of interest). In future evaluations of inter-rater agreement in ROB-ME assessments, it would be useful to pair individuals with content and methods expertise.

5. Conclusions

There was a high raw percentage agreement between pairs for most ROB-ME items, but these estimates were uncertain. Furthermore, Gwet's ACs, which are corrected for chance, indicated substantial variation in assessments. More tutorials and training are needed to help researchers apply the ROB-ME tool more consistently.

Authors' contributions

All authors met the ICMJE conditions for authorship. MJP and JEM conceived the study design. RK prepared the ROB-ME templates. JEM randomly assigned assessors to pairs and SRs to pairs. RK, AGC, EK, SM, ATM, P-YN, IJS, and MAW conducted ROB-ME assessments. JEM analysed the data. RK wrote the first draft of the manuscript. MJP and JEM drafted sections of the

manuscript. All authors contributed to revisions of the article. All authors approved the final version of the submitted article.

Data availability

The data and analytic code for this study are available on the Open Science Framework at https://osf.io/zsb96/.

Declaration of Competing Interest

MJP is an editorial board member for the *Journal of Clinical Epidemiology*. MJP led and JEM and RK contributed to the development of the ROB-ME tool.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.111244.

References

[1] Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. Chichester, UK: John Wiley & Sons; 2019.
[2] Page MJ, Higgins JP, Sterne JA. Assessing risk of bias due to missing results in a synthesis. Cochrane Handbook for Systematic reviews of Interventions. 2nd ed. Chichester, UK: John Wiley & Sons; 2019: 349−74.
[3] Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010;14:1−220.
[4] Kirkham JJ, Altman DG, Chan A-W, Gamble C, Dwan KM, Williamson PR. Outcome reporting bias in trials: a methodological approach for assessment and adjustment in systematic reviews. BMJ 2018;362:k3802.
[5] Page MJ, Sterne JAC, Boutron I, Hróbjartsson A, Kirkham JJ, Li T, et al. ROB-ME: a tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. BMJ 2023;383: e076754.
[6] de Rezende LFM, Rey-López JP, de Sá TH, Chartres N, Fabbri A, Powell L, et al. Reporting bias in the literature on the associations of health-related behaviors and statins with cardiovascular disease and all-cause mortality. PLoS Biol 2018;16(6):e2005761.
[7] Page MJ, Bero L, Kroeger CM, Dai Z, McDonald S, Forbes A, et al. Investigation of risk of bias due to unreported and selectively included results in meta-analyses of nutrition research: the robust study protocol. F1000Res 2019;8:1760.
[8] Kanukula R, McKenzie JE, Bero L, Dai Z, McDonald S, Kroeger CM, et al. Investigation of bias due to selective inclusion of study effect estimates in meta-analyses of nutrition research. medRxiv 2022:2022.11.01.22281823.
[9] Kanukula R, McKenzie JE, Bero L, Dai Z, McDonald S, Kroeger CM, et al. Methods used to select results to include in meta-analyses of nutrition research: a meta-research study. J Clin Epidemiol 2022;142:171−83.
[10] Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629−34.
[11] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol 2008;61:991−6.
[12] Sterne JA, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. J Clin Epidemiol 2001;54:1046−55.
[13] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377−81.
[14] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.
[15] Gwet KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters: Advanced Analytics. 4th ed. Piedmont, CA: LLC; 2014.
[16] StataCorp L. Stata Statistical Software: Release 15. College Station, TX: Statacorp; 2017: [cited 2021 January 27].
[17] Klein D. Implementing a general framework for assessing interrater agreement in Stata. Stata J 2018;18(4):871−901.
[18] Ayoub-Charette S, Liu Q, Khan TA, Au-Yeung F, Mejia SB, de Souza RJ, et al. Important food sources of fructose-containing sugars and incident gout: a systematic review and meta-analysis of prospective cohort studies. BMJ Open 2019;9(5):e024171.
[19] Bermejo LM, López-Plaza B, Santurino C, Cavero-Redondo I, Gómez-Candela C. Milk and dairy product consumption and bladder cancer risk: a systematic review and meta-analysis of observational studies. Adv Nutr 2019;10(suppl_2):S224−38.
[20] Cheng S, Zheng Q, Ding G, Li G. Mediterranean dietary pattern and the risk of prostate cancer: a meta-analysis. Medicine 2019;98(27): e16341.
[21] Choo VL, Viguiliouk E, Mejia SB, Cozma AI, Khan TA, Ha V, et al. Food sources of fructose-containing sugars and glycaemic control: systematic review and meta-analysis of controlled intervention studies. BMJ 2018;363:k4644.
[22] Eaton JC, Rothpletz-Puglia P, Dreker MR, Iannotti L, Lutter C, Kaganda J, et al. Effectiveness of provision of animal-source foods for supporting optimal growth and development in children 6 to 59 months of age. Cochrane Database Syst Rev 2019;(2): CD012818.
[23] George ES, Marshall S, Mayr HL, Trakman GL, Tatucu-Babet OA, Lassemillante A-CM, et al. The effect of high-polyphenol extra virgin olive oil on cardiovascular risk factors: a systematic review and meta-analysis. Crit Rev Food Sci Nutr 2019;59(17):2772−95.
[24] Ghaedi E, Mohammadi M, Mohammadi H, Ramezani-Jolfaie N, Malekzadeh J, Hosseinzadeh M, et al. Effects of a Paleolithic diet on cardiovascular disease risk factors: a systematic review and meta-analysis of randomized controlled trials. Adv Nutr 2019; 10(4):634−46.
[25] Haider LM, Schwingshackl L, Hoffmann G, Ekmekcioglu C. The effect of vegetarian diets on iron status in adults: a systematic review and meta-analysis. Crit Rev Food Sci Nutr 2018;58(8):1359−74.
[26] Hou R, Wei J, Hu Y, Zhang X, Sun X, Chandrasekar EK, et al. Healthy dietary patterns and risk and survival of breast cancer: a meta-analysis of cohort studies. Cancer Causes Control 2019;30: 835−46.
[27] Huang Y, Zheng S, Wang T, Yang X, Luo Q, Li H. Effect of oral nut supplementation on endothelium-dependent vasodilation−a meta-analysis. Vasa 2018;47:203−7.
[28] Iguacel I, Miguel-Berges ML, Gómez-Bruton A, Moreno LA, Julián C. Veganism, vegetarianism, bone mineral density, and fracture risk: a systematic review and meta-analysis. Nutr Rev 2019; 77(1):1−18.
[29] Kang K, Sotunde OF, Weiler HA. Effects of milk and milk-product consumption on growth among children and adolescents aged 6−18 years: a meta-analysis of randomized controlled trials. Adv Nutr 2019;10(2):250−61.

[30] Kibret KT, Chojenta C, Gresham E, Tegegne TK, Loxton D. Maternal dietary patterns and risk of adverse pregnancy (hypertensive disorders of pregnancy and gestational diabetes mellitus) and birth (preterm birth and low birth weight) outcomes: a systematic review and meta-analysis. Publ Health Nutr 2019;22(3):506−20.
[31] Kodama S, Horikawa C, Fujihara K, Ishii D, Hatta M, Takeda Y, et al. Relationship between intake of fruit separately from vegetables and triglycerides-A meta-analysis. Clin Nutr ESPEN 2018;27:53−8.
[32] Kojima G, Avgerinou C, Iliffe S, Walters K. Adherence to Mediterranean diet reduces incident frailty risk: systematic review and meta-analysis. J Am Geriatr Soc 2018;66(4):783−8.
[33] Larsson SC, Drca N, Jensen-Urstad M, Wolk A. Chocolate consumption and risk of atrial fibrillation: two cohort studies and a meta-analysis. Am Heart J 2018;195:86−90.
[34] Li L, Lietz G, Seal C. Buckwheat and CVD risk markers: a systematic review and meta-analysis. Nutrients 2018;10(5):619.
[35] Li R, Yu K, Li C. Dietary factors and risk of gout and hyperuricemia: a meta-analysis and systematic review. Asia Pac J Clin Nutr 2018; 27(6):1344−56.
[36] Li W, Ruan W, Peng Y, Wang D. Soy and the risk of type 2 diabetes mellitus: a systematic review and meta-analysis of observational studies. Diabetes Res Clin Pract 2018;137:190−9.
[37] Lopez PD, Cativo EH, Atlas SA, Rosendorff C. The effect of vegan diets on blood pressure in adults: a meta-analysis of randomized controlled trials. Am J Med 2019;132(7):875−883.e7.
[38] Maki KC, Palacios OM, Koecher K, Sawicki CM, Livingston KA, Bell M, et al. The relationship between whole grain intake and body weight: results of meta-analyses of observational studies and randomized controlled trials. Nutrients 2019;11(6):1245.
[39] Malmir H, Saneei P, Larijani B, Esmaillzadeh A. Adherence to Mediterranean diet in relation to bone mineral density and risk of fracture: a systematic review and meta-analysis of observational studies. Eur J Nutr 2018;57:2147−60.
[40] Matía-Martín P, Torrego-Ellacuría M, Larrad-Sainz A, Fernández-Pérez C, Cuesta-Triana F, Rubio-Herrera MÁ. Effects of milk and dairy products on the prevention of osteoporosis and osteoporotic fractures in Europeans and non-Hispanic Whites from North America: a systematic review and updated meta-analysis. Adv Nutr 2019; 10(suppl_2):S120−43.
[41] Mena-Sánchez G, Becerra-Tomás N, Babio N, Salas-Salvadó J. Dairy product consumption in the prevention of metabolic syndrome: a systematic review and meta-analysis of prospective cohort studies. Adv Nutr 2019;10(suppl_2):S144−53.
[42] Milajerdi A, Namazi N, Larijani B, Azadbakht L. The association of dietary quality indices and cancer mortality: a systematic review and meta-analysis of cohort studies. Nutr Cancer 2018;70(7):1091−105.
[43] Mishali M, Prizant-Passal S, Avrech T, Shoenfeld Y. Association between dairy intake and the risk of contracting type 2 diabetes and cardiovascular diseases: a systematic review and meta-analysis with subgroup analysis of men versus women. Nutr Rev 2019;77(6): 417−29.
[44] Mohseni R, Abbasi S, Mohseni F, Rahimi F, Alizadeh S. Association between dietary inflammatory index and the risk of prostate cancer: a meta-analysis. Nutr Cancer 2019;71(3):359−66.
[45] Musa-Veloso K, Poon T, Harkness LS, O'Shea M, Chu Y. The effects of whole-grain compared with refined wheat, rice, and rye on the postprandial blood glucose response: a systematic review and meta-analysis of randomized controlled trials. Am J Clin Nutr 2018; 108(4):759−74.
[46] Namazi N, Larijani B, Azadbakht L. Dietary inflammatory index and its association with the risk of cardiovascular diseases, metabolic syndrome, and mortality: a systematic review and meta-analysis. Horm Metab Res 2018;50(05):345−58.
[47] Picasso MC, Lo-Tayraco JA, Ramos-Villanueva JM, Pasupuleti V, Hernandez AV. Effect of vegetarian diets on the presentation of metabolic syndrome or its components: a systematic review and meta-analysis. Clin Nutr 2019;38(3):1117−32.
[48] Qin Z-Z, Xu J-Y, Chen G-C, Ma Y-X, Qin L-Q. Effects of fatty and lean fish intake on stroke risk: a meta-analysis of prospective cohort studies. Lipids Health Dis 2018;17(1):1−7.
[49] Rees K, Takeda A, Martin N, Ellis L, Wijesekara D, Vepa A, et al. Mediterranean-style diet for the primary and secondary prevention of cardiovascular disease. Cochrane Database Syst Rev 2019;3(3): CD009825.
[50] Ren Y, Liu Y, Sun X-Z, Wang B-Y, Zhao Y, Liu D-C, et al. Chocolate consumption and risk of cardiovascular diseases: a meta-analysis of prospective studies. Heart 2019;105:49−55.
[51] Shab-Bidar S, Golzarand M, Hajimohammadi M, Mansouri S. A posteriori dietary patterns and metabolic syndrome in adults: a systematic review and meta-analysis of observational studies. Publ Health Nutr 2018;21(9):1681−92.
[52] Shafiei F, Salari-Moghaddam A, Larijani B, Esmaillzadeh A. Adherence to the Mediterranean diet and risk of depression: a systematic review and updated meta-analysis of observational studies. Nutr Rev 2019;77(4):230−9.
[53] Teoh SL, Lai NM, Vanichkulpitak P, Vuksan V, Ho H, Chaiyakunapruk N. Clinical evidence on dietary supplementation with chia seed (Salvia hispanica L.): a systematic review and meta-analysis. Nutr Rev 2018;76(4):219−42.
[54] Voon PT, Lee ST, Ng TKW, Ng YT, Yong XS, Lee VKM, et al. Intake of palm olein and lipid status in healthy adults: a meta-analysis. Adv Nutr 2019;10(4):647−59.
[55] Wang L, Liu C, Zhou C, Zhuang J, Tang S, Yu J, et al. Meta-analysis of the association between the dietary inflammatory index (DII) and breast cancer risk. Eur J Clin Nutr 2019;73(4):509−17.
[56] Xiao Y, Ke Y, Wu S, Huang S, Li S, Lv Z, et al. Association between whole grain intake and breast cancer risk: a systematic review and meta-analysis of observational studies. Nutr J 2018;17:1−12.
[57] Xu Y, Yang J, Du L, Li K, Zhou Y. Association of whole grain, refined grain, and cereal consumption with gastric cancer risk: a meta-analysis of observational studies. Food Sci Nutr 2019;7(1): 256−65.
[58] Zhang Z, Chen G-C, Qin Z-Z, Tong X, Li D-P, Qin L-Q. Poultry and fish consumption in relation to total cancer mortality: a meta-analysis of prospective studies. Nutr Cancer 2018;70(2):204−12.
[59] de Magalhães Cunha C, Costa PR, de Oliveira LP, Queiroz VAO, Pitangueira JC, Oliveira AM. Dietary patterns and cardiometabolic risk factors among adolescents: systematic review and meta-analysis. Br J Nutr 2018;119(8):859−79.
[60] Page MJ, McKenzie JE, Higgins JP. Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review. BMJ Open 2018;8(3):e019703.
[61] Hartling L, Bond K, Vandermeer B, Seida J, Dryden DM, Rowe BH. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. PLoS One 2011;6:e17242.
[62] Hartling L, Hamm MP, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. J Clin Epidemiol 2013;66:973−81.
[63] Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, Fuentes J, et al. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. PLoS One 2014; 9:e96920.
[64] Hartling L, Ospina M, Liang Y, Dryden DM, Hooton N, Seida JK, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ 2009;339:b4012.
[65] Hartling L, Hamm M, Milne A, Vandermeer B, Santaguida PL, Ansari M, et al. Validity and inter-rater reliability testing of quality

assessment instruments. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012.
[66] Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225−34.
[67] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ 2017;358:j4008.
[68] Boutron I, Page MJ, Higgins JPT, Altman DG, Lundh A, Hróbjartsson A. Considering bias and conflicts of interest among the included studies. In: Cochrane Handbook for Systematic Reviews of Interventions. 2nd ed. Chichester, UK: John Wiley & Sons; 2019: 177−204.
[69] Kolaski K, Logan LR, Ioannidis JPA. Guidance to best tools and practices for systematic reviews. Syst Rev 2023;12(1):96.
[70] Minozzi S, Dwan K, Borrelli F, Filippini G. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. J Clin Epidemiol 2022;141:99−105.
[71] Jeyaraman MM, Robson RC, Copstein L, Al-Yousif N, Pollock M, Xia J, et al. Customized guidance/training improved the psychometric properties of methodologically rigorous risk of bias instruments for non-randomized studies. J Clin Epidemiol 2021;136:157−67.
[72] da Costa BR, Beckett B, Diaz A, Resta NM, Johnston BC, Egger M, et al. Effect of standardized training on the reliability of the Cochrane risk of bias assessment tool: a prospective study. Syst Rev 2017;6(1): 1−8.
[73] Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ 2006;333:597−600.
[74] Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol 2005;58:894−901.