Journal of Clinical Epidemiology 161 (2023) 39–45

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review

Thomas Frederick Crocker[a,*], Natalie Lam[a,1], Magda Jordão[a], Caroline Brundle[a], Matthew Prescott[a], Anne Forster[a], Joie Ensor[b,2], John Gladman[c], Andrew Clegg[a]

[a]*Academic Unit for Ageing and Stroke Research (University of Leeds), Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK*

[b]*Centre for Prognosis Research, Keele School of Medicine, Keele University, Keele, Staffordshire, UK*

[c]*Centre for Rehabilitation & Ageing Research, Academic Unit of Injury, Inflammation and Recovery Sciences, University of Nottingham and Health Care of Older People, Nottingham University Hospitals NHS Trust, Nottingham, UK*

Accepted 20 June 2023; Published online 24 June 2023

Abstract

Objectives: To report our experience using version 2 of the Cochrane risk-of-bias tool for randomized trials (RoB 2).

Study Design and Setting: Two reviewers independently applied RoB 2 to results of interest in a large systematic review of complex interventions and reached consensus. We recorded the time taken, and noted and discussed our difficulties using the tool, and the resolutions we adopted. We explored the time taken with regression analysis and summarized our experience of implementing the tool.

Results: We assessed risk of bias in 860 results of interest in 113 studies. Staff resource averaged 358 minutes per study (SD 183). Number of results ($\beta = 22$) and reports ($\beta = 14$) per study and experience of the team ($\beta = -6$) significantly affected assessment time. To implement the tool consistently, we developed cut points for missingness and considerations of balance regarding missingness, assumed some concerns with intervention deviations unless otherwise prevented or investigated, some concerns with measurements from unblinded self-reporting participants, and judged low risk of selection for certain dichotomous outcomes despite the absence of an analysis plan.

Conclusion: The RoB 2 tool and guidance are useful but resource-intensive and challenging to implement. Critical appraisal tools and reporting guidelines should detail risk of bias implementation. Improved guidance focusing on implementation could assist reviewers.

Keywords: RoB2; Risk of bias; Process duration; Research methods; Certainty assessment; Systematic reviews

1. Introduction

Systematic reviews that synthesize evidence of the effectiveness of interventions are a cornerstone of clinical guidelines and evidence-based medicine. Evaluating risk of bias (RoB) is an essential element of systematic reviews and one step toward establishing the degree of confidence or certainty in the synthesis [1,2]. Methods for evaluating RoB have evolved from study quality checklists to an increasing

Funding: This project is funded by the National Institute for Health Research (NIHR) Health Technology Assessment programme (NIHR128862). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The grant applicants designed the overarching systematic review and network meta-analysis; the funders approved the protocol. The funders have not been involved in any aspect of the collection, analysis or interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Data availability: A de-identified version of the data set upon which this article is based will be made will be openly available indefinitely under a Creative Commons attribution license from the University of Leeds Data Repository. https://doi.org/10.5518/1386

[1] Current address: Mental Health and Addiction Research Group, Department of Health Sciences, Faculty of Science, ARRC Building, University of York, Heslington, York, UK.

[2] Current address: Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK.

* Corresponding author. Bradford Institute for Health Research, Bradford Royal Infirmary, Duckworth Lane, Bradford BD9 6RJ, UK. Tel.: +44-1274-383406.

E-mail address: tom.crocker@bthft.nhs.uk (T.F. Crocker).

https://doi.org/10.1016/j.jclinepi.2023.06.015

What is new?

Key findings

- It took 5h 58m of staff time to assess risk of bias per study on average using version 2 of the Cochrane tool for assessing risk of bias in randomized trials (RoB 2) in a systematic review of community-based complex interventions.
- Variation in time per study was largely explained by models including the number of results of interest, the experience of the reviewers with the use of the RoB 2 tool, and, for individual assessments, number of reports.
- Despite the extensive guidance, we had difficulty implementing aspects of most domains.

What this adds to what was known?

- This is the first research to identify the impact of number of results of interest and number of reports on the time taken to assess RoB 2, and adds to a limited body of published evidence about how reviewers have implemented the RoB 2 tool.

What is the implication and what should change now?

- Improved guidance is needed to further assist review authors in implementing the RoB 2 tool and to encourage reporting details of their approach to implementation.

focus on factors that affect the internal validity of the results of studies. In 2008, the Cochrane Collaboration published a tool to evaluate RoB in randomized controlled trials (RCTs) which became the standard for such assessments [3–5].

Version 2 of the Cochrane risk-of-bias tool for randomized trials (RoB 2) was published in 2019 [6,7]. This revised version, developed through an expert consensus process, introduced substantial changes. For example, each relevant result of a study was now assessed instead of the study as a whole and an overall judgment was to be reached. There was some restructuring of the domains of bias in response to theoretical developments. Additionally, a series of signaling questions (SQs) and accompanying algorithm were provided for each domain to try to improve reviewers' agreement [8].

RoB 2 has been widely cited in the 3 years since its publication (over 8,000 citations according to Google Scholar) suggesting widespread uptake. A recent meta-epidemiological study identified 196 completed systematic reviews that applied RoB 2 [9]. Two studies have estimated the time taken to apply RoB 2, finding it demanding with problematic reliability and recommending the development of operational criteria specific to the review to improve implementation [10,11].

We used RoB 2 to assess RoB in a large, robust systematic review with network meta-analysis. This article reports our experiences, including details to help reviewers planning to use RoB 2, how we operationalized aspects of it, and recommendations for those maintaining and developing the tool.

2. Methods

2.1. Overarching systematic review

We undertook a systematic review and network meta-analysis of community-based complex interventions to sustain independence in older people [12]. The review followed standard procedures, was prospectively registered on PROSPERO (CRD42019162195), and the protocol published [13].

Briefly, the methods of relevance to our use of RoB 2 were as follows:

Following a comprehensive search, we included RCTs or cluster-RCTs that measured outcomes at least 24 weeks after baseline. Participants were older people living at home (≥65 years). Eligible interventions were community-based complex interventions targeted at the individual, focused on sustaining independence. Eligible comparators were usual care, placebo, attention control, or a different complex intervention which met our criteria.

Two researchers independently screened records (title and abstract), assessed eligibility, and extracted data.

Our outcomes of interest comprised six dichotomous outcomes and seven continuous outcomes. Each comparison between two trial arms (e.g., experimental and control interventions) for an outcome of interest at a particular timepoint for which an effect estimate was reported or could be calculated is referred to henceforth as a *result of interest*. Results of interest were extracted for three timeframes.

2.2. Risk of bias assessment using RoB 2

Two reviewers independently assessed RoB in each result of interest from each included study using RoB 2 [6–8]. We were interested in the effect of assignment to the intervention ('intention-to-treat' effect). Disagreements were resolved by consensus between the reviewers or through discussion with the Programme Management Group which included expert clinicians, trialists, and statisticians.

For individually randomized studies, we assessed RoB in five domains: (1) the randomization process; (2) deviations from intended interventions; (3) missing outcome data; (4) measurement of the outcome; and (5) selection of the reported result. Domain 1 was assessed at the study

level and the other domains were assessed at the result level. Cluster-RCTs were assessed similarly, except with two domains of allocation bias in place of domain 1: (1a) the randomization process and (1b) the identification or recruitment of participants [14].

For each domain, we made a judgment of high RoB, low RoB, or some concerns. We used the SQs and algorithm and considered whether to override the algorithm result, recording our reasons and supporting evidence. We reached an overall judgment at least as severe as the most severe domain risk.

Reviewers who conducted RoB 2 assessments read the guidance and watched the Cochrane RoB 2: Learning Live webinars [8,14,15]. Two reviewers had previous experience of using the original Cochrane RoB tool (T.C., N.L.) and three were new to assessing RoB (C.B., M.J., M.P.).

2.3. Evaluation of RoB 2 usage

We recorded details of the time taken per study to conduct assessments (per reviewer) and reach consensus. We discussed and noted our difficulties with using the tool and the resolutions we adopted.

We produced graphs, summary statistics, and conducted multivariable linear regression to explore whether the time taken to assess RoB for a study (per individual, per consensus meeting, overall) was influenced by:

- the number of results to be assessed (increasing time);
- experience using RoB 2 (decreasing time);
- number of reports per study (increasing time).

Additionally, we anticipated variability between individual reviewers and so included this as a categorical variable for regressions of individual assessment time.

In these analyses, we only included studies with time data for two individual assessments. We calculated resource used in person-hours, and full-time equivalent as one person working 7.5 hours per day, 5 days per week, with 33 days leave per year (including public holidays): 142.3 hours per month.

We summarized our approach and reasoning to RoB 2 implementation.

3. Results

3.1. Overarching systematic review

Our literature searches produced 40,112 records after deduplication. We included 129 studies consisting of 496 reports, of which 113 reported results of interest (see Appendix A). Among these studies, there were 860 results of interest for which we assessed RoB,[1] ranging from 1 to 33 per study (median 6).

[1] 34 were unsuitable for inclusion, 826 results are presented in Appendix B.

3.2. RoB 2 assessments

In every result of interest, we judged there to be at least some concerns about overall RoB (28%), with 72% at high RoB. A description of RoB by domain is provided in Appendix B.

3.3. RoB 2 assessment process time

Mean time per study to conduct an individual assessment (per reviewer) was 127 minutes (2h 7m; standard deviation [SD] 67) and 54 minutes (SD 43) for the consensus meeting (see Table 1). We had complete timing data for 99 studies; overall these included 35,472 minutes of worktime (591.2 person-hours or 2.1 months of 2 × full-time equivalent work) including each individual assessment and two people in a consensus meeting (5h 58m per study, 47m per result).

3.3.1. Factors influencing process time

Graphs and results of regression analyses are presented in Appendix C. The number of results of interest per study affected time to conduct individual assessments and consensus meetings, adding 22 minutes per result (95% confidence interval [CI] 18 to 26) to overall worktime. Number of reports affected time to conduct individual assessments but not the consensus meetings, adding 14 minutes per report (95% CI 6 to 22) to overall worktime. Experience reduced time taken to conduct individual assessments and consensus meetings. Additionally, there was substantial variation between individual reviewers.

Thirty-two studies were assessed by two experienced reviewers who had each previously assessed at least 25 studies in this review using RoB 2. For these studies, overall worktime was 5 hours 15 minutes per study or 44 minutes per result on average. Regression analysis estimated this to be 178 minutes per study (2h 58m; 95% CI 139 to 218) plus 19 minutes per result (95% CI 15 to 24) (adjusted R^2 .73).

3.4. Challenges implementing RoB 2

Although the guidance for RoB 2 is extensive, reviewers found certain aspects of the guidance to be lacking specificity sufficient to operationalize it.

3.4.1. Deviations from the intended interventions

We were uncertain what evidence should be considered sufficient to indicate there were ‘probably no’ deviations from the intended intervention that arose because of the trial context (SQ 2.3). This signaling question often decided the judgment for the domain because we were unable to answer that both participants and personnel were unaware of their allocation for any of these complex intervention

Table 1. Descriptive data for time taken to conduct risk of bias assessments and consensus meetings

	Individual assessments[a]	Consensus meetings[b]	Resource for overall process[c]
Studies with complete data (n)	106[d]	105[e]	99[f]
Results of interest per study	7.4 (5.7) [1 to 33]	7.8 (6.0) [1 to 33]	7.6 (5.8) [1 to 33]
Reports per study	3.8 (2.8) [1 to 12]	3.9 (2.8) [1 to 12]	3.9 (2.8) [1 to 12]
Time per study (mins)	127 (67) [32 to 421]	54 (43) [6 to 271]	358 (183) [98 to 976]
Time per result of interest (mins)[g]	17	6.9	47

Data are presented as mean (SD) [min to max] unless otherwise stated.

[a] Studies with timing data for two complete individual assessments only. Times are presented per reviewer so total resource use is double.

[b] Studies with timing data for consensus meetings (following two independent individual assessments). Times are presented for the meeting in which two reviewers were present so total resource use is double.

[c] Studies with timing data for two complete individual assessments and a consensus meeting. Times are presented for total resource use counting both reviewers.

[d] Seven studies missing: four with missing timing data for at least one reviewer, and three with more than two reviewers involved in the assessment.

[e] Eight studies with missing timing data for the consensus meeting.

[f] Fourteen studies not analyzable for the overall process. Of the 106 studies with timing data for (only) two complete individual assessments, seven did not have timing data for the consensus meeting.

[g] Only means are presented as times were recorded per study.

studies (SQ 2.1 and 2.2). The guidance provided an example of trial enrollment and randomization potentially leading control participants to feel in need and unlucky and thus seek other interventions they would not have otherwise. This seemed like a risk that was always plausible although, perhaps, often unlikely, and one that was very difficult for trialists to investigate.

We answered 'probably no' for four studies: three where authors had specifically investigated control group behaviors and found no concerns; and one where we considered the stepped-wedge design would make such deviation unlikely. We answered 'no information' for most studies, answering 'probably yes' for nine studies with evidence of contamination.

3.4.2. *Missing outcome data*

We found several difficulties in implementing the guidance regarding the assessment of missing outcome data. Firstly, standardizing the assessment of 'nearly all' data being available (SQ 3.1). We operationalized this as at most 5% of participants missing as a proportion of: the number allocated for continuous outcomes; or, recorded cases for binary outcomes (e.g., 1 person missing if 20 people died regardless of sample size).

SQ 3.2 asks "Is there evidence that the result was not biased by missing outcome data?" [8]. We judged there was insufficient evidence, even when study authors had conducted multiple imputations with multiple measured variables to correct for bias. Our outcomes of interest were so distal it was unlikely *all* relevant variables were included in such models (we were uncertain whether this could ever be a plausible assumption).

For almost all results where some data were missing, it was plausible that missingness *could* have depended on the true value of the outcome (SQ 3.3). This was because outcomes were health outcomes and reasons for missingness typically included mortality, care home admission, or withdrawal. SQ 3.4 delineates between a judgment of some concerns and high risk. It asks whether it is *likely* missingness depended on the true value of the missing result; the elaboration indicates we should answer '(probably) yes' if "reported reasons for missing outcome data provide evidence that missingness in the outcome depends on its true value" [8]. Therefore, when we first applied the tool we invariably answered 'probably yes' to SQ 3.4 based on our reason for SQ 3.3. Almost all results were therefore judged high risk. It seemed inappropriate and at odds with the preceding guidance that we would judge a high RoB due to missingness when losses were relatively small and balanced in numbers and reasons; for example, for a continuous outcome where 5.1% of participants had died in both arms.

To decide between a judgment of some concerns and high risk, we first considered whether more than 45% of participants/cases were missing overall, regardless of how balanced, as an arbitrary upper limit. Secondly, we considered whether differences in total numbers or reasons missing differed by our threshold values of 5%, or if insufficient detail regarding reasons was given. In any of these cases, we would make a judgment of high risk. For remaining results, where between 5% and 45% of participants/cases were missing overall, we sought increasing balance in the numbers and reasons of missingness to judge the result some concerns only.

We decided against answering 'no information' to SQ 3.4 (which would have resulted in a judgment of high risk) in the common situation where numbers and reasons for losses were presented by group for a trial overall but not for each specific result of interest, although we were uncertain this was how we were supposed to interpret the guidance.

3.4.3. Measurement of the outcome

For participant-reported outcomes (PROs) (''involving judgment''), we had to judge the likelihood that knowledge of the intervention influenced assessment (high risk) or not (some concerns) (SQ 4.5). Based on the guidance, we considered the degree of participant judgment for each outcome, whether certain interventions (e.g., alternative medicine) or closely related outcomes (e.g., activities of daily living training for activities of daily living outcomes) should indicate high risk, and whether studies with active comparators should only be rated 'some concerns' despite some of these features. We were concerned that such judgments were predicated on assumptions rather than evidence, risking our own biases influencing the assessment. In the absence of stronger empirical evidence regarding the factors influencing RoB for PROs, we decided that knowledge of the intervention could influence assessment but that this was unlikely ('some concerns' in the absence of other problems). Our reasoning was that unblinded observers were a greater risk than unblinded self-reporting participants, many of the interventions may not be recognized as departures from usual care by participants, and measurements would often be temporally distant from intervention receipt. We were able to judge this domain low risk for some PRO results for which we concluded that the participants were probably unaware of the intervention received, usually in cluster trials.

3.4.4. Selection of the reported result

Domain five considers selective reporting of results. To reach a judgment of low risk, the algorithm requires the result of interest to have been analyzed in accordance with a prespecified analysis plan. Such plans are rarely available. When a dichotomous outcome that could only be measured in one way (e.g., mortality) was reported as the number of participants, we considered that this was equivalent to a situation where we gathered and reanalyzed individual patient data, and therefore judged the risk as low, regardless of a prespecified plan.

4. Discussion

This article detailed the substantial resource taken to conduct RoB 2 assessments in a systematic review (358 mins/study) and estimated that multiple results and study reports increased time per study, while experience reduced time taken. We have detailed the way we operationalized guidance for missing outcome data, deviations due to trial context, knowledge of the intervention influencing assessment, and selection of the reported result. Overall, we found the SQs and algorithms helpful. However, sections of the guidance seemed unnecessarily discursive and theoretical with insufficient practical advice for interpreting the SQs. We also found the wording for some SQs misleading. Additionally, completing the SQs and the supporting free-text boxes for each SQ and each result in each study was time-consuming.

Our results relate to one particularly challenging systematic review with network meta-analysis of pragmatic trials of complex interventions that required deliverer and participant involvement. Therefore, the time taken and some of our difficulties may not manifest for other reviews. However, the findings are from a real-world review without special support from the Cochrane Methods Support Unit. The order in which we reviewed studies was not randomly selected and so estimates of the effect of experience are limited in this regard. We did not attempt to clarify uncertainties regarding RoB with study authors, as recommended for domain 5 [8]; such action may have affected the time we took as well as our judgments. The approaches we took to implementing RoB 2 were based on extensive reading of the guidance and discussion among the authors who comprise experienced trialists and systematic reviewers. Nonetheless, they may not fit with the intentions of the tool authors, so we advise caution in following these approaches.

We are aware of only one other study that reports upon the implementation of the RoB 2 tool in a systematic review: Minozzi et al. analyzed the impact an implementation document had on times and inter-rater reliability of their RoB 2 assessments in a review of cannabis and cannabinoids for people with multiple sclerosis [11]. The supplementary implementation document details their approach to similar challenges. Like Minozzi et al. we identified an improvement in speed over time, although to a lesser extent. We additionally estimated how number of results and reports per study affected time. Minozzi et al. implemented a 90% cut point for judging nearly all data being available, while we implemented a 5% rule relating to missingness (SQ 3.1). More specific guidance from the RoB 2 tool authors would help to develop consistency between review teams, in the absence of which reviewers are likely to develop their own cut points. We decided that none of the analyses that attempted to correct for missing outcome data were sufficient to judge a low RoB whereas Minozzi et al. accepted these. For bias due to assessment of the outcome being influenced by knowledge of the intervention received we both decided that PROs requiring judgment should be treated as some concerns rather than high risk, although their position was informed by relevant evidence.

We agree with Minozzi et al. [11] that review authors should develop guidance specific to their review to establish how to assess issues such as those described in this article and assure consistency across assessed results. Review authors can use our findings to plan the time it will take to assess RoB, although this may vary substantially. Review users should be aware that RoB 2 implementation will affect judgment and thus certainty assessments.

Our findings suggest refined guidance from the tool developers is warranted, with a focus on operationalizing the tool. We would welcome further specific examples directly

related to the SQs, and more examples of 'judgment calls' rather than extremes. For example, the What Works Clearinghouse Standards Handbook provides boundaries for unacceptable RoB for combinations of overall and differential attrition under cautious and optimistic assumptions [16], the former being similar to our implementation for SQ 3.4. However, we recognize that it is important to limit the overall volume of guidance. Sometimes the theoretical background included factors that were not to be assessed; these and empirical evidence could be moved to an appendix. Reporting guidelines and critical appraisal tools should also include details of RoB implementation (e.g., Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 [17], A MeaSurement Tool to Assess systematic Reviews 2 [18]).

5. Conclusion

The RoB 2 tool is a positive development from the original RoB tool, with the addition of SQs and an algorithm likely to improve consistency if carefully followed. Assessing individual results is useful for differentiating within-study RoB between outcomes and timepoints, particularly in domains such as missing outcome data. This combined with an overall RoB judgment assists in progressing to an assessment of the certainty of the evidence. Nonetheless, conducting assessments with RoB 2 is a substantial and challenging undertaking. We recommend reporting guidelines and critical appraisal tools reflect this, requiring implementation details. Furthermore, the burden on reviewers should be reduced by improving the guidance with greater emphasis on the application of, and location of dividing lines for, each signaling question.

CRediT authorship contribution statement

Thomas Frederick Crocker: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Visualization, Supervision, Funding acquisition, Writing - original draft, Writing - review & editing. **Natalie Lam:** Conceptualization, Methodology, Investigation, Resources, Data curation, Writing − original draft, Writing − review & editing, Project administration. **Magda Jordão:** Conceptualization, Investigation, Data curation, Writing − review & editing. **Caroline Brundle:** Investigation, Writing − review & editing. **Matthew Prescott:** Investigation, Writing − review & editing. **Anne Forster:** Methodology, Writing − review & editing, Supervision, Funding acquisition. **Joie Ensor:** Methodology, Writing − review & editing, Supervision, Funding acquisition. **John Gladman:** Methodology, Writing − review & editing, Supervision, Funding acquisition. **Andrew Clegg:** Methodology, Conceptualization, Writing − review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

Andrew Clegg declares funding through the NIHR HTA programme, NIHR Programme Grants for Applied Research, NIHR HS&DR programme, NIHR Applied Research Collaboration Yorkshire & Humber, and Health Data Research UK; Anne Forster declares NIHR Senior Investigator Award 2017-present, NIHR Programme Grant 10% of salary, NIHR HS&DR grant 8% of salary, HTA grant 5% of salary, National Institute for Health (USA) payment for panel membership 2021, 2022, participation in Programme Steering Committees for NIHR 202,339 Improving the lives of stroke survivors with data, and NIHR202020 Research Title Personalised Exercise-Rehabilitation FOR people with Multiple long-term conditions (multimorbidity)-The PERFORM trial, University of Leeds Governor representative on the Governors Board of Bradford Teaching Hospitals NHS Foundation Trust, member of HSDR Researcher-Led panel, member of NIHR Doctoral Fellowship Panel member of Policy Research Unit assessment panel. Other authors declare no potential conflicts of interest.

Acknowledgments

We thank the members of the wider systematic review and network meta-analysis project team who attended Project Management Group meetings: Professor of biostatistics Richard Riley, applied statistics lecturer Ram Bajpai, and medical statistics research assistant Matthew Bond, aging research assistant Eleftheria Patetsini, aging research assistant Ridha Ramiz; or who otherwise co-authored the project report on which this is based, including: information specialist Deirdre Andre, aged care researcher Alison Ellwood, rehabilitation research programme manager John Green, geriatric academic fellow Matthew Hale, geriatric medicine doctor Jessica Morgan, cardiovascular aging clinical lecturer Oliver Todd, and anesthetics doctor Rebecca Walford. We are grateful to Jasmin Manik for her support with testing the RoB 2 tool algorithms.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.06.015.

References

[1] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. Chichester: John Wiley & Sons; 2019.

[2] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.

[3] Higgins JP, Altman DG. Assessing risk of bias in included studies. In: Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Chichester: John Wiley & Sons; 2008: 187−241.

[4] Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.

[5] Jørgensen L, Paludan-Müller AS, Laursen DRT, Savović J, Boutron I, Sterne JAC, et al. Evaluation of the Cochrane tool for assessing risk of bias in randomized clinical trials: overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. Syst Rev 2016;5:80.

[6] Higgins JPT, Savović J, Page MJ, Elbers RG, Sterne JAC. Chapter 8: assessing risk of bias in a randomized trial Cochrane. In: Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al, editors. Cochrane handbook for systematic reviews of interventions Version 6.3. Cochrane; 2022. Available at www.training.cochrane.org/handbook/. Accessed January 12, 2023.

[7] Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.

[8] Higgins JPT, Savović J, Page MJ, Sterne JAC, RoB2 Development Group. Revised Cochrane risk-of-bias tool for randomized trials (RoB 2) 2019. Available at https://www.riskofbias.info/. Accessed January 12, 2023.

[9] Minozzi S, Gonzalez-Lorenzo M, Cinquini M, Berardinelli D, Cagnazzo C, Ciardullo S, et al. Adherence of systematic reviews to Cochrane RoB2 guidance was frequently poor: a meta epidemiological study. J Clin Epidemiol 2022;152:47−55.

[10] Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol 2020;126:37−44.

[11] 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.

[12] Crocker TF, Lam N, Ensor J, Jordão M, Bajpai R, Bond M, et al. Community-based complex interventions to sustain independence in older people, stratified by frailty: a systematic review and network meta-analysis. Health Technol Assess 2023. in press.

[13] Crocker TF, Clegg A, Riley RD, Lam N, Bajpai R, Jordão M, et al. Community-based complex interventions to sustain independence in older people, stratified by frailty: a protocol for a systematic review and network meta-analysis. BMJ Open 2021;11:e045637.

[14] Eldridge S, Campbell MK, Campbell MJ, Drahota AK, Giraudeau B, Reeves BC, et al. Revised Cochrane risk of bias tool for randomized trials (RoB 2): Additional considerations for cluster-randomized trials (RoB 2 CRT) 2021. Available at https://www.riskofbias.info/. Accessed January 12, 2023.

[15] Cochrane Training. RoB 2: Learning Live webinar series. Available at https://training.cochrane.org/rob-2-learning-live-webinar-series. Accessed January 12, 2023.

[16] What Works Clearinghouse. What Works Clearinghouse Procedures and Standards Handbook Version 5.0. Available at. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation; 2022. https://ies.ed.gov/ncee/wwc/handbooks. Accessed May 30, 2023.

[17] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.

[18] 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.