Development of an Instrument for Assessing Risk of Bias of Randomized Trials in Systematic Reviews

Development of an Instrument for Assessing Risk of Bias of Randomized Trials in Systematic Reviews

Ying Wang, MPharm

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

McMaster University © Copyright by Ying Wang, September 2024

McMaster University DOCTOR OF PHILOSOPHY (2024) Hamilton, Ontario (Health Research Methodology)

TITLE: Development of an Instrument for Assessing Risk of Bias of Randomized Trials in Systematic Reviews

AUTHOR: Ying Wang, MPharm

SUPERVISOR: Dr. Gordon H Guyatt

NUMBER OF PAGES: ix, 233

Abstract

Assessment of risk of bias in the included randomized controlled trials (RCTs) has become an essential step in systematic reviews, which informs the decision of whether to rate down certainty of evidence due to risk of bias applying the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach. Many instruments exist for rating risk of bias in RCTs; however, even those most commonly used that developed by the Cochrane group, suffer from limitations. In particular, the revised Cochrane instrument, while reflecting methodological advances, sacrificed simplicity and practicability.

The objective of this thesis is to use rigorous methodology to develop a simple-structured RCT risk of bias instrument that is easy for systematic review authors to use. The thesis begins with a chapter introducing the background and the structure of the thesis. The thesis subsequently describes a systematic survey of existing RCT risk of bias instruments for their included items, through which we collected potential candidate items for the new instrument. We then present a summary of empirical evidence investigating how the possible risk of bias issues influence the estimates of intervention effects in RCTs, which assisted with the item selection for the new instrument. Then, this thesis describes the detailed process for instrument development and providing the new instrument. This thesis ends with a chapter summarizing key findings, discussing strengths and limitations, and exploring directions for future research.

Acknowledgements

First and foremost, I would like to express my huge thanks to my supervisor Dr. Gordon H Guyatt. I'm so lucky to have Dr. Gordon H Guyatt to be my supervisor, who actually led me into the area of evidence-based medicine and deeply influenced my way of thinking and working habit. He serves as a wonderful role model for me as how to be a methodologist. He gave me top-level guidance and training especially on systematic review and guideline methodology. I appreciate his tremendous support and encouragement during my PhD. Gordon, I will continue to practice the ideas and knowledge of evidence-based medicine that you taught me.

I also want to express my great thanks to my supervisory committee, Dr. Romina Brignardello-Petersen and Dr. Reed AC Siemieniuk, for their professional advice and great support. Romina, I appreciate for your valuable feedback and also the challenging questions you raised, which definitely helped improving the quality of my work. From you, I have learned the rigorous work attitude of methodologists. Reed, I have been grateful for your great support since the earliest project we worked together. You are one of the first methodologists I know. I learnt a lot about review and guideline methodology from you.

To all the co-authors, thanks for your efforts, commitments, and valuable suggestions on this thesis and other projects. Dear panel members, we cannot develop the instrument without your contributions and the team collaboration.

To the faculty members at the Department of Health Research Methods, Evidence, and Impact, it's really my great pleasure to be in such a wonderful academic environment and work with you. To all my dear friends at McMaster, thank you for being so nice and

friendly and helping me through a difficult time when I first came to Canada.

My parents deserve my deepest gratitude. I wouldn't be here without their support. Dad, you are the person who gives me the biggest support and understanding for my career pursuit. Your spirit of exploring and learning new things has deeply influenced me since I was very young. Mom, you give me the warmest love in life. Thank you for your tolerance and understanding of me.

Table of Contents

Abstract iii

Acknowledgements iv

Lists of Figures vii

Lists of Tables viii

Declaration of Academic Achievement ix

Chapter 1: Introduction to The Thesis 1

Chapter 2: Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues 7

Chapter 3: Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors 36

Chapter 4: Development of the Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT) 123

Chapter 5: Discussion and Conclusion to This Thesis 227

Lists of Figures

Page

CHAPTER 1

None

CHAPTER 2

Figure 1 Study selection flow chart 11

Figure 2 Item selection and classification process 13

CHAPTER 3

Figure 1 Flow diagram of selection of studies 42

CHAPTER 4

None

CHAPTER 5

None

Lists of Tables

Page

CHAPTER 1

None

CHAPTER 2

Table 1 Included instruments 12

CHAPTER 3

Table 1 Summary of characteristics of included meta-epidemiological studies 43

Table 2 Impact of potential risk of bias elements on effect estimates in randomized trials 46

CHAPTER 4

Table 1 Initially selected core items and optional items and judgement regarding whether they met the six item selection criteria 143

Table 2 ROBUST-RCT core items and two steps 145

Table 3 ROBUST-RCT optional items 146

CHAPTER 5

None

Declaration of Academic Achievement

This is a “sandwich thesis” comprised of five chapters.

Chapter 1 is unpublished. YW is the sole author.

Chapter 2 is published in *Journal of Clinical Epidemiology*. GHG and YW conceived the idea of this chapter. RC conducted the literature search. YW, MG, QW, LH, AI, CH, LY, MH, ZY and DZ collected data. SAO, DB, MB, LLG, PG, RJ, SK, LL, PR, KFS, DZ, YW and GHG classified the items. YW analyzed data. YW drafted the first version of the article and all others reviewed and revised. All authors approved the final version of the article.

Chapter 3 is published in *Journal of Clinical Epidemiology*. GHG and YW conceived the idea of this chapter. YW, RC, and QW collected data. SP and YW analyzed data. YW and GHG assessed certainty of evidence. YW and GHG drafted the first version of the article and all others reviewed and revised. All authors approved the final version of the article.

Chapter 4 is under review at *The British Medical Journal*. GHG and YW developed the idea. GHG, YW, RBP, RAS, and DZ were operations committee members. GHG, YW, RBP, RAS, DZ, MB, PG, EAA, SAO, DB, CG, LLG, JLH, PR, KFS, DJT, SK, LML were the panel members. YW and GHG drafted the first version of the article and all others reviewed and revised. All authors approved the final version of the article.

Chapter 5 is unpublished. YW is the sole author.

Chapter 1: Introduction to The Thesis

Study limitations in randomized controlled trials (RCTs) could result in bias (1). Assessment of risk of bias in the included RCTs has become an essential step in systematic reviews, which informs the decision of whether to rate down certainty of evidence due to risk of bias applying the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach (2, 3). Many instruments exist for rating risk of bias in RCTs (4); however, even those developed by the Cochrane group and most commonly used (5, 6) suffer from limitations. In particular, the revised Cochrane instrument (6), while reflecting advances in risk of bias methodology, sacrificed simplicity and practicability (7-9).

The objective of this thesis is to use rigorous methodology to develop an instrument for rating risk of bias in RCTs that is easy for use by systematic review authors.

Chapter 2 of the thesis is a systematic survey of instruments addressing risk of bias in RCTs that published from 2010 to October 2021. We extracted the items included in these instruments. After excluding the items that two reviewers agreed clearly did not address risk of bias, for the remaining items, we conducted a survey of 13 experts in risk of bias methodology and evidence-based medicine. Through this survey, we classified the items into three categories: items that most of the 13 experts thought address risk of bias; items that most thought address other issues (applicability, imprecision, reporting quality or others) rather than risk of bias; and items that experts had major disagreement about whether or not they address risk of bias. This chapter provided the candidate items for the new instrument: panelists sequentially discussed items in the three categories (as part of chapter 4). The item classification results informed the extent to which the items meet one of the six criteria for our item selection for the new instrument “item addressing clearly risk of bias issue rather than others” (as part of chapter 4).

Chapter 3 is a systematic survey of meta-epidemiological studies evaluating impact of possible risk of bias items on estimates of intervention effects in RCTs. This study followed advanced systematic review methodology. Incorporating both the GRADE approach (2) and the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) (10), we developed an approach to rating certainty of inference from meta-epidemiological studies. We presented a summary of findings table including inferences regarding the impact of all possible risk of bias items on the estimates of effects and our certainty in the inferences. The results informed the extent to which the items meet one of the item selection criteria for the new instrument "empirical evidence supports item influence on effect estimates" (as part of chapter 4).

Chapter 4 describes the development of the new instrument, named Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT). We followed a rigorous instrument development process: we assembled a panel of experts with diverse backgrounds; established ground rules for the new instrument; conducted preparatory studies to support the instrument development (as described in chapter 2 and 3); held panel meetings to reach consensus on item selection and instructions; drafted the instrument document and user manual; and conducted pre-testing with the systematic reviewers and based on the feedback improved the instrument.

Chapter 4 presents the final version of the ROBUST-RCT and the manual. ROBUST-RCT includes six core items each of which includes two steps: first evaluating what happened in individual trials and second judging the associated risk of bias. ROBUST-RCT provides eight optional items that may be relevant in specific cases. We believe that ROBUST-RCT is simple and easy to use by systematic reviewers with different levels of expertise.

Chapter 5 summarizes main findings of the thesis, discusses its strengths and limitations, and explores directions for future studies.

References

1. Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs included in systematic reviews. J Clin Epidemiol. 2001;54(7):651-4.

2. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.

3. 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(4):407-15.

4. Wang Y, Ghadimi M, Wang Q, Hou L, Zeraatkar D, Iqbal A, et al. Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues. J Clin Epidemiol. 2022;152:218-25.

5. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.

6. Sterne JAC, Savovic 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.

7. Kuehn R, Wang Y, Guyatt G. Overly complex methods may impair pragmatic use of core evidence-based medicine principles. BMJ Evid Based Med. 2024.

8. Martimbianco ALC, Sa KMM, Santos GM, Santos EM, Pacheco RL, Riera R. Most Cochrane systematic reviews and protocols did not adhere to the Cochrane's risk of bias 2.0 tool. Rev Assoc Med Bras (1992). 2023;69(3):469-72.

9. 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.

10. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ. 2020;192(32):E901-

E6.

Chapter 2: Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues

Cited as and reprinted from: Wang Y, Ghadimi M, Wang Q, Hou L, Zeraatkar D, Iqbal A, et al. Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues. J Clin Epidemiol. 2022 Dec;152:218-225.

ELSEVIER

Journal of Clinical Epidemiology 152 (2022) 218–225

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues

Ying Wang[a,*], Maryam Ghadimi[a], Qi Wang[b], Liangying Hou[b], Dena Zeraatkar[a,c], Atiya Iqbal[a], Cameron Ho[d], Liang Yao[a], Malini Hu[e], Zhikang Ye[a], Rachel Couban[f], Susan Armijo-Olivo[g,h], Dirk Bassler[i], Matthias Briel[a,j], Lise Lotte Gluud[k], Paul Glasziou[l], Rod Jackson[m], Sheri A. Keitz[n], Luz M. Letelier[o], Philippe Ravaud[p], Kenneth F. Schulz[q], Reed A.C. Siemieniuk[a], Romina Brignardello-Petersen[a], Gordon H. Guyatt[a]

[a]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada
[b]Department of Social Medicine and Health Management, School of Public Health, Lanzhou University, Lanzhou, China
[c]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[d]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, Ontario, Canada
[e]Michael G. DeGroote School of Medicine, McMaster University, Hamilton, Ontario, Canada
[f]Department of Anesthesia, McMaster University, Hamilton, Ontario, Canada
[g]University of Applied Sciences, Faculty of Business and Social Sciences, Osnabrück, Germany
[h]Faculty of Rehabilitation Medicine, Department of Physical Therapy, University of Alberta, Edmonton Canada
[i]Department of Neonatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[j]Meta-Research Centre Basel, Department of Clinical Research, University Hospital Basel, Switzerland
[k]Gastro Unit, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark
[l]Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Queensland, Australia
[m]Section of Epidemiology & Biostatistics at the School of Population Health, Faculty of Medical and Health Sciences, University of Auckland, New Zealand
[n]Department of Medicine, Lahey Hospital & Medical Center, Burlington, MA, USA
[o]Department of Internal Medicine, Escuela de Medicina, Pontificia Universidad Catolica de Chile, Alameda 340, Santiago, Chile
[p]Epidemiology and Statistics Sorbonne Paris Cité Research Center (CRESS), INSERM, Université Paris Descartes, Paris, France
[q]School of Medicine, University of North Carolina at Chapel Hill, USA

Accepted 21 October 2022; Published online 28 October 2022

Abstract

Objectives: To establish whether items included in instruments published in the last decade assessing risk of bias of randomized controlled trials (RCTs) are indeed addressing risk of bias.

Study Design and Setting: We searched Medline, Embase, Web of Science, and Scopus from 2010 to October 2021 for instruments assessing risk of bias of RCTs. By extracting items and summarizing their essential content, we generated an item list. Items that two reviewers agreed clearly did not address risk of bias were excluded. We included the remaining items in a survey in which 13 experts judged the issue each item is addressing: risk of bias, applicability, random error, reporting quality, or none of the above.

Results: Seventeen eligible instruments included 127 unique items. After excluding 61 items deemed as clearly not addressing risk of bias, the item classification survey included 66 items, of which the majority of respondents deemed 20 items (30.3%) as addressing

Declaration of interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions: GHG and YW conceived the study idea. RC conducted the literature search. YW, MG, QW, LH, AI, CH, LY, MH, and ZY conducted study selection. YW and DZ extracted data and summarized the essential content of each item. YW and GHG judged which items to be excluded from the item classification survey, and which items to be included in the survey. SAO, DB, MB, LLG, PG, RJ, SK, LL, PR, KFS, DZ, YW, and GHG participated in the item classification survey, of which RJ, SK, and LL are experts in evidence-based medicine/risk of bias educators, others have expertise in risk of bias methodology. YW conducted data analysis. YW drafted the manuscript. MG, QW, LH, DZ, AI, CH, LY, MH, ZY, RC, SAO, DB, MB, LLG, PG, RJ SK, LL, PR, KFS, RACS, RBP, and GHG reviewed, revised, and approved this manuscript.

* Corresponding author. Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada. Tel.: +1 289 680 1832.

E-mail address: yingwwy@163.com (Y. Wang).

https://doi.org/10.1016/j.jclinepi.2022.10.018

0895-4356/© 2022 Elsevier Inc. All rights reserved.

risk of bias; the majority deemed 11 (16.7%) as not addressing risk of bias; and there proved substantial disagreement for 35 (53.0%) items.

Conclusion: Existing risk of bias instruments frequently include items that do not address risk of bias. For many items, experts disagree on whether or not they are addressing risk of bias. © 2022 Elsevier Inc. All rights reserved.

Keywords: Risk of bias; Methodological quality; Instrument; Randomized controlled trials; Systematic reviews; Systematic survey

1. Introduction

Randomized controlled trials (RCTs) and systematic reviews of RCTs provide the most trustworthy evidence regarding the effects of health care interventions [1,2]. Due to flaws in design and execution, the effect estimates presented in RCTs can, however, be biased.

Health researchers have long acknowledged the importance of assessing "quality" of RCTs, and have developed many instruments to address this issue [3–6]. However, in 2004 the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group initially chose the word "quality" to represent a multidimensional concept in which risk of bias is the most critical dimension [2,5]. In the last decade, researchers developing instruments addressing limitations of RCTs typically described their goal as assessing risk of bias. These risk of bias instruments have found their primary application in systematic reviews, often as one of the five critical domains for assessing certainty of evidence using the GRADE approach [7].

The conceptual coherence of instruments for assessing risk of bias of RCTs in systematic reviews depends on whether their items are restricted to those truly addressing risk of bias, rather than other domains of the GRADE system or other dimensions of quality. Although a consensus exists regarding the definition of bias—systematic deviation from the truth—those creating the instruments may not have had, as they chose items, this definition forefront in their minds. Conceptual confusion between bias and imprecision (or random error) [8,9], and particularly between bias and applicability (external validity [10] or, in the GRADE system, indirectness [11]) exists. Reporting quality of trial reports is another dimension of quality that may be confused with bias [12]. Researchers have not yet addressed the extent to which the items included in existing RCT risk of bias instruments are truly addressing risk of bias.

We have therefore conducted a systematic survey of instruments published between 2010 and 2021 assessing risk of bias of RCTs (typically used RCT risk of bias instruments were published in the last decade). Our primary objective is to document all unique items included in these instruments, and to establish whether the included items are indeed addressing risk of bias.

2. Methods

2.1. Literature search

Our search included Medline, Embase, Web of Science, and Scopus, from 2010 to October 2021. We developed the search strategy in collaboration with an experienced research librarian (Appendix 1). We also scanned reference lists of existing systematic surveys of similar topics [13–16] to identify other potentially eligible studies.

2.2. Study selection

Pairs of reviewers independently screened titles and abstracts followed by full texts. They resolved disagreement by discussion or by consultation with an adjudicator.

Eligible instruments were published in or after 2010 and met the following criteria:

(1) explicitly assessed risk of bias or internal validity of RCTs, or
(2) assessed quality or methodological quality of RCTs (included terms such as quality, methodological quality, critical appraisal, trial checklist, or methodological checklist).

We did not apply any restrictions on the types of RCTs (individually randomized parallel-group trials, cluster-randomized parallel-group trials, or individually randomized crossover trials) that the instruments aimed to assess. Eligible instruments could assess risk of bias of both RCTs and other study designs, but our data extraction and analysis focused only on items appropriate for RCTs. We included only instruments focusing on RCTs of health care topics. We included both instruments that could apply to all health care topics and instruments that specifically targeted a single health care topic (e.g., psychotherapy).

2.3. Data extraction

For each eligible instrument, one reviewer (Y.W) extracted data using a standardized, predesigned data extraction form, and a second reviewer (D.Z) independently checked the abstraction.

We abstracted the following information from each instrument:

What is new?

Key findings

- From the 17 instruments published in the last decade assessing risk of bias of RCTs, we identified 127 unique items of which we deemed 61 as clearly not addressing risk of bias. Of the remaining 66 items, the majority of respondents deemed 20 items (30.3%) as addressing risk of bias; the majority deemed 11 (16.7%) as not addressing risk of bias; and there proved substantial disagreement for 35 (53.0%) items.

What this adds to what was known?

- Existing RCT risk of bias instruments include many items that in fact address other issues including applicability, random error, and reporting quality. The instruments include many items in which experts disagree on the optimal characterization across these categories, with the most vexing classification issues arising in differentiating between risk of bias and applicability.

What is the implication and what should change now?

- Existing instruments may not be suitable for assessing risk of bias of RCTs in systematic reviews under the Grading of Recommendations Assessment, Development, and Evaluation framework. A new instrument specifically developed for assessing risk of bias in the context of systematic reviews may prove valuable. Disagreement regarding classification of risk of bias vs. other domains, in particular applicability, remains. Thus, achieving consensus is likely to prove elusive, at least in the short term. Acknowledging the legitimacy of alternative viewpoints may prove a more productive way forward.

- Name/title of the instrument and development organization
- Target study design of the instrument (only RCTs, or both RCTs and other study designs; all RCTs, or only individually randomized parallel-group trials, cluster-randomized parallel-group trials, or individually randomized crossover trials)
- Target health care topic of the instrument or generic instrument
- Originally developed or adapted from another instrument
- Stated objective of the instrument
- Target user of the instrument
- Whether risk of bias assessment is outcome specific
- Items and their response options as well as elaboration or examples

2.4. Data synthesis and analysis

One reviewer (Y.W), after reviewing all related information regarding the item (the item itself, response options, elaboration or examples), summarized the essential content of each item in each instrument. A second reviewer (D.Z) independently checked the summarization. They solved disagreement by consultation with an adjudicator (G.H.G). The process included decisions as to which similarly worded items are addressing the same underlying concept and which are not. Result of this process is a list of all unique items included in these instruments.

Two reviewers (Y.W and G.H.G) screened the list of all unique items included in the instruments and judged: which items are clearly not addressing risk of bias or clearly irrelevant—these items were excluded from a following item classification survey; the remaining items were included in the survey.

We assembled a committee with 8 individuals (7 with expertise in the methodology of risk of bias assessment and 1 with expertise in evidence-based medicine/risk of bias education). To complement the committee, we invited 12 other experts to join the panel—10 risk of bias methodological experts and 2 experts in evidence-based medicine/risk of bias education, which were randomly selected from a list of risk of bias methodological experts and a list of internationally recognized experts in evidence-based medicine education, stratified by gender and region.

We collected the potential risk of bias methodological experts from author lists of risk of bias methodological papers. Methodological papers in which authors stated explicitly or implicitly indicated that what they addressed was a risk of bias issue, were eligible. We identified risk of bias methodological papers from four resources: references of RCT risk of bias instruments; references of guidance documents describing the use of RCT risk of bias instruments; eligible papers that are volunteered by the committee; references of eligible papers that are identified from above three resources. Eligible risk of bias methodological experts should be the first or the last or the corresponding author of at least one eligible paper and coauthor of at least two other eligible papers.

Overall the panel included 17 individuals with expertise in the methodology of risk of bias assessment and 3 individuals with expertise in risk of bias education. We invited these panel members to participate in the item classification survey, of whom 13 (10 experts and 3 educators) participated.

In this survey, respondents judged the issue addressed by each item: risk of bias, applicability, random error, reporting quality, or none of the above. For each item, we counted the number of respondents that chose each option, and

classified items into three categories (we chose the thresholds prior to receiving responses):

Category 1: items that the majority thought are addressing risk of bias (i.e., ≥10 respondents chose risk of bias).

Category 2: items that the majority thought are not addressing risk of bias (i.e., ≤3 respondents chose risk of bias).

Category 3: items with substantial disagreement regarding whether they are addressing risk of bias (i.e., 4-9 respondents chose risk of bias).

3. Results

3.1. Search results

Figure 1 presents the study selection process. From database searches, after removing duplicates, we screened titles and abstracts for 14,960 records. We reviewed full texts for 162 records, of which 10 [1,17–25] proved eligible. We identified 7 other eligible instruments [26–32] from systematic surveys of similar topics. We ultimately included 17 instruments (Table 1), of which 11 [1,17,18,21,23,24,26,29–32] explicitly assessed risk of bias or internal validity of RCTs, and 6 [19,20,22,25,27,28] assessed quality or methodological quality of RCTs.

3.2. Characteristics of included instruments

Appendix 2 presents characteristics of the included instruments. Of the 17 instruments, 12 [1,19–24,27–31] addressed only RCTs and 5 [17,18,25,26,32] addressed both RCTs and nonrandomized studies. All instruments included items for individually randomized parallel-group trials; 4 [17,23,25,30] additionally included items for individually randomized crossover trials; 2 [23,30] included items for cluster-randomized parallel-group trials; and 1 [30] included an item for stepped-wedge RCTs. Seven instruments addressed one specific health care topic (pharmacist interventions [24], drug adverse events [17], psychotherapy [19], epidural infusion analgesia [20], behavioral interventions [18], exercise training [22], and interventional pain management techniques [21]), while the other 10 were not specific to a health care topic. Seven instruments [1,18–20,23,25,30] were developed independent of any

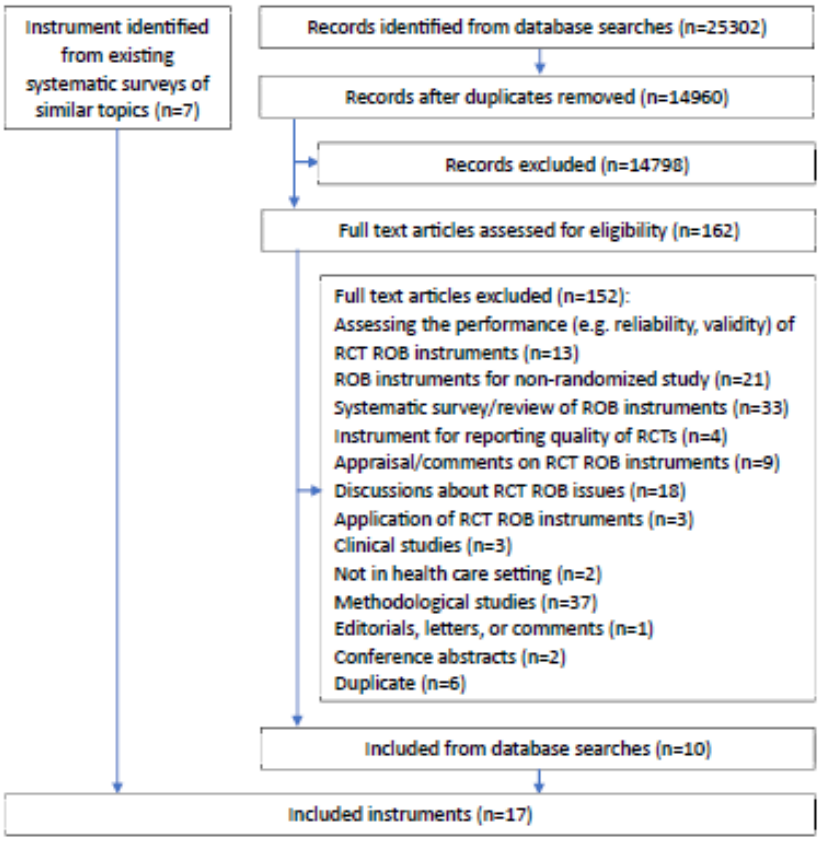

Fig. 1. Study selection flow chart.

Table 1. Included instruments

Study ID	Name of instrument	Development organization
CASP 2020 [28]	CASP Randomized Controlled Trial Standard Checklist	CASP
CLARITY Group tool [29]	Tool to Assess Risk of Bias in Randomized Controlled Trials	CLARITY group
EPOC 2017 [26]	Suggested risk of bias criteria for EPOC reviews	EPOC
Faillie 2017 [17]	Risk of Bias Assessment Checklist for Studies Included in Systematic Reviews of Drug Adverse Events	NR
Higgins 2011 [1]	Cochrane Collaboration's risk of bias assessment tool	Cochrane Bias Methods Group, Cochrane Statistical Methods Group
JBI 2020 [30]	JBI Critical Appraisal Checklist for Randomized Controlled Trials	JBI
Kennedy 2019 [18]	Evidence Project risk of bias tool	Evidence Project
Kocsis 2010 [19]	RCT-PQRS	American Psychiatric Association Committee on Research on Psychiatric Treatments
Liu 2010 [20]	EATC	NR
Manchikanti 2014 [21]	IPM-QRB	NR
NICE 2012 [31]	Methodology checklist: randomised controlled trials	NICE
NIHBL 2013 [32]	Quality Assessment of Controlled Intervention Studies	NIHBL
SIGN checklist [27]	SIGN methodology checklist 2: randomized controlled trials	SIGN
Smart 2015 [22]	TESTEX	NR
Sterne 2019 [23]	Revised Cochrane risk-of-bias tool 2.0	Cochrane group
Stone 2021 [25]	MASTER Scale	NR
Tonin 2019 [24]	Guide for Risk of Bias Judgment in Pharmacy Services	NR

CASP, Critical Appraisal Skills Programme; CLARITY, Clinical Advances Through Research and Information Translation; EATC, Epidural Analgesia Trial Checklist; EPOC, Cochrane Effective Practice and Organization of Care; IPM-QRB, Interventional Pain Management Techniques Quality Appraisal of Reliability and Risk of Bias Assessment tool; JBI, Joanna Briggs Institute; MASTER, MethodologicAl STandards for Epidemiological Research; NICE, National Institute for Health and Care Excellence; NIHBL, National Heart, Lung, and Blood Institute; NR, not reported; RCT-PQRS, RCT of Psychotherapy Quality Rating Scale; SIGN, Scottish Intercollegiate Guidelines Network; TESTEX, Tool for the assEssment of Study qualiTy and reporting in EXercise.

other instrument; the other 10 were adapted from another instrument or guidance or included items from other instruments. Twelve instruments [1,17,18,21–26,29,30,32] specified or implied that the target users are systematic reviewers; the others did not make clear the target audience.

Seven instruments [1,23–26,29,31] specified that risk of bias assessment should be outcome specific (assess risk of bias for each outcome, or for objective outcomes and subjective outcomes separately). The Cochrane risk of bias tool 2.0 [23] further suggested to assess risk of bias for each single effect estimate of interest.

3.3. Items and what the items actually addressed

Figure 2 presents the process involved in identifying and categorizing items and its results. From these instruments we identified 127 unique items, of which we judged 61 as clearly not addressing risk of bias (e.g., the item "whether the study addresses a clearly focused research question") or irrelevant (either too broad or vague) (e.g., the item "statistical methodology") and thus were excluded from further consideration (Appendix 3). The survey included the remaining 66 items (Appendix 4).

Appendix 4 presents the item classification survey results. The majority deemed 20 items (30.3%) as addressing risk of bias (category 1) (e.g., the item "random sequence generation"); 11 items (16.7%) as not addressing risk of bias (category 2) (e.g., the item "whether the sample size is big enough"); there was substantial disagreement on the panel as to whether the remaining 35 items (53.0%) addressed risk of bias (category 3) (e.g., the item "whether the co-interventions balance between groups"). Of the 35 items on which there was substantial disagreement, the disagreement was primarily between risk of bias and applicability in 15; between multiple categories in 14; between risk of bias and random error in 3; and between risk of bias and none of the other labeled categories in 3. Sixteen instruments [1,17–22,24–32] included at least one item that clearly did not address risk of bias or had items deemed by the majority of survey respondents as not addressing risk of bias. Among the 11 instruments [1,17,18,21,23,24,26,29–32] that explicitly assessed risk of bias or internal validity, 10 [1,17,18,21,24,26,29–32] included items that clearly did not address risk of bias or had items deemed as not addressing risk of bias by the majority. The same was true of all six instruments

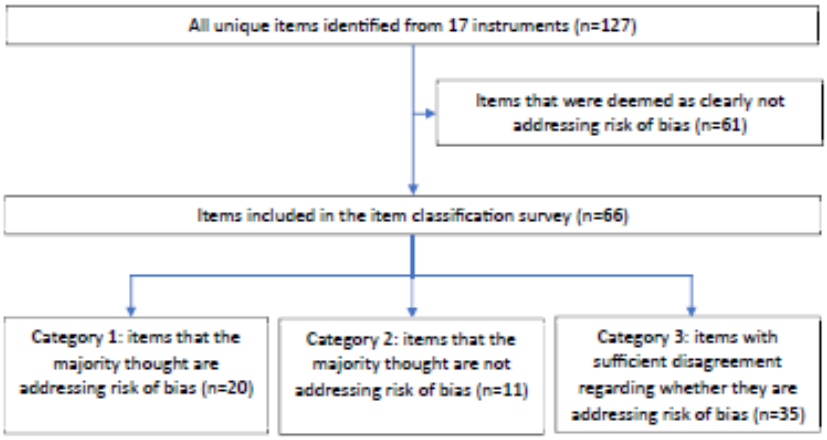

Fig. 2. Item selection and classification process.

[19,20,22,25,27,28] assessing the quality or the methodological quality of RCTs. The Cochrane risk of bias tool 2.0 [23] proved the only instrument without items that are clearly not addressing the risk of bias.

3.4. Response options for single items

For response options for single items, 3 instruments [1,24,26] used "low risk of bias", "high risk of bias", and "unclear risk of bias"; 7 [17,18,27,28,30–32] used "yes", "no", and "unclear/can't tell/others/not applicable/not reported"; 2 [23,29] used "definitely yes", "probably yes", "probably no", and "definitely no"; 5 [19–22,25] used numbers with specification of what each number means (Appendix 5).

4. Discussion

From the 17 instruments published in the last decade assessing risk of bias of RCTs, we identified 127 unique items of which we deemed about half of the items as clearly not addressing risk of bias or irrelevant. A majority of a panel of 10 experts in risk of bias methodology and 3 experts in risk of bias education deemed 20 (30.3%) of the remaining 66 items as addressing risk of bias and 11 (16.7%) as not addressing the risk of bias. The panel expressed appreciably divergent views regarding the remaining 35 (53.0%) items.

Although 12 instruments specified or implied that their target users are systematic reviewers, it is unclear whether they have had this target forefront in their minds when they developed the instruments and chose items. Ideally to be suitable for use in the context of systematic reviews that apply the GRADE approach, risk of bias instruments should include only items addressing risk of bias rather than other GRADE domains such as imprecision [9] or applicability (indirectness) [10]. Our results demonstrated that uncertainty remains regarding whether issues are best classified as risk of bias vs. precision or, particularly, applicability. Varying perspective exists in differentiating risk of bias from applicability in items related to co-interventions, failure in implementing the intended interventions, and nonadherence to the assigned intervention.

Because risk of bias can differ between outcomes, the risk of bias assessment should be outcome specific. Seven instruments specified that risk of bias assessment should be outcome specific. Most instruments did not address this issue specifically.

Risk of bias assessment involves two steps: first, to determine what happened in the trial, and then to judge the extent to which any limitation is likely to increase bias. Instrument addressing risk of bias could involve an initial assessment of the first step with, for instance, the second step in which a systematic review team would decide whether to rate down for risk of bias across all studies. Alternatively individuals completing the instrument could make their judgments of the magnitude of risk of bias associated with each item for the outcome under consideration. The 3 instruments using "low risk of bias", "high risk of bias", and "unclear risk of bias" as response options for single items [1,24,26], asked individuals completing the instrument to make a judgment regarding the magnitude of risk of bias for each item. Of the nine instruments using "definitely yes", "probably yes", "probably no", and "definitely no" [23,29], or using "yes", "no", and "unclear"/"can't tell"/"others"/"not applicable"/"not reported" as response options [17,18,27,28,30–32], 7 are consistent in determining only what happened in the trial [18,23,27,28,30–32]. The associated elaborations for the other two instruments [17,29], however, suggest the intent is rating magnitude of risk of bias.

Previous authors have conducted systematic surveys of instruments for assessing methodological quality or risk of bias of RCTs [13–16,33]. Two studies identified existing

instruments for assessing methodological quality of different types of studies including RCTs [14,16]. One study identified instruments evaluating methodological quality of RCTs in health care research and summarized their content, construction, development, and psychometric properties [15]. None of these studies, however, addressed details regarding the included items.

One study identified instruments evaluating methodological quality or risk of bias of RCTs in general health and physical therapy research, summarized their psychometric properties, and classified their included items as evaluating "reporting" vs. "conduct" and addressing different types of bias vs. imprecision [13]. In the item classification of evaluating "reporting" vs. "conduct", two reviewers classified 48% of the items included in general health research instruments and 33% of the items in physical therapy research instruments as evaluating "reporting" rather than "conduct". In the item classification of addressing bias vs. imprecision, they classified 3.1% of the items in general health research instruments and 4.2% of the items in physical therapy research instruments as addressing imprecision rather than bias. The authors updated the work, only focusing on the rehabilitation area [33]. These are the only previous studies that addressed details regarding the items. However, only two reviewers made the judgment, and they omitted applicability—a particularly important domain that needs to be distinguished from risk of bias.

Strengths of this study include the comprehensive search, transparent eligibility criteria, and thorough extraction and summary of information regarding the items. We are the first to conduct an item classification survey which included a systematically assembled panel of risk of bias methodological experts and educators. In the survey we distinguished items addressing risk of bias from addressing other easily confused concepts. Notably, we are the first to distinguish risk of bias from applicability in the item classification scheme.

Our study has several limitations. First, the panel making judgments regarding whether items represented risk of bias included only 13 individuals. A larger group of both experts and educators would have yielded a broader representation of opinion. Second, one might perceive our restriction to instruments published in the last decade as a limitation. This restriction, however, ensures that the instruments benefited from relatively current thinking regarding risk of bias assessment.

In conclusion, existing instruments that intended, at least in part, to assess risk of bias, include many items that in fact address other issues including applicability, random error, and reporting quality. Thus, existing instruments may not be suitable for assessing risk of bias in systematic reviews under the GRADE framework. A new instrument specifically developed for assessing risk of bias in the context of systematic reviews would likely prove valuable.

Another notable finding of our work is that existing instruments include many items in which experts disagree on the optimal characterization across these categories, with the most vexing classification issues arising in differentiating between risk of bias and applicability. There are clearly gray areas in which one might reasonably consider an issue as representing a risk of bias or applicability concern. How one deals with nonadherence to an intervention, or a control group receiving the intervention of interest, represent two such areas.

One might reasonably argue that moving forward in this area requires establishing a consensus on classification of risk of bias vs. issues such as applicability/directness. Doing so may involve coming to agreement regarding the concept of risk of bias, and differences between assessing risk of bias in the context of systematic reviews vs. single RCT. The problem with seeking such a consensus is that, for several such issues, alternative positions are reasonable, and possibly equally reasonable. In such situations, achieving consensus may not be a realistic or appropriate objective in the short term.

An alternative to seeking consensus is that authors providing methodologic guidance—be it in the form of a risk of bias instrument or other guidance—highlight areas of controversy, clearly state their position on the controversy, and acknowledge the matter remains a legitimate issue of doubt or dispute. Systematic review authors following methodologic guidance—through a specific risk of bias instrument, through GRADE guidance, or other guidance—will then achieve a balanced understanding of the "gray area" problem. They can then decide what position they find most compelling, and conduct their reviews accordingly.

Appendix A

Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2022.10.018.

References

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

[2] Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs included in systematic reviews. J Clin Epidemiol 2001;54:651–4.

[3] Dechartres A, Charles P, Hopewell S, Ravaud P, Altman DG. Reviews assessing the quality or the reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. J Clin Epidemiol 2011;64:136–44.

[4] Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled Clin trials 1996;17:1–12.

[5] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Controlled Clin trials 1995;16:62–73.

[6] Verhagen AP, de Vet HC, de Bie RA, Kessels AG, Boers M, Bouter LM, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 1998;51:1235–41.

[7] 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.

[8] Altman DG, Dore CJ. Baseline comparisons in randomized clinical trials. Stat Med 1991;10:797–9.

[9] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence–imprecision. J Clin Epidemiol 2011;64:1283–93.

[10] Sedgwick P. Randomised controlled trials: internal versus external validity. BMJ 2014;348:g1742.

[11] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence–indirectness. J Clin Epidemiol 2011;64:1303–10.

[12] Juni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42–6.

[13] Armijo-Olivo S, Fuentes J, Ospina M, Saltaji H, Hartling L. Inconsistency in the items included in tools used in general health research and physical therapy to evaluate the methodological quality of randomized controlled trials: a descriptive analysis. BMC Med Res Methodol 2013;13:116.

[14] Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res 2020;7:7.

[15] Olivo SA, Macedo LG, Gadotti IC, Fuentes J, Stanton T, Magee DJ. Scales to assess the quality of randomized controlled trials: a systematic review. Phys Ther 2008;88:156–75.

[16] Zeng X, Zhang Y, Kwong JS, Zhang C, Li S, Sun F, et al. The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review. J Evid Based Med 2015;8:2–10.

[17] Faillie JL, Ferrer P, Gouverneur A, Driot D, Berkemeyer S, Vidal X, et al. A new risk of bias checklist applicable to randomized trials, observational studies, and systematic reviews was developed and validated to be used for systematic reviews focusing on drug adverse events. J Clin Epidemiol 2017;86:168–75.

[18] Kennedy CE, Fonner VA, Armstrong KA, Denison JA, Yeh PT, O'Reilly KR, et al. The Evidence Project risk of bias tool: assessing study rigor for both randomized and non-randomized intervention studies. Syst Rev 2019;8:3.

[19] Kocsis JH, Gerber AJ, Milrod B, Roose SP, Barber J, Thase ME, et al. A new scale for assessing the quality of randomized clinical trials of psychotherapy. Compr Psychiatry 2010;51:319–24.

[20] Liu SS, Togioka BM, Hurley RW, Vu CM, Hanna MN, Murphy JD, et al. Methodological quality of randomized controlled trials of postoperative epidural analgesia: validation of the Epidural Analgesia Trial Checklist as a specific instrument to evaluate methodology. Reg Anesth Pain Med 2010;35:549–55.

[21] Manchikanti L, Hirsch JA, Cohen SP, Heavner JE, Falco FJ, Diwan S, et al. Assessment of methodologic quality of randomized trials of interventional techniques: development of an interventional pain management specific instrument. Pain Physician 2014;17:E263–90.

[22] Smart NA, Waldron M, Ismail H, Giallauria F, Vigorito C, Cornelissen V, et al. Validation of a new tool for the assessment of study quality and reporting in exercise training studies: TESTEX. Int J Evid Based Healthc 2015;13:9–18.

[23] Sterne JAC, Savovic 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.

[24] Tonin FS, Lopes LA, Rotta I, Bonetti AF, Pontarolo R, Correr CJ, et al. Usability and sensitivity of the risk of bias assessment tool for randomized controlled trials of pharmacist interventions. Int J Clin Pharm 2019;41:785–92.

[25] Stone JC, Glass K, Clark J, Ritskes-Hoitinga M, Munn Z, Tugwell P, et al. The MethodologicAl STandards for Epidemiological Research (MASTER) scale demonstrated a unified framework for bias assessment. J Clin Epidemiol 2021;134:52–64.

[26] Cochrane Effective Practice and Organisation of Care (EPOC). Suggested risk of bias criteria for EPOC reviews. Available at https://epoc.cochrane.org/resources/epoc-resources-review-authors. Accessed August 28, 2022.

[27] Scottish Intercollegiate Guidelines Network (SIGN). Methodology checklist 2: controlled trials. Available at https://www.sign.ac.uk/what-we-do/methodology/checklists/. Accessed August 28, 2022.

[28] Critical Appraisal Skills Programme (CASP). CASP randomised controlled trial standard checklist. Available at https://casp-uk.net/casp-tools-checklists/. Accessed August 28, 2022.

[29] Tool to Assess Risk of Bias in Randomized Controlled Trials. Contributed by the CLARITY group at McMaster University. Available at https://www.evidencepartners.com/wp-content/uploads/2021/03/Tool-to-Assess-Risk-of-Bias-in-Randomized-Controlled-Trials-DistillerSR.pdf. Accessed August 28, 2022.

[30] Joanna Briggs Institute (JBI). JBI critical appraisal checklist for randomized controlled trials. Available at https://joannabriggs.org/critical-appraisal-tools. Accessed August 28, 2022.

[31] National Institute for Health and Care Excellence (NICE). Methodology checklist: randomised controlled trials. Available at https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-2549703709/chapter/appendix-c-methodology-checklist-randomised-controlled-trials. Accessed August 28, 2022.

[32] National Heart, Lung, and Blood Institute (NIHBL). Quality assessment of controlled intervention studies. Available at https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools. Accessed August 28, 2022.

[33] Armijo-Olivo S, Patrini M, Oliveira-Souza AIS, Dennett L, Arienti C, Dahchi M, et al. Tools to assess the risk of bias and reporting quality of randomized controlled trials in rehabilitation. Arch Phys Med Rehabil 2021;102:1606–13.

Appendix 1. Search strategy

MEDLINE
Database: OVID Medline Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present
Search Strategy:
---

1 (risk adj3 bias).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
2 methodological quality.mp.
3 internal* valid*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
4 systematic error.mp.
5 ((assess* or evaluat* or apprais* or measur*) adj3 (quality or valid*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
6 or/1-5
7 critical* apprais*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
8 ((assess* or evaluat* or apprais* or methodolog*) adj3 (scale* or tool* or checklist* or instrument* or criteri*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
9 7 or 8
10 6 and 9
11 exp Randomized Controlled Trial/
12 exp Randomized Controlled Trials as Topic/
13 (randomi?ed adj3 trial*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, floating sub-heading word, keyword heading word, organism supplementary concept word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]
14 RCT.mp.
15 or/11-14
16 10 and 15
17 6 and 9 and 15
18 limit 17 to yr="2010 -Current"

Embase
Search Strategy:
---

1 (risk adj3 bias).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
2 methodological quality.mp.
3 internal* valid*.mp.
4 systematic error.mp. or exp systematic error/
5 ((assess* or evaluat* or apprais* or measur*) adj3 (quality or valid*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
6 or/1-5
7 critical* apprais*.mp.
8 ((assess* or evaluat* or apprais* or methodolog*) adj3 (scale* or tool* or checklist* or instrument* or criteri*)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
9 7 or 8
10 6 and 9
11 exp randomized controlled trial/
12 (randomi?ed adj3 trial*).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word, candidate term word]
13 RCT.mp.
14 or/11-13
15 10 and 14
16 6 and 9 and
17 limit 16 to yr="2010 -Current"

Web of Science (Clarivate)

# 17	#16 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=2010-2021*	Edit
# 16	#14 AND #9 AND #6 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 15	#14 AND #10 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 14	#13 OR #12 OR #11 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 13	TS=RCT *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 12	TS=Randomized Controlled Trial *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI*	Edit

	Timespan=1976-2021	
# 11	TS=(randomi?ed near/3 trial*) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 10	#9 AND #6 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 9	#8 OR #7 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 8	TS=((assess* or evaluat* or apprais* or methodolog*) near/3 (scale* or tool* or checklist* or instrument* or criteri*)) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 7	TS=critical* apprais* *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 6	#5 OR #4 OR #3 OR #2 OR #1 *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 5	TS=((assess* or evaluat* or apprais* or measur*) Near/3 (quality or valid*)) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 4	TS=systematic error *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 3	TS=internal* valid* *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 2	TS=methodological quality *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit
# 1	ts=(risk near/3 bias) *Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI* *Timespan=1976-2021*	Edit

Scopus

(("randomised controlled trial") OR (rct)) AND ((("critical appraisal" OR "critically appraise") OR (TITLE ((assess* OR evaluat* OR apprais* OR methodolog*) W/3 (scale* OR tool* OR checklist* OR instrument* OR criteri*)))) AND ((risk W/3 bias) OR ("methodological quality") OR ("internal validity" OR "internally valid") OR ("systematic error") OR (TITLE

((assess* OR evaluat* OR apprais* OR measur*) W/3 (quality OR valid*))))) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010))

Appendix 2. Characteristics of included instruments

Study ID	Target study design	Target health care topic	Originally developed or adapted from another instrument 1=originally developed 2=adapted from another instrument	Objective of the instrument	Target user of the instrument	Whether risk of bias assessment is outcome specific
CASP 2020 [1]	RCTs	Comprehensive	2 (The CASP RCT checklist was originally based on JAMA Users' guides to the medical literature 1994, and piloted with healthcare practitioners. This version has been updated taking into account the CONSORT 2010 guideline)	"11 questions to help you make sense of a RCT"	NR	NR
CLARITY Group tool [2]	RCTs	Comprehensive	2 (Modified from Cochrane risk of bias tool 1.0)	Assess risk of bias of RCTs	Systematic reviewers	Yes
EPOC 2017 [3]	RCTs, non-randomized trials, controlled before-after studies	Comprehensive	2 (Further information can be obtained from the Cochrane handbook)	NR	Systematic reviewers	Yes
Faillie 2017 [4]	RCTs, cohort studies, case-control studies, nested case-control studies, systematic reviews	Drug adverse events	2 (In the pilot version of the checklist, the items and domains were derived from previous quality assessment instruments (Downs and Black scale, Cochrane risk of bias tool, and AMSTAR), reviews of the literature, and recommendations from the Agency for Healthcare Research and Quality)	"The developed checklist aims to assess the risk of bias for the different types of study used to assess drug safety"	Systematic reviewers	NR
Higgins 2011 (Cochrane risk of bias tool 1.0) [5]	RCTs	Comprehensive	1	"Assess the potential for bias in RCTs included in systematic reviews or meta-analyses"	Systematic reviewers	Yes
JBI 2020 [6]	Individual participants parallel groups RCTs, crossover RCTs, cluster RCTs, stepped-wedge RCTs	Comprehensive	1	"Assess the methodological quality of a study and determine the extent to which a study has addressed the possibility of bias in its design,	Systematic reviewers	NR

				conduct and analysis"		
Kennedy 2019 [7]	RCTs, non-randomized intervention studies	Behavioral interventions	1	"A tool for assessing risk of bias in both randomized and non-randomized intervention studies"	Systematic reviewers	No
Kocsis 2010 [8]	RCTs	Psychotherapy	1	"A rating scale designed to assess the quality of RCTs of psychotherapy"	NR	NR
Liu 2010 [9]	RCTs	Postoperative epidural analgesia	1	"It is a grading tool for pain studies that was designed to be specific for RCTs that have epidural infusion analgesia as a primary intervention for randomization."	NR	No
Manchikanti 2014 [10]	RCTs	Interventional pain management techniques	2 (The final list of items included 9 of 12 items from Cochrane review criteria and 13 new items)	"Our objective was to develop an instrument specifically for interventional pain management, to assess the methodological quality of randomized trials of interventional techniques"	Systematic reviewers	NR
NICE 2012 [11]	RCTs	Comprehensive	NR	"This checklist is designed to assess the internal validity of the study; that is, whether the study provides an unbiased estimate of what it claims to show"	NR	Yes
NIHBL 2013 [12]	Controlled intervention studies	Comprehensive	2 (The tools were based on quality assessment methods, concepts, and other tools developed by researchers in the Agency for Healthcare Research and Quality's Evidence-based Practice Centers, the Cochrane Collaborative, the USPSTF, the National Health Service Centre for Reviews and Dissemination)	"The questions on the assessment tool were designed to help reviewers focus on the key concepts for evaluating a study's internal validity"	Systematic reviewers	NR
SIGN checklist	RCTs	Comprehensive	NR	NR	NR	NR

[13]						
Smart 2015 [14]	RCTs	Exercise training trials	2 (The group used the PEDro scale as a template for the development of a new scale)	"We have developed the TESTEX scale as an exercise sciences-specific scale, designed…to assess the study quality and reporting of randomized controlled trials of exercise training"	Systematic reviewers	NR
Sterne 2019 (Cochrane risk of bias tool 2.0) [15]	Individually-randomized parallel-group trials, cluster-randomized parallel-group trials, individually randomized crossover trials	Comprehensive	1	NR	Systematic reviewers	Yes (risk of bias should be done for each effect estimate)
Stone 2021 (MASTER scale) [16]	All analytic study designs of exposures or interventions	Comprehensive	1	"The MASTER scale presents a unified framework for assessment of methodological quality/bias assessment of multiple analytic study designs within a systematic review"; "The MASTER scale contains a list of major methodological safeguards that should be present in epidemiological studies"	Systematic reviewers	Yes
Tonin 2019 [17]	RCTs	Pharmacist interventions	2 (Adapted from the Cochrane Handbook)	"A Guide for risk of bias Judgment in Pharmacy Services will help in the interpretation and judgment of bias criteria in randomized controlled trials of clinical pharmacy interventions"	Systematic reviewers	Yes

CASP=Critical Appraisal Skills Programme; RCT=randomized controlled trial; JAMA=Journal of the American Medical Association; CONSORT=Consolidated Standards of Reporting Trials; NR=not reported; NA=not applicable; CLARITY=Clinical Advances Through Research and Information Translation; EPOC=Cochrane Effective Practice and Organisation of Care; AMSTAR=Assessing the Methodological Quality of Systematic Reviews; JBI=Joanna Briggs Institute; NICE=National Institute for Health and Care Excellence; NIHBL=National Heart, Lung, and Blood Institute; USPSTF=United States Preventive Services Task Force; SIGN=Scottish Intercollegiate Guidelines Network; PEDro=Physiotherapy Evidence Database; TESTEX=Tool for the assEssment of Study qualiTy and reporting in Exercise; MASTER=MethodologicAl STandards for Epidemiological Research.

References

1. Critical Appraisal Skills Programme (CASP). CASP Randomised Controlled Trial Standard Checklist. https://casp-uk.net/casp-tools-checklists/.
2. Tool to Assess Risk of Bias in Randomized Controlled Trials. Contributed by the CLARITY Group at McMaster University. https://www.evidencepartners.com/wp-content/uploads/2021/03/Tool-to-Assess-Risk-of-Bias-in-Randomized-Controlled-Trials-DistillerSR.pdf.
3. Cochrane Effective Practice and Organisation of Care (EPOC). Suggested risk of bias criteria for EPOC reviews. https://epoc.cochrane.org/resources/epoc-resources-review-authors.
4. Faillie JL, Ferrer P, Gouverneur A, Driot D, Berkemeyer S, Vidal X, et al. A new risk of bias checklist applicable to randomized trials, observational studies, and systematic reviews was developed and validated to be used for systematic reviews focusing on drug adverse events. J Clin Epidemiol 2017;86:168-75.
5. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
6. Joanna Briggs Institute (JBI). JBI Critical Appraisal Checklist for Randomized Controlled Trials. https://joannabriggs.org/critical-appraisal-tools.
7. Kennedy CE, Fonner VA, Armstrong KA, Denison JA, Yeh PT, O'Reilly KR, et al. The Evidence Project risk of bias tool: assessing study rigor for both randomized and non-randomized intervention studies. Systematic reviews 2019;8:3.
8. Kocsis JH, Gerber AJ, Milrod B, Roose SP, Barber J, Thase ME, et al. A new scale for assessing the quality of randomized clinical trials of psychotherapy. Comprehensive psychiatry 2010;51:319-24.
9. Liu SS, Togioka BM, Hurley RW, Vu CM, Hanna MN, Murphy JD, et al. Methodological quality of randomized controlled trials of postoperative epidural analgesia: validation of the Epidural Analgesia Trial Checklist as a specific instrument to evaluate methodology. Reg Anesth Pain Med 2010;35:549-55.
10. Manchikanti L, Hirsch JA, Cohen SP, Heavner JE, Falco FJ, Diwan S, et al. Assessment of methodologic quality of randomized trials of interventional techniques: development of an interventional pain management specific instrument. Pain Physician 2014;17:E263-90.
11. National Institute for Health and Care Excellence (NICE) Methodology checklist: randomised controlled trials. https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-2549703709/chapter/appendix-c-methodology-checklist-randomised-controlled-trials.
12. National Heart, Lung, and Blood Institute (NIHBL). Quality Assessment of Controlled Intervention Studies. https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools.
13. Scottish Intercollegiate Guidelines Network (SIGN). Methodology Checklist 2: Controlled Trials. https://www.sign.ac.uk/what-we-do/methodology/checklists/.
14. Smart NA, Waldron M, Ismail H, Giallauria F, Vigorito C, Cornelissen V, et al. Validation of a new tool for the assessment of study quality and reporting in exercise training studies: TESTEX. International journal of evidence-based healthcare 2015;13:9-18.
15. Sterne JAC, Savovic 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.
16. Stone JC, Glass K, Clark J, Ritskes-Hoitinga M, Munn Z, Tugwell P, et al. The MethodologicAl STandards for Epidemiological Research (MASTER) scale demonstrated a unified framework for bias assessment. J Clin Epidemiol 2021;134:52-64.
17. Tonin FS, Lopes LA, Rotta I, Bonetti AF, Pontarolo R, Correr CJ, et al. Usability and sensitivity of the risk of bias assessment tool for randomized controlled trials of pharmacist interventions. Int J Clin Pharm 2019;41:785-92.

Appendix 3. Items that were deemed as clearly not addressing risk of bias or irrelevant, thus were excluded from the item classification survey (61 items)
Other bias
Whether the study addresses a clearly focused research question
Whether the effects of intervention are reported comprehensively
Whether the precision of the estimate of the intervention or treatment effect is reported
Do the benefits of the experimental intervention outweigh the harms and costs
Can the results be applied to your local population/in your context
Whether study objective is clearly specified and appropriate
Whether the number of participants is clearly reported throughout the study
Whether the number of drop-outs/withdrawals due to drug safety outcome clearly stated for each treatment arm
Whether the definition of the drug safety outcome is clearly stated
Whether the severity of the drug safety outcome clearly stated
Whether the method for ascertaining the drug safety outcome is adequately constructed and equal for all participants
Whether the number of drug adverse events and the number of patients with a drug adverse event is reported in both treatment arms
Whether the time frequency of drug safety outcome assessment during the follow-up period is appropriate
Whether the time between the exposure to a drug and the onset of the adverse event is reported
Whether the process of determining that the adverse event is linked to the drug is appropriate
Blinding of the process of determining the adverse event is linked to the drug
Whether the statistical methods used to analyze the drug safety outcome appropriate
Whether a survival analysis performed when there are individual differences in length of follow-up
Whether the composite outcome of drug safety adequately constructed
Whether the results are consistent in primary and secondary analyses
Reliability of diagnostic methodology
Description of relevant comorbidities
Description of numbers of subjects screened, included, and excluded
Treatments (including control/comparison groups) are sufficiently described or referenced to allow for replication

Treatment being studied is treatment being delivered
Therapist training and level of experience in the treatment(s) under investigation
Therapist supervision while treatment is being provided
Description of concurrent treatments
Discussion of safety and adverse events during study treatment(s)
Assessment of long-term post-termination outcome
Appropriate consideration of therapist and site effects
A priori relevant hypotheses that justify comparison group(s)
Comparison groups from same population and time frame as experimental group
Balance of allegiance to types of treatment by practitioners
Conclusions of study justified by sample, measures, and data analysis, as presented
Precise definition of outcome
Confirmed presence of functioning catheter (e.g. injection of LA for sensory level)
Epidural catheter inserted at location congruent to surgical incision (e.g. thoracic placement for thoracic or abdominal surgery)
Assessment of VAS pain at rest and with activity
Proper presentation of VAS pain scores (i.e., mean T SD for continuous VAS and median/95% confidence interval for numeric rating scale or categorical data)
Appropriate duration of epidural analgesia (e.g. postoperative day 92 [per JAMA meta-analysis] for postoperative pain studies)
Appropriate description and mention of adverse effects (opioids: postoperative nausea and vomiting, pruritus, respiratory depression, urinary retention, sedation; local anesthetic: hypotension, motor block, urinary retention)
Eligibility criteria specified
Between-group statistical comparisons reported
Point measures and measures of variability for all reported outcome measures
Activity monitoring in control groups
Relative exercise intensity remained constant
Exercise volume and energy expenditure
Trial design guidance and reporting

Type and design of trial
Setting/physician
Imaging
Statistical methodology
Inclusiveness of population - 7a. For epidural procedures:
Inclusiveness of population - 7b. For facet or sacroiliac joint interventions:
Duration of pain
Previous treatments conservative management including drug therapy, exercise therapy, physical therapy, etc.
Duration of follow-up with appropriate interventions
Computation errors or contradictions were absent
Dose of intervention/ exposure was sufficient to influence the outcome

Appendix 4. Item classification survey results

Category 1: items that the majority thought are addressing risk of bias (i.e. ≥10 respondents chose risk of bias)

Classification	Items included in the survey	Number of respondents that chose each option				
		Risk of bias	Applicability	Random error	Reporting quality	None of the above
Randomization	Random sequence generation	13	0	0	0	0
	Allocation concealment	13	0	0	0	0
Baseline differences	Whether baseline differences between groups suggest a problem with the randomization process (i.e. baseline differences that are not compatible with chance)	11	0	2	0	0
Blinding	Blinding of health care providers	13	0	0	0	0
	Blinding of participants	13	0	0	0	0
	Blinding of data collectors	13	0	0	0	0
	Blinding of outcome assessors	13	0	0	0	0
	Blinding of data analysts	13	0	0	0	0
Outcome measurement	If unblinding of outcome assessors, whether the outcome assessment is likely to be influenced by outcomes assessors' knowledge of intervention status (it depends on the observers' preconceptions and degree of judgement involved in assessing an outcome)	12	1	0	0	0
	Whether the outcome measurement (data collection) differs between groups	11	1	0	0	1
	Whether the follow-up time similar between groups	12	1	0	0	0
Missing outcome data	Whether the proportion of participants with missing outcome data is large enough to influence results	11	2	0	0	0
	Whether the reasons for missing outcome data provide evidence that missingness in the outcome relates to its true value	11	1	0	0	1
	Whether the proportion of missing outcome data is similar between groups	11	1	0	0	1
	Whether the reasons for missing outcome data differ between groups	12	1	0	0	0

Selective reporting	Whether the data that produce this result analyzed in accordance with a pre-specified analysis plan that is finalized before unblinded outcome data are available for analysis	10	0	0	3	0
Analysis	For estimating effect of assignment to intervention, whether an appropriate analysis is used (appropriate analysis is intention-to-treat analysis; inappropriate analyses include as-treated analysis, per-protocol analysis, etc.)	10	1	0	1	1
Cluster randomized controlled trials	Whether those identifying actual/ potential participants, those recruiting participants, and potential participants are aware of cluster allocation at recruitment	10	0	0	0	3
	If those identifying actual/ potential participants, those recruiting participants, or potential participants are aware of cluster allocation at recruitment, whether the selection of individual participants is likely to be affected by knowledge of the intervention assigned to the cluster	12	1	0	0	0
	Whether baseline imbalances suggest differential identification or recruitment of individual participants between intervention groups	12	1	0	0	0

Category 2: items that the majority thought are not addressing risk of bias (i.e. ≤3 respondents chose risk of bias)

Classification	Items included in the survey	Number of respondents that chose each option				
		Risk of bias	Applicability	Random error	Reporting quality	None of the above
Non-adherence to the assigned intervention	Whether there is crossover to the comparator intervention	3	6	0	0	4
Outcome measurement	Reliability of outcome measurement	3	3	7	0	0
	Whether the length of follow-up is adequate to identify the outcome of interest	2	8	0	0	3
Selective reporting	All of the study's prespecified outcomes have been reported	3	1	0	8	1
	The study report failed to include results for a key outcome that would be expected to have been reported for such a study	1	4	0	4	4
Early termination	Stop early for other reasons (e.g., lack of funds, ethical issues)	1	4	2	0	6
Others	Whether the sample size is big enough	1	1	9	0	2
	Funding	3	1	0	2	7
	Whether the results are comparable for all sites where the study is carried out	1	8	1	0	3
	Is the study design free of run-in/lead-in period before inclusion/randomization of participants	1	8	1	0	3
	Appropriateness of the inclusion and exclusion criteria	1	11	0	0	1

Category 3: items with sufficient disagreement regarding whether they are addressing risk of bias (i.e. 4-9 respondents chose risk of bias)

Classification	Items included in the survey	Number of respondents that chose each option				
		Risk of bias	Applicability	Random error	Reporting quality	None of the above
Baseline differences	Whether the baseline prognostic factors balance between groups	9	0	4	0	0
	Whether the baseline outcome measurements balance between groups	8	0	5	0	0
Co-interventions	Whether the co-interventions balance between groups	8	1	3	0	1
	For estimating effect of assignment to intervention, if unblinding of participants or health care providers, whether there is occurrence of co-interventions that are inconsistent with the trial protocol which affects outcome that arise because of the trial context	8	2	0	0	3
	For estimating effect of assignment to intervention, if unblinding of participants or health care providers, whether the co-interventions that are inconsistent with the trial protocol which affects outcome that arise because of the trial context balance between groups	9	4	0	0	0
	For estimating effect of adhering to intervention, if unblinding of participants or health care providers, whether the co-interventions that are inconsistent with the trial protocol which affects outcome balance between groups	7	5	0	0	1
Failure in implementing the intervention as intended	For estimating effect of assignment to intervention, if unblinding of participants or health care providers, whether there is failure in implementing the intervention as intended in the protocol which affects outcome that arise because of the trial context	5	7	0	0	1
	For estimating effect of assignment to intervention, if unblinding of participants or health care providers, whether the failure in implementing the intervention as intended in the protocol which affects outcome that arise because of the trial context balance between groups	5	6	0	0	2
	For estimating effect of adhering to intervention, whether there is failure in implementing the intervention as intended in the protocol which affects outcome	4	8	0	0	1
Non-adherence to the assigned	For estimating the effect of assignment to intervention, if unblinding of participants or health care providers, whether there is non-adherence to their assigned intervention by trial	5	8	0	0	0

intervention	participants (imperfect compliance with a sustained intervention, cessation of intervention, crossover to the comparator intervention and switch to another active intervention) which affects outcome that arise because of the trial context					
	For estimating the effect of assignment to intervention, if unblinding of participants or health care providers, whether the non-adherence to their assigned intervention by trial participants (imperfect compliance with a sustained intervention, cessation of intervention, crossover to the comparator intervention and switch to another active intervention) which affects outcome that arise because of the trial context balance between groups	6	6	0	0	1
	For estimating effect of adhering to intervention, whether there is non-adherence to their assigned intervention by trial participants (imperfect compliance with a sustained intervention, cessation of intervention, crossover to the comparator intervention and switch to another active intervention) which affects outcome	4	7	1	0	1
	Whether the rate of non-adherence to the assigned intervention is high	4	8	0	0	1
Outcome measurement	Validity of outcome measurement	5	5	2	0	1
	Non-differential measurement error of outcome measurement	4	1	5	0	3
	Whether the method of outcome measurement (data collection) is sensitive to plausible intervention effects (e.g. important ranges of outcome values fall inside levels that are detectable using the measurement method)	4	4	2	0	3
Missing outcome data	Whether the sensitivity analysis shows that results are little changed under a range of plausible assumptions about the relationship between missingness in the outcome and its true value, or whether the analysis that corrects for bias for missing outcome data suggests no bias in result	8	1	1	0	3
	Whether missing data have been imputed using appropriate methods	4	2	0	0	7
	Whether the analysis that accounts for participant characteristics explains the relationship between missingness and the true value of the outcome	5	2	0	3	3
Selective reporting	The outcome of interest has been reported but it was not prespecified	4	3	0	4	2
	Whether there is selective reporting of a particular outcome measurement from multiple	8	2	0	2	1

	measurements assessed within an outcome domain					
	Whether there is selective reporting of a particular analysis from multiple analyses of a specific outcome measurement	9	2	0	2	0
Analysis	For estimating effect of assignment to intervention, if an inappropriate analysis is used (appropriate analysis is intention-to-treat analysis; inappropriate analyses include as-treated analysis, per-protocol analysis, etc.), whether the proportion of participants who are analyzed in the wrong group or excluded from the analysis (deviation from intention-to-treat analysis) is big enough to influence result	9	2	0	1	1
	For estimating effect of adhering to intervention, if deviations from intended intervention occur, whether an appropriate analysis is used (appropriate analyses include instrumental variable analyses and inverse probability weighting; inappropriate analyses include intention-to-treat analysis, per-protocol analysis, as treated analysis, etc.)	4	5	0	1	3
Early termination	Stop early for benefit	8	2	2	0	1
	Stop early for harm	5	3	2	0	3
	Stop early for futility	4	3	2	0	4
Others	Conflicts of interest	4	1	0	2	6
Crossover randomized controlled trials	Chronic stable condition	6	4	0	0	3
	Whether there is sufficient time (i.e. long washout period) for any carryover effects to have disappeared before outcome assessment in the second period	9	2	0	0	2
	Period effect (i.e. systematic differences between responses in the second compared with the first period that are not due to the interventions being compared): 1. whether the number of participants allocated to each of the two sequences equal or nearly equal; 2. whether the analysis accounts for period effect	8	3	0	0	2
	Is a result based on data from both periods sought, but unavailable on the basis of carryover having been identified	5	2	0	2	4
Custer	Unit-of-analysis issue (i.e. whether the analysis accounts for cluster design)	4	0	5	1	3

randomized controlled trials	Whether all the individual participants identified and recruited before randomization of clusters	6	1	0	1	5
Stepped-wedge randomized controlled trials	Whether the analysis accounts for the effects of time trends	9	1	0	1	2

Appendix 5. Response options

Study ID	Response options for SINGLE item	Response options for OVERALL risk of bias assessment
CASP 2020	yes, no, can't tell	NR
CLARITY Group tool	definitely yes, probably yes, probably no, definitely no	NR
EPOC 2017	low risk, unclear risk, high risk	low risk, unclear risk, high risk
Faillie 2017	yes, no, unclear, NA	low, unclear, high
Higgins 2011 (Cochrane risk of bias tool 1.0)	low risk, unclear risk (either lack of information or uncertainty over the potential for bias), high risk	low risk, unclear risk, high risk
JBI 2020	yes, no, unclear, NA	include, exclude, seek further information
Kennedy 2019	yes, no, NA, NR	"we decided to stop reporting the overall summary score and instead leave the tool as a simple checklist"
Kocsis 2010	0 (poor execution or description and an inappropriate element of the study's design), 1 (a moderately described and executed study element, a poorly described but well executed element, or a well-described but poorly executed element), 2 (a well-described and executed design element, which, if nonstandard, is also well justified)	1 (exceptionally poor), 2 (very poor), 3 (moderately poor), 4 (average), 5 (moderately good), 6 (very good), 7 (exceptionally good)
Liu 2010	yes=1 point, no=0 point	0-8 point
Manchikanti 2014	numerical scoring	total maximum 48
NICE 2012	yes (always indicates that the study has been designed/conducted in such a way as to minimize the risk of bias for that item), no, unclear (the item is not reported or not clearly reported), NA (randomized controlled trial cannot give an answer of "yes" no matter how well it has been done)	low risk of bias, unclear/unknown risk, high risk of bias (not for overall assessment for a study, but for each potential type of bias)
NIHBL 2013	yes, no, other (CD, NR, NA)	good, fair, poor
SIGN checklist	yes, no, can't tell, does not apply	high quality, acceptable, low quality, unacceptable
Smart 2015	1 point, 0 point	0-15 points
Sterne 2019 (Cochrane risk of bias tool 2.0)	yes, probably yes, probably no, no, no information ("yes" and "no" typically imply that firm evidence is available; the "probably" responses typically imply that a judgment has been made. The "no information" response should be used only when insufficient details are available to allow a different response, and when, in the absence of these details, it would be unreasonable to respond "probably yes" or "probably no")	low risk of bias, some concerns, high risk of bias

Stone 2021 (MASTER scale)	1 = Safeguard present, 0 = Safeguard absent	0-36 points
Tonin 2019	low risk, high risk, unclear risk	low risk, high risk, unclear risk

CASP=Critical Appraisal Skills Programme; NR=not reported; CLARITY=Clinical Advances Through Research and Information Translation; EPOC=Cochrane Effective Practice and Organisation of Care; NA=not applicable; JBI=Joanna Briggs Institute; NICE=National Institute for Health and Care Excellence; NIHBL=National Heart, Lung, and Blood Institute; CD=cannot determine; SIGN=Scottish Intercollegiate Guidelines Network; MASTER=MethodologicAl STandards for Epidemiological Research.

Chapter 3: Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors

Cited as and reprinted from: Wang Y, Parpia S, Couban R, Wang Q, Armijo-Olivo S, Bassler D, et al. Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors. J Clin Epidemiol. 2024 Jan;165:111211.

ELSEVIER

Journal of Clinical Epidemiology 165 (2024) 111211

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors

Ying Wang[a,*], Sameer Parpia[b], Rachel Couban[c], Qi Wang[d], Susan Armijo-Olivo[e,f], Dirk Bassler[g], Matthias Briel[h], Romina Brignardello-Petersen[a], Lise Lotte Gluud[i], Sheri A. Keitz[j], Luz M. Letelier[k], Philippe Ravaud[l], Kenneth F. Schulz[m], Reed A.C. Siemieniuk[a], Dena Zeraatkar[a], Gordon H. Guyatt[a]

[a]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada
[b]Department of Oncology, McMaster University, Hamilton, Ontario, Canada
[c]Department of Anesthesia, McMaster University, Hamilton, Ontario, Canada
[d]School of Public Health, Capital Medical University, Beijing, China
[e]University of Applied Sciences, Faculty of Business and Social Sciences, Osnabrück, Germany
[f]Department of Physical Therapy, Faculty of Rehabilitation Medicine, University of Alberta, Edmonton Canada
[g]Department of Neonatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[h]Department of Clinical Research, Meta-Research Centre Basel, University Hospital Basel, Basel, Switzerland
[i]Gastro Unit, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark
[j]Department of Medicine, Lahey Hospital & Medical Center, Burlington, MA, USA
[k]Department of Internal Medicine, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile
[l]Epidemiology and Statistics Sorbonne Paris Cité Research Center (CRESS), INSERM, Université Paris Descartes, Paris, France
[m]School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Accepted 1 November 2023; Published online 7 November 2023

Abstract

Objectives: To investigate the impact of potential risk of bias elements on effect estimates in randomized trials.

Study Design and Setting: We conducted a systematic survey of meta-epidemiological studies examining the influence of potential risk of bias elements on effect estimates in randomized trials. We included only meta-epidemiological studies that either preserved the clustering of trials within meta-analyses (compared effect estimates between trials with and without the potential risk of bias element within each meta-analysis, then combined across meta-analyses; between-trial comparisons), or preserved the clustering of substudies within trials (compared effect estimates between substudies with and without the element, then combined across trials; within-trial comparisons). Separately for studies based on between- and within-trial comparisons, we extracted ratios of odds ratios (RORs) from each study and combined them using a random-effects model. We made overall inferences and assessed certainty of evidence based on Grading of Recommendations, Assessment, development, and Evaluation and Instrument to assess the Credibility of Effect Modification Analyses.

Results: Forty-one meta-epidemiological studies (34 of between-, 7 of within-trial comparisons) proved eligible. Inadequate random sequence generation (ROR 0.94, 95% confidence interval [CI] 0.90–0.97) and allocation concealment (ROR 0.92, 95% CI 0.88–0.97) probably lead to effect overestimation (moderate certainty). Lack of patients blinding probably overestimates effects for patient-reported outcomes (ROR 0.36, 95% CI 0.28–0.48; moderate certainty). Lack of blinding of outcome assessors results in effect overestimation for subjective outcomes (ROR 0.69, 95% CI 0.51–0.93; high certainty). The impact of patients or outcome assessors blinding on other outcomes, and the impact of blinding of health-care providers, data collectors, or data analysts, remain uncertain. Trials stopped early for benefit probably overestimate effects (moderate certainty). Trials with imbalanced cointerventions may overestimate effects, while trials with missing outcome data may underestimate effects (low certainty). Influence of baseline imbalance, compliance, selective reporting, and intention-to-treat analysis remain uncertain.

Conclusion: Failure to ensure random sequence generation or adequate allocation concealment probably results in modest overestimates of effects. Lack of patients blinding probably leads to substantial overestimates of effects for patient-reported outcomes. Lack of

* Corresponding author. Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4L8, Canada. Tel./fax: +1-289-680-1832.

E-mail address: yingwwy@163.com (Y. Wang).

https://doi.org/10.1016/j.jclinepi.2023.11.001
0895-4356/© 2023 Elsevier Inc. All rights reserved.

blinding of outcome assessors results in substantial effect overestimation for subjective outcomes. For other elements, though evidence for consistent systematic overestimate of effect remains limited, failure to implement these safeguards may still introduce important bias. © 2023 Elsevier Inc. All rights reserved.

Keywords: Risk of bias; Meta-epidemiological studies; Empirical evidence; Randomized trials; Systematic survey; Systematic review

1. Introduction

Randomized controlled trials (RCTs) represent the optimal study design for assessing effects of health care interventions [1]. Nevertheless, flaws in design and execution of RCTs may lead to overestimation or underestimation of treatment effects: i.e., bias [2]. Investigators have developed a number of risk of bias instruments for RCTs [2–5]. These instruments vary in the items they include, raising concerns of which items are most important—or even appropriate—to include [6].

Meta-epidemiological studies provide evidence of how potential risk of bias elements influence effect estimates in RCTs; therefore, potentially informing the choice of items for risk of bias instruments. Meta-epidemiological studies typically collect meta-analyses of RCTs and, within each meta-analysis compare effect estimates between trials with and without methodological safeguards against bias, and then combine across meta-analyses (i.e., based on between-trial comparisons) [7,8]. Another type of meta-epidemiological study collects trials with multiple arms some of which include risk of bias safeguards and other that do not, within each trial compares effect estimates between comparisons with and without risk of bias safeguards, and then combines across trials (i.e., based on within-trial comparisons).

Because meta-epidemiological studies based on between-trial comparisons are limited by the possibility of study-level confounding, meta-epidemiological studies based on within-trial comparisons provide more credible evidence [9].

Conducting a systematic survey of meta-epidemiological studies is likely to optimally inform risk of bias assessment, particularly if it includes appropriate statistical pooling of results across studies. Indeed, investigators have conducted several systematic surveys of meta-epidemiological studies [10–13]. These systematic surveys, however, suffer from limitations, including failure to differentiate within- from between-trial comparisons, a crucial issue in determining credibility of estimates of risk of bias elements on actual bias [9]. Moreover, only one of these systematic surveys has conducted statistical pooling of results, and that survey [11] — along with the others conducted to date—did not include new meta-epidemiological studies that have appeared in recent years [14–18], including the MetaBLIND study [17], which is the largest meta-epidemiological study examining impact of blinding on effect estimates in RCTs. In addition to systematically summarizing evidence, assessing the certainty of evidence is crucial in interpretation and application of results [19]. Failure to address the certainty of evidence represents another limitation of systematic surveys of meta-epidemiological studies conducted thus far.

Therefore, to optimally summarize evidence regarding risk of bias assessment in RCTs, we conducted an updated systematic survey of meta-epidemiological studies, and assessed the associated certainty of evidence.

2. Methods

2.1. Literature search

This systematic survey followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Appendix 1). Appendix 2 is the protocol of this systematic survey. Our study team reviewed meta-epidemiological studies included in the three existing systematic surveys that searched through 2015 [10,11,13] and assessed their eligibility. In addition, in collaboration with a research librarian, we conducted an electronic literature search of Medline, Embase, Web of Science, and the Cochrane Database of Systematic Reviews from 2015 until January 20, 2023 (Appendix 3). The team also reviewed the reference lists of eligible meta-epidemiological studies to identify other potentially eligible studies.

2.2. Study selection

Two reviewers (YW, QW) independently screened abstracts followed by full texts. They resolved discrepancies by consensus or, if necessary, involved a third reviewer (GG, SP).

We included meta-epidemiological studies investigating the influence of potential risk of bias elements on effect estimates in RCTs of any health care topic. For studies using between-trial comparisons, we included only those that preserved the clustering of trials within meta-analyses in which comparisons were made only between trials within meta-analyses (all studies had similar population, intervention and outcome). For studies based on within-trial comparisons, only those preserving the clustering of substudies within trials (the comparisons were made only between substudies within trials which have same population, intervention and outcome), were eligible.

Eligible meta-epidemiological studies analyzed the influence of potential risk of bias elements quantitatively as a difference in effect estimates and combined across the

What is new?

Key findings

- On average, failure to ensure random sequence generation or adequate allocation concealment probably result in modest overestimates of treatment effects. Lack of patients blinding probably leads to substantial overestimates of effects for patient-reported outcomes. Lack of blinding of outcome assessors results in substantial effect overestimation for subjective outcomes.

What this adds to what was known?

- Our results are consistent with previous systematic surveys of meta-epidemiological studies in that all concluded evidence supports overestimation as a result of inadequate random sequence generation or allocation concealment.
- Our results are similar to one prior systematic survey in the conclusion of possible effect overestimation with lack of/unclear blinding of patients or lack of/unclear blinding of outcome assessors for subjective outcomes.
- Our results also suggested possible effect overestimation in trials with imbalanced cointerventions and trials stopped early for benefit, and possible effect underestimation in trials with missing outcome data.
- We provided an explicit rating of certainty of evidence regarding these inferences based on explicit criteria developed from the Instrument to assess the Credibility of Effect Modification Analyses and the Grading of Recommendations, Assessment, development, and Evaluation approach.

What is the implication and what should change now?

- Our empirical evidence, together with theoretical considerations, trial reporting practice, and the ease with which reviewers can make the required judgements, can inform the choice of items for risk of bias instruments.

meta-analyses/trials to obtain a single result. Authors used the following effect measures.

- Ratio of odds ratios (RORs) = $OR_{methodologically\ inferior}/OR_{superior}$
- Ratio of hazard ratios (RHRs) = $HR_{methodologically\ inferior}/HR_{superior}$
- Ratio of risk ratios (RRRs) = $RR_{methodologically\ inferior}/RR_{superior}$
- Difference in standardized mean differences (dSMD) = $SMD_{methodologically\ inferior} - SMD_{superior}$

This study focused on potential risk of bias elements. Another paper presented the details of how we chose the potential risk of bias elements [6]. Briefly, we collected items that were included in risk of bias instruments, and classified them into three categories: items the majority thought address risk of bias (category 1); items the majority thought do not address risk of bias (category 2); and items with substantial disagreement regarding whether they are addressing risk of bias (category 3).

Items in category 1 and 3 were the elements on which we focused in this systematic survey.

- Random sequence generation
- Allocation concealment
- Baseline imbalance
- Blinding of health-care providers
- Blinding of patients
- Blinding of data collectors
- Blinding of outcome assessors
- Blinding of data analysts
- Double blinding (as reported by the authors)
- Imbalance in cointerventions
- Compliance/adherence with intervention
- Missing outcome data
- Selective reporting
- Intention-to-treat analysis
- Early termination: stop early for benefit, stop early for harm, stop early for futility

We excluded single systematic reviews and meta-analyses that presented meta-regression, subgroup or sensitivity analyses based on risk of bias elements; meta-epidemiological studies examining the impact of only overall risk of bias rather than individual risk of bias elements; studies addressing non-health care topics (e.g., animal studies); and protocols or conference abstracts.

2.3. Data extraction

Paired reviewers (YW, QW) independently extracted data, resolving disagreement by consensus or if necessary involving a senior reviewer (GG, SP). Reviewers extracted the following information from each meta-epidemiological study.

- Based on between or within-trial comparisons
- Number of included systematic reviews, meta-analyses, trials, and participants
- Identification of systematic reviews or trials: source, eligibility criteria, topic area, sampling frame
- Choice of meta-analyses within systematic reviews (a systematic review may include several meta-analyses with different interventions or outcomes, how authors decided which to include in their analysis)

- Statistical approach or model (common models that preserve clustering include: meta-meta-analytic approach (i.e., two-step approach) [7], multivariable multilevel model [20], Bayesian hierarchical model [21], linear or logistic regression model preserving clustering [22], and label-invariant model [23])
- Focusing on only undesirable outcomes (e.g., death), desirable outcomes (e.g., recovery), or both
- Approach to distinguishing between intervention and control groups
- Approach to dealing with overlap of trials across meta-analyses
- Potential risk of bias elements considered
- Type of comparison (e.g., high/unclear vs. low risk of bias, high vs. low, unclear vs. low, high vs. low/unclear, definitely/probably no vs. definitely/probably yes, etc.)
- Type of outcome, control, and intervention
- Adjustment or no adjustment for other risk of bias elements or confounders
- Average bias: ROR, RHR, RRR, or dSMD, and 95% confidence or credible interval

2.4. Data synthesis and analysis

We analyzed data such that a ROR/RHR/RRR <1 or dSMD <0 indicates methodologically inferior trials yielded a larger effect estimate (more beneficial or less harmful effect of the experimental intervention) compared to methodological superior trials. For meta-epidemiological studies that present effects in the opposite direction (the original study compared methodological superior trials vs. inferior trials, or focused on desirable outcomes), we recalculated the effect measures as the inverse of the reported ROR/RHR/RRR, or 0 - reported dSMD.

To synthesize the average bias for binary and continuous outcomes, we converted dSMDs to log(RORs) by multiplying by $\pi/\sqrt{3} = 1.814$ [24]. We combined the RORs from different meta-epidemiological studies using the DerSimonian and Laird with Hartung-Knapp adjustment random-effects model [25]. The analysis included the I^2 statistic as a measure of inconsistency. We separately conducted analyses for between- and within-trial comparisons. To assess the magnitude of overlap between meta-epidemiological studies, we compared the source and topic area of included meta-analyses among meta-epidemiological studies. All analyses were performed in R (version 4.2.2, R Foundation for Statistical Computing), using the meta and metafor packages.

Meta-epidemiological studies reported different comparisons based on how they assessed risk of bias and dealt with trials with unclear reporting: high/unclear vs. low risk of bias, high vs. low, unclear vs. low, high vs. low/unclear, definitely/probably no vs. definitely/probably yes, etc. We conducted separate analyses for each comparison.

Meta-epidemiological studies reported results for different types of outcomes, controls, or interventions. Our team set rules a priori for selection of results into our primary analyses (Appendix 4).

The influence of risk of bias element may be confounded by other risk of bias elements, and studies sometimes provided both adjusted and unadjusted estimates. Appendix 4 presents the basis for our choices of which estimates to include in our analyses. Briefly, we selected adjusted results in preference of unadjusted results, and selected results adjusted for more other risk of bias elements in preference of results adjusted for less elements.

For blinding elements, meta-epidemiological studies may compare trials based on risk of bias judgment (e.g., high vs. low risk of bias, in which unblinded studies may still be considered low risk of bias), or compare trials based on whether blinding has been implemented (e.g., unblinded vs. blinded). If a study reported both, we selected the comparison based on blinding status rather than risk of bias judgment.

We conducted a number of subgroup analyses. Appendix 5 presents the detailed methods. We applied the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) to assess credibility of subgroup effect [9].

2.5. Inferences and certainty of inferences

Systematic review authors frequently use the Grading of Recommendations, Assessment, development, and Evaluation (GRADE) approach [19] to assess certainty of evidence in systematic reviews of clinical topics, suggesting its possible use for assessing certainty of evidence in systematic surveys of meta-epidemiological studies. ICEMAN [9], which is increasingly used to assess credibility of subgroup effect in systematic reviews or RCTs, is also relevant here, because when meta-epidemiological studies address whether effects differ between studies with or without risk of bias elements, they are conceptually addressing the presence or absence of risk of bias elements as an effect modifier or subgroup effect. In addressing the certainty of evidence regarding the impact of risk of bias elements on magnitude of effect, we developed an approach that consider both GRADE and ICEMAN criteria.

Following the GRADE approach [19], we first made inferences, for each comparison (e.g., high vs. low risk of bias for random sequence generation), regarding whether each risk of bias element biases the effect (ROR was less than 0.95 or greater than 1.05) or, rather, has little or no bias (ROR was between 0.95 and 1.05).

For each comparison, we then rated the certainty of the inference using a strategy based on principles from the GRADE approach [19]. Our approach to rating certainty used GRADE categories of high, moderate, low and very low certainty, and considered criteria to assess credibility of subgroup effects within meta-analyses from the ICEMAN [9].

In accordance with ICEMAN [9], certainty ratings based on between-trial comparisons started as moderate certainty,

and those based on with-trial comparisons started as high certainty. For other ICEMAN criteria, we presumed that study authors had made a priori hypotheses specifying a direction (risk of bias leads to larger effects); noted that authors typically used random effect models; and acknowledge that there were typically a large number of effect modifiers considered, but did not typically rate down overall certainty for this limitation.

Regarding precision of estimates, when ROR was less than 0.95 (i.e., effect overestimation) or more than 1.05 (i.e., effect underestimation), we rated down once for imprecision if the 95% confidence interval (95% CI) crossed 1.00 (serious imprecision); twice if the 95% CI crossed the threshold on the opposite side (e.g., ROR <0.95 and 95% CI includes 1.05, very serious imprecision); and thrice if it crossed 1.25 or 0.80 (e.g., ROR <0.95 and 95% CI included 1.25, extremely serious imprecision). When 0.95≤ ROR ≤1.05 (i.e., little or no bias), we rated down once for imprecision if 95% CI crossed 0.95 or 1.05; twice if crossed 0.90 or 1.10; and thrice if crossed 0.80 or 1.25.

We reasoned that when there were only a small number of comparisons (e.g., a small number of meta-analyses), even if a risk of bias was convincingly demonstrated, generalization to the entire body of randomized trials was tenuous. Thus, following GRADE terminology, we rated down for indirectness. We rated down once if meta-epidemiological studies included 10−20 comparisons; twice if less than 10 comparisons. For inconsistency ratings, we used GRADE criteria of similarity of point estimates, extent of overlap of 95% CI across studies, and I^2.

We made final inferences regarding the impact of each risk of bias element with consideration of all comparisons and the consistency of the results across comparisons. When all comparisons reported results that supported the same inference (for instance, all reported pooled estimates of 0.95 or less, supporting a conclusion of overestimated effects in methodological inferior trials), we rated certainty on the basis of the highest certainty comparison. When comparisons were inconsistent - that is, they suggested different inferences - we made inferences based primarily on the comparisons with higher certainty. When different comparisons with high certainty suggested different inferences, we concluded there remains high uncertainty regarding the impact of the risk of bias element. When all comparisons proved of very low certainty, we concluded that we are very uncertain of the impact the element.

3. Results

3.1. Search results

Our search identified 3,484 records from databases (Fig. 1). After removing duplicates, of 1,987 remaining, we screened 321 full texts, of which 24 proved eligible. Our team identified 17 other eligible studies from existing systematic surveys of similar topics or reference lists of eligible studies, resulting in a total of 41 eligible meta-epidemiological studies [14−18,20,22,26−60].

Two meta-epidemiological studies developed databases by combining data from other meta-epidemiological studies, removed overlapping meta-analyses and trials, and then conducted their own analyses [40,61]. Wood et al. [61] combined data from three meta-epidemiological studies, while the Bias in Randomized and Observational studies (BRANDO) project [40] combined the same three along with another four meta-epidemiological studies. We included the BRANDO study and excluded the Wood et al. since it did not report any additional risk of bias element, comparison, or subgroup. Regarding the seven meta-epidemiological studies included in the BRANDO study, we included three [20,22,42] that examined additional elements, comparisons, or subgroups of interest.

3.2. Characteristics of included studies

Of the 41 meta-epidemiological studies, 34 were based on meta-analyses (i.e., between-trial comparisons) [14−18,20,22,26−33,35−45,47,48,53−55,57−60], and 7 were based on trials (i.e., within-trial comparisons) [34,46,49−52,56] (Table 1).

The 34 studies based on between-trial comparisons included a total of 1,740 meta-analyses with 17,259 trials (Appendix 6). The median number of meta-analyses included per study was 27.

The seven studies based on within-trial comparisons included 238 trials (Appendix 6). Five studies included trials with blinded and unblinded assessment of the same outcome (two binary outcomes [50,56], two time-to-event outcomes [46,52], and one measurement scale outcomes [51]), thus examining the influence of blinding of outcome assessors. One study included four- or three-armed trials that can be divided as a patient-blinded substudy and a patient-unblinded substudy, thus assessing blinding of patients [49]. One study examined influence of intention-to-treat analysis by including trials that reported both an effect estimate based on intention-to-treat and an effect estimate based on per-protocol analysis [34]. Appendix 7 shows definition of the potential risk of bias elements.

3.3. Random sequence generation

Nineteen meta-epidemiological studies [14,16,18,22,28,30,31,37,39−41,44,45,48,53−55,57,59,60] assessed influence of random sequence generation on effect estimates, all of which were based on between-trial comparisons. Fifteen studies [16,18,28,30,31,37,39−41,44,45,48,53,60] suggested on average, trials with inadequate or unclear random sequence generation probably overestimate effects compared to trials with adequate random sequence generation (ROR 0.94, 95% CI 0.90−0.97; moderate certainty) (Table 2). Two studies [28,40] suggested trials with unclear

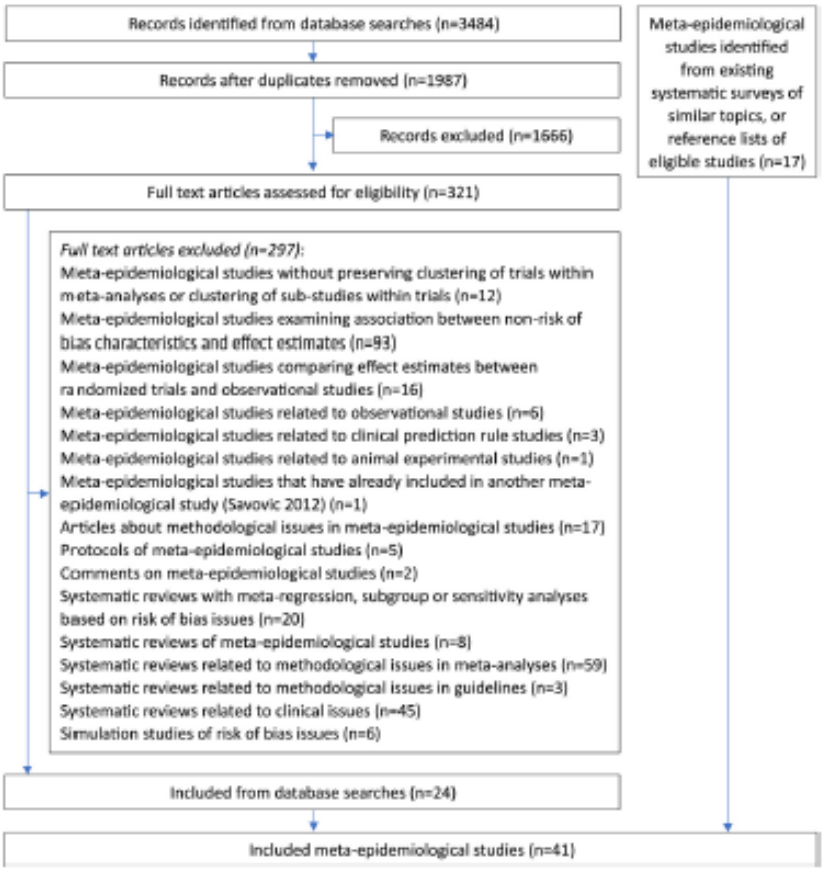

Fig. 1. Flow diagram of selection of studies. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

random sequence generation probably overestimate effects (ROR 0.89, 95% CI 0.83−0.97; moderate certainty) (Table 2). Appendix 8 presents the forest plots.

3.4. Allocation concealment

Twenty-two studies based on between-trial comparisons assessed the influence of allocation concealment [14,16−18,22,28,30−32,35,39−41,44,45,48,53−55,57−60]. Sixteen studies [16−18,28,30−32,35,39−41,44,45,53,57,60] suggested inadequate or unclear allocation concealment probably leads to an effect overestimation compared to adequate concealment (ROR 0.92, 95% CI 0.88−0.97; moderate certainty) (Table 2). Three studies [28,40,48] suggested unclear allocation concealment probably results in an effect overestimation compared to adequate concealment (ROR 0.85, 95% CI 0.74−0.98; moderate certainty) (Table 2). The effect was also observed in studies that authors described as double-blind, and trials with both adequate random sequence generation and blinding (moderate or low certainty) (Appendix 9).

One study compared different methods of allocation concealment [32]. Compared to central randomization, trials using sealed envelopes with no further elaboration (low certainty) and trials simply reported as "randomized" with no further details (moderate certainty) likely overestimate effects (Appendix 9).

3.5. Baseline imbalance

Four studies of between-trial comparisons examined the influence of baseline imbalance [31,39,42,57]. Although results suggested trials with imbalanced baseline characteristics may report larger effect estimates, the certainty of evidence is very low (Table 2). Thus, we remain very uncertain of the impact of baseline imbalance.

3.6. Blinding of health-care providers

Five studies of between-trial comparisons assessed the effect of blinding of health-care providers [14,17,38,42,44]. Comparison of high vs. low risk of bias suggested an effect overestimation associated with unblinding of health-care providers with very low certainty,

Table 1. Summary of characteristics of included meta-epidemiological studies

Characteristics of meta-epidemiological studies based on between-trial comparisons	
Topic area of systematic reviews	
Any	26.5% (9/34)
Specific topics	73.5% (25/34)
Choice of meta-analyses within systematic reviews	
Meta-analyses with primary outcome (or most clinically important outcome)	17.6% (6/34)
Meta-analyses with largest number of trials	8.8% (3/34)
Meta-analyses with primary outcome with largest number of trials	11.8% (4/34)
Meta-analyses with primary outcome or outcome with largest number of trials	20.6% (7/34)
Meta-analyses with specific outcome (mortality, pain)	8.8% (3/34)
Meta-analyses with most homogeneous group of interventions	2.9% (1/34)
Randomly chose one	2.9% (1/34)
Analyzed for different outcomes separately	5.9% (2/34)
Not reported	20.6% (7/34)
Statistical approach	
Meta-meta-analytic approach	58.8% (20/34)
Multivariable multilevel model	17.6% (6/34)
Bayesian hierarchical model	11.8% (4/34)
Regression model preserving clustering	11.8% (4/34)
Type of outcomes	
Binary outcomes	50.0% (17/34)
Continuous outcomes	26.5% (9/34)
Both - analyzed separately	8.8% (3/34)
Both - combined together	14.7% (5/34)
Type of outcomes	
Undesirable outcomes	67.6% (23/34)
Desirable outcomes	8.8% (3/34)
Both	2.9% (1/34)
Not reported	20.6% (7/34)
Approach to distinguishing between intervention and control groups	
Included only meta-analyses with placebo, no treatment or standard care as control group	20.6% (7/34)
Followed classification in original systematic reviews	14.7% (5/34)
Excluded meta-analyses if intervention and control cannot be ascertained	8.8% (3/34)
Followed descriptions in trial reports	2.9% (1/34)
Others	5.9% (2/34)
Not reported	47.1% (16/34)
Characteristics of meta-epidemiological studies based on within-trial comparisons	
Topic area of trials	
Any	85.7% (6/7)
Oncology	14.3% (1/7)
Choice of outcome included in analysis	
Primary outcome (or most clinically important outcome)	71.4% (5/7)
Specific outcome	14.3% (1/7)
Not reported	14.3% (1/7)
Statistical approach	
Random effects meta-analysis	100% (7/7)
Type of outcomes	
Binary outcomes	28.6% (2/7)
Measurement scale outcomes	28.6% (2/7)
Time-to-event outcomes	28.6% (2/7)
Both binary and continuous outcomes - combined together	14.3% (1/7)

(*Continued*)

Table 1. Continued

Type of outcomes	
Undesirable outcomes	57.1% (4/7)
Not reported	42.9% (3/7)
Characteristics assessed	
Blinding of outcome assessors	71.4% (5/7)
Blinding of participants	14.3% (1/7)
Intention-to-treat analysis	14.3% (1/7)
Approach to distinguishing between intervention and control groups	
Excluded if unclear	42.9% (3/7)
Not reported	57.1% (4/7)

regardless of the type of outcome (Table 2). Other three comparisons reported little or no impact of blinding of health-care providers on effect estimates with very low certainty evidence (Table 2). Our overall inference is, therefore, that we remain very uncertain of the impact of blinding of health-care providers on estimates of intervention effects.

3.7. Blinding of patients

Ten studies examined influence of blinding of patients, nine of which were based on between-[14,17,27,29,30,35,38,42,44] and one within-trial comparison [49]. The within-trial comparison provided moderate certainty evidence of overestimates of treatment effects with unblinding of patients in trials measuring patient-reported outcomes (ROR 0.36, 95% CI 0.28−0.48) (Table 2), but not in trials when outcomes were measured by observers (very low certainty) (Table 2), with moderate credibility of the difference between the two (interaction $P < 0.0001$) (Appendix 10). Between-trial comparisons provided only very low certainty evidence (Table 2).

Overall inference is that we have moderate certainty evidence that failure to blind patients probably overestimates treatment effects for patient-reported outcomes, while we are very uncertain of its influence on other type of outcomes.

3.8. Blinding of data collectors

One study [14] of between-trial comparisons raised the possibility that trials with unblinding of data collectors may overestimate effect; however, the certainty of evidence is very low (Table 2). We, therefore, conclude we remain very uncertain of the influence of blinding of data collectors.

3.9. Blinding of outcome assessors

Sixteen studies (eleven were based on between-trial comparisons [14,16−18,27,30,31,38,42,60] and five on within-trial comparisons [46,50−52,56]) assessed the influence of blinding of outcome assessors.

Within-trial comparisons showed high certainty evidence of effect overestimation with unblinded outcome assessment for subjective outcomes (Table 2) (Appendix 6 shows the definition of types of outcomes).

Between-trial comparisons provided only very low certainty evidence of the impact of blinding of outcome assessors on any types of outcomes, objective outcomes and subjective outcomes (Table 2). Subgroup analysis raised the possibility that unblinded outcome assessment overestimated effects more for subjective outcomes than for objective outcomes with moderate credibility (Appendix 10).

We conclude that we have high certainty evidence that failure to blind outcome assessors overestimates treatment effects when outcome measures are subjective. In other situations, the evidence is very uncertain.

3.10. Blinding of data analysts

One study [27] of between-trial comparisons provided only very low certainty evidence regarding blinding of data analysts (Table 2). Thus, we remain very uncertain of the effect of blinding of data analysts.

3.11. Double blinding

Seventeen studies based on between-trial comparisons assessed influence of double blinding (as reported by the authors or blinding of two groups) [14,16−18,22,31,33,38,40,41,45,48,53−55,57,58,60]. Seven studies comparing high vs. low risk of bias [14,18,31,40,48,54,55,58] suggested lack of double blinding probably leads to an effect overestimation for subjective outcomes with moderate certainty (ROR 0.70, 95% CI 0.53−0.89) (Table 2), and any outcomes with low certainty (ROR 0.89, 95% CI 0.77−1.04).

Twelve studies comparing high/unclear vs. low risk of bias [16−18,31,33,38,40,41,45,53,57,60] provided moderate certainty evidence of overestimates of effect with lack of double blinding for any outcomes (ROR 0.93, 95% CI 0.88−0.99), objective outcomes (ROR 0.94, 95% CI

0.88−0.99), and subjective outcomes (ROR 0.90, 95% CI 0.83−0.96). Other comparisons provided only very low certainty evidence.

We conclude that we have moderate certainty evidence that trials not reported as double blind are probably associated with overestimates of treatment effects, regardless of the type of outcomes.

3.12. Imbalance in cointerventions

One study [47] of between-trial comparisons suggested that trials with imbalanced or unclear cointerventions may estimate a larger effect compared to trials with no or balanced cointerventions (ROR 0.86, 95% CI 0.73−0.99; low certainty) (Table 2).

3.13. Compliance with intervention

Two studies [47,60] of between-trial comparison provided only very low certainty evidence regarding compliance with intervention (Table 2).

3.14. Missing outcome data

Six studies based on between-trial comparisons compared trials with a higher vs. trials with a lower proportion of missing outcome data using different thresholds [15,22,36,40,42,53]. Trials with more than 5% of missing outcome data may underestimate effects compared to trials with less than 5% of missing data (ROR 1.46, 95% CI 1.16−1.85; low certainty) (Table 2). Studies using other missing proportion thresholds provided only very low certainty evidence (Table 2). Ten studies comparing trials based on risk of bias assessment for the missing outcome data item [15,16,18,31,41,44,45,57,59,60], provided only very low certainty evidence (Table 2).

We conclude that we have low certainty evidence that missing outcome data may leads to underestimates of effects.

3.15. Selective reporting

Five studies of between-trial comparisons assessed influence of selective reporting [31,44,57,59,60]. Although results suggested that trials with selective reporting may overestimate effects, the certainty of evidence is very low (Table 2). Thus, the evidence for selective reporting is uncertain.

3.16. Intention-to-treat analysis

Four studies of between-trial comparisons [15,20,26,42] and one study of within-trial comparisons [34] examined the impact of intention-to-treat analysis on effect estimates. The study of within-trial comparisons found [34], for trials with significant difference between intervention and control groups, per-protocol analysis produced a larger effect estimate compared to intention-to-treat or modified intention-to-treat analysis (in which patients were excluded if a certain minimum dose of intervention was not received) (ROR 0.93, 95% CI 0.90−0.98; high certainty) (Table 2). However, for trials with no significant difference between intervention and control groups, per-protocol analysis and intention-to-treat or modified intention-to-treat analysis produced similar effect estimates (ROR 0.99, 95% CI 0.97−1.01; high certainty) (Table 2). Between-trial comparisons also suggested different inferences.

Since comparisons with high certainty evidence suggested different inferences, our overall inference is we remain very uncertain of the impact of intention-to-treat analysis on effect estimates.

3.17. Stopping early for benefit

One study of between-trial comparisons suggested trials that stopped early for benefit probably report larger effect estimates compared to trials without early termination (RRR 0.71, 95% CI 0.65−0.77; moderate certainty) (Table 2) [43].

4. Discussion

Meta-epidemiological studies provide moderate certainty evidence that inadequate random sequence generation or allocation concealment probably leads to modest effect overestimation. They also provide moderate certainty evidence that lack of patients blinding probably overestimates effects for patient-reported outcomes; we are, however, very uncertain of its influence on other types of outcomes. Studies provide high certainty evidence that unblinded outcome assessment overestimates effects for subjective outcomes; we remain, however, very uncertain of its impact on other outcomes. We remain very uncertain of the impact of blinding of health-care providers, data collectors, or data analysts on effect estimates.

Missing outcome data may lead to underestimates of effects, and trials with imbalanced or unclear cointerventions may overestimate treatment effects (both low certainty). One study provides moderate certainty evidence that trials that stopped early for benefit probably overestimate effects.

We remain very uncertain of the impact of baseline imbalance, compliance, selective reporting and intention-to-treat analysis on effect estimates.

One strength is we included only meta-epidemiological studies that preserved the clustering of trials within meta-analyses or the clustering of substudies within trials, since they are much more credible than studies without clustering.

Similar to systematic reviews of health care interventions, statistical pooling enhances precision of effect estimates, allows statistical assessment of variability in effects, and simplifies interpretation of results from

Table 2. Impact of potential risk of bias elements on effect estimates in randomized trials

Comparison	Number of ME studies (between- or within-trial comparisons)	Average bias (95% CI)	Direction of bias (certainty of evidence[a])	Overall inference for the item
Random sequence generation				**Overestimation (Moderate certainty)**
High vs. low risk of bias	7 (between-trial)	ROR 0.93 (0.75–1.16)	Overestimation (Very low[b])	
High/unclear vs. low risk of bias	15 (between-trial)	ROR 0.94 (0.90–0.97)	Overestimation (Moderate)	
High vs. low/unclear risk of bias	2 (between-trial)	ROR 0.91 (0.76–1.09)	Overestimation (Very low[b])	
Unclear vs. low risk of bias	2 (between-trial)	ROR 0.89 (0.83–0.97)	Overestimation (Moderate)	
Allocation concealment				**Overestimation (Moderate certainty)**
High vs. low risk of bias	9 (between-trial)	ROR 0.95 (0.69–1.29)	Little or no bias (Very low[c])	
High/unclear vs. low risk of bias	16 (between-trial)	ROR 0.92 (0.88–0.97)	Overestimation (Moderate)	
High vs. low/unclear risk of bias	2 (between-trial)	ROR 0.92 (0.80–1.07)	Overestimation (Very low[b])	
Unclear vs. low risk of bias	3 (between-trial)	ROR 0.85 (0.74–0.98)	Overestimation (Moderate)	
Definitely/probably unconcealed vs. definitely/probably concealed	1 (between-trial)	ROR 0.87 (0.66–1.14)	Overestimation (Very low[b,d])	
Baseline imbalance				**Very uncertain**
Imbalance vs. balance	1 (between-trial)	ROR 0.78 (0.47–1.29)	Overestimation (Very low[c])	
Imbalance/unclear vs. balance	4 (between-trial)	ROR 0.93 (0.80–1.09)	Overestimation (Very low[b])	
Imbalance vs. balance/unclear	1 (between-trial)	ROR 0.86 (0.52–1.44)	Overestimation (Very low[c,d])	
Blinding of health-care providers				**Very uncertain**
High vs. low risk of bias	1 (between-trial)	Health-care provider decision outcomes: ROR 0.51 (0.24–1.06)	Overestimation (Very low[b,e])	
		Observer-reported outcomes: ROR 0.76 (0.61–0.96)	Overestimation (Very low[e])	
		Patient-reported outcomes: ROR 0.55 (0.12–2.62)	Overestimation (Very low[c,e])	
High/unclear vs. low risk of bias	3 (between-trial)	Any outcomes: ROR 0.98 (0.89–1.08)	Little or no bias (Very low[b])	
Definitely/probably unblinded vs. definitely/probably blinded	1 (between-trial)	Any outcomes: ROR 0.97 (0.80–1.17)	Little or no bias (Very low[c])	
		Health-care provider decision outcomes: ROR 0.97 (0.77–1.18)	Little or no bias (Very low[c])	
		Blinded patient- or observer-reported outcomes: ROR 0.96 (0.64–1.45)	Little or no bias (Very low[c,d])	
Definitely/probably unblinded or unclear vs. definitely/probably blinded	1 (between-trial)	Any outcomes: ROR 1.01 (0.86–1.19)	Little or no bias (Very low[b])	
		Health-care provider decision outcomes: ROR 1.03 (0.84–1.23)	Little or no bias (Very low[b,d])	
		Blinded patient- or observer-reported outcomes: ROR 1.03 (0.67–1.54)	Little or no bias (Very low[c])	

(*Continued*)

Table 2. Continued

Comparison	Number of ME studies (between- or within-trial comparisons)	Average bias (95% CI)	Direction of bias (certainty of evidence[a])	Overall inference for the item
Blinding of patients				*Any outcomes:* **Very uncertain** *Patient-reported outcomes:* **Overestimation (Moderate certainty)** *Observer-reported or objective outcomes:* **Very uncertain**
High vs. low risk of bias	1 (between-trial)	Patient-reported outcomes: ROR 0.48 (0.12–1.92)	Overestimation (Very low[c,e])	
		Observer-reported outcomes: ROR 0.76 (0.61–0.93)	Overestimation (Very low[e])	
High/unclear vs. low risk of bias	7 (between-trial)	Any outcomes: ROR 0.92 (0.80–1.07)	Overestimation (Very low[b])	
		Patient-reported outcomes: ROR 0.86 (0.14–5.38)	Overestimation (Very low[c])	
		Objective outcomes: ROR 0.87 (0.43–1.77)	Overestimation (Very low[c])	
Definitely/probably unblinded vs. definitely/probably blinded	1 (between-trial)	Any outcomes: ROR 1.04 (0.79–1.37)	Little or no bias (Very low[c])	
		Patient-reported outcomes: ROR 1.10 (0.72–1.69)	Underestimation (Very low[c,d,f])	
		Blinded observer-reported outcomes: ROR 1.00 (0.70–1.44)	Little or no bias (Very low[c,d])	
Definitely/probably unblinded or unclear vs. definitely/probably blinded	1 (between-trial)	Any outcomes: ROR 0.94 (0.74–1.19)	Overestimation (Very low[b])	
		Patient-reported outcomes: ROR 0.91 (0.61–1.35)	Overestimation (Very low[c,d])	
		Blinded observer-reported outcomes: ROR 1.07 (0.74–1.56)	Underestimation (Very low[c,d,f])	
High vs. low risk of bias (in adequately concealed trials)	1 (within 12 trials)	Patient-reported outcomes: ROR 0.36 (0.28–0.48)[g]	Overestimation (Moderate[a])	
	1 (within 1 trial)	Blinded observer-reported outcomes: ROR 0.96 (0.67–1.39)[g]	Little or no bias (Very low[c,e])	
Blinding of data collectors				**Very uncertain**
High vs. low risk of bias	1 (between-trial)	Any outcomes: ROR 0.81 (0.58–1.13)	Overestimation (Very low[b,e])	
Blinding of outcome assessors				*Any outcomes:* **Very uncertain** *Objective outcomes:* **Very uncertain** *Subjective outcomes:* **Overestimation (High certainty)**
High vs. low risk of bias	3 (between-trial)	Any outcomes: ROR 1.00 (0.56–1.79)	Little or no bias (Very low[c])	
		Objective outcomes: ROR 0.82 (0.63–1.07)[g]	Overestimation (Very low[b,e])	
		Subjective outcomes: ROR 0.07 (0.02–0.33)[g]	Overestimation (Very low[a])	
High/unclear vs. low risk of bias	9 (between-trial)	Any outcomes: ROR 0.97 (0.89–1.06)	Little or no bias (Very low[b])	
		Objective outcomes: ROR 0.93 (0.73–1.19)	Overestimation (Very low[b])	
		Subjective outcomes: ROR 0.91 (0.64–1.29)	Overestimation (Very low[c])	
High vs. low/unclear risk of bias	1 (between-trial)	Any outcomes: ROR 1.24 (0.64–2.48)	Underestimation (Very low[c,d,f])	
Definitely/probably unblinded vs. definitely/probably blinded	1 (between-trial)	Subjective observer reported outcomes: ROR 1.01 (0.85–1.20)	Little or no bias (Very low[b])	
Definitely/probably unblinded or unclear vs. definitely/probably blinded	1 (between-trial)	Any outcomes: ROR 1.01 (0.89–1.14)	Little or no bias (Very low[b])	
		Objective outcomes: ROR 0.94	Overestimation	

(*Continued*)

Table 2. Continued

Comparison	Number of ME studies (between- or within-trial comparisons)	Average bias (95% CI)	Direction of bias (certainty of evidence[a])	Overall inference for the item
		(0.61–1.26)	(Very low[c,d])	
		Subjective observer reported outcomes: ROR 1.04 (0.89–1.23)	Little or no bias (Very low[b])	
High vs. low risk of bias (binary outcomes)	2 (within 23 trials)	Subjective outcomes: ROR 0.69 (0.51–0.93)	Overestimation (High)	
	1 (within 17 trials)	Clearly subjective outcomes: ROR 0.55 (0.32–0.95)	Overestimation (Moderate[d])	
	1 (within 5 trials)	Moderately subjective outcomes: ROR 0.93 (0.56–1.54)	Overestimation (Very low[c,e])	
High vs. low risk of bias (measurement scale outcomes)	1 (within 16 trials)	Subjective outcomes: ROR 0.66 (0.48–0.90)	Overestimation (Moderate[d])	
	1 (within 13 trials)	Clearly subjective outcomes: ROR 0.59 (0.40–0.86)	Overestimation (Moderate[d])	
	1 (within 3 trials)	Moderately subjective outcomes: ROR 0.93 (0.56–1.57)	Overestimation (Very low[c,e])	
High vs. low risk of bias (time-to-event outcomes)	2 (within 31 trials)	Any outcomes: RHR 0.97 (0.61–1.56)	Little or no bias (Very low[c])	
	1 (within 11 trials)	Clearly subjective outcomes: RHR 0.88 (0.69–1.12)	Overestimation (Very low[b,c])	
Blinding of data analysts				**Very uncertain**
High/unclear vs. low risk of bias	1 (between-trial)	ROR 0.93 (0.61–1.39)	Overestimation (Very low[c,e])	
Double blinding				*Any outcomes:* **Overestimation (Moderate certainty)** *Objective outcomes:* **Overestimation (Moderate certainty)** *Subjective outcomes:* **Overestimation (Moderate certainty)**
High vs. low risk of bias	7 (between-trial)	Any outcomes: ROR 0.89 (0.77–1.04)	Overestimation (Low[c])	
		Objective outcomes: ROR 0.93 (0.80–1.08)	Overestimation (Very low[b])	
		Subjective outcomes: ROR 0.70 (0.53–0.89)	Overestimation (Moderate)	
High/unclear vs. low risk of bias	12 (between-trial)	Any outcomes: ROR 0.93 (0.88–0.99)	Overestimation (Moderate)	
		Objective outcomes: ROR 0.94 (0.88–0.99)	Overestimation (Moderate)	
		Subjective outcomes: ROR 0.90 (0.83–0.96)	Overestimation (Moderate)	
High vs. low/unclear risk of bias	2 (between-trial)	Any outcomes: ROR 1.06 (0.63–1.78)	Underestimation (Very low[c,f])	
Unclear vs. low risk of bias	1 (between-trial)	Any outcomes: ROR 0.90 (0.69–1.15)	Overestimation (Very low[b])	
		Objective outcomes: ROR 1.11 (0.73–1.70)	Underestimation (Very low[c,d,f])	
		Subjective outcomes: ROR 0.75 (0.51–1.09)	Overestimation (Very low[b,c])	
Definitely/probably unblinded vs. definitely/probably blinded	1 (between-trial)	Objective outcomes: ROR 1.12 (0.93–1.33)	Underestimation (Very low[b,d,f])	
Imbalance in cointerventions				**Overestimation (Low certainty)**
Imbalance/unclear vs. balance	1 (between-trial)	ROR 0.86 (0.73–0.99)	Overestimation (Low[c])	
Compliance with intervention				**Very uncertain**
Unacceptable/unclear noncompliance vs. no/acceptable noncompliance	2 (between-trial)	ROR 1.00 (0.65–1.54)	Little or no bias (Very low[c,d])	

(*Continued*)

Table 2. Continued

Comparison	Number of ME studies (between- or within-trial comparisons)	Average bias (95% CI)	Direction of bias (certainty of evidence[a])	Overall inference for the item
Missing outcome data				**Underestimation (Low certainty)**
Missing vs. no missing	2 (between-trial)	ROR 0.92 (0.66–1.28)	Overestimation (Very low[c])	
Missing ≥5% vs. <5%	1 (between-trial)	ROR 1.46 (1.16–1.85)	Underestimation (Low[f])	
Missing ≥10% vs. <10%	2 (between-trial)	ROR 1.14 (0.69–1.90)	Underestimation (Very low[c,f])	
Missing ≥15% vs. <15%	1 (between-trial)	ROR 1.16 (1.00–1.34)	Underestimation (Very low[b,f])	
Missing ≥20% vs. <20%	2 (between-trial)	ROR 1.16 (0.40–3.38)	Underestimation (Very low[c,f])	
Missing ≥25% vs. <25%	1 (between-trial)	ROR 1.09 (0.90–1.34)	Underestimation (Very low[b,f])	
Missing ≥30% vs. <30%	1 (between-trial)	ROR 1.08 (0.83–1.39)	Underestimation (Very low[b,f])	
Missing ≥35% vs. <35%	1 (between-trial)	ROR 1.11 (0.68–1.82)	Underestimation (Very low[c,f])	
Missing ≥40% vs. <40%	1 (between-trial)	ROR 0.91 (0.50–1.66)	Overestimation (Very low[c])	
Per 1% point increase in missing	1 (between-trial)	ROR 1.02 (0.94–1.12)	Little or no bias (Very low[b,d])	
High vs. low risk of bias	3 (between-trial)	ROR 0.64 (0.04–11.28)	Overestimation (Very low[c])	
High/unclear vs. low risk of bias	9 (between-trial)	ROR 0.99 (0.90–1.10)	Little or no bias (Very low[b])	
High vs. low/unclear risk of bias	2 (between-trial)	ROR 1.02 (0.90–1.15)	Little or no bias (Very low[b])	
Selective reporting				**Very uncertain**
High vs. low risk of bias	2 (between-trial)	ROR 0.82 (0.19–3.56)	Overestimation (Very low[c])	
High/unclear vs. low risk of bias	4 (between-trial)	ROR 0.83 (0.55–1.23)	Overestimation (Very low[b])	
High vs. low/unclear risk of bias	1 (between-trial)	ROR 0.85 (0.71–1.04)	Overestimation (Very low[d,h])	
Intention-to-treat analysis				**Very uncertain**
Not ITT vs. ITT	1 (between-trial)	ROR 0.80 (0.69–0.94)	Overestimation (Moderate)	
Not ITT/unclear vs. ITT	3 (between-trial)	ROR 1.05 (0.85–1.31)	Little or no bias (Very low[c])	
Not ITT vs. ITT/unclear	1 (between-trial)	ROR 0.90 (0.79–1.02)	Overestimation (Low[h])	
Not ITT vs. unclear	1 (between-trial)	ROR 0.92 (0.70–1.23)	Overestimation (Very low[b])	
PP vs. ITT	1 (within 133 trials)	ROR 0.97 (0.95–1.00)	Little or no bias (Moderate[h])	
PP vs. mITT	1 (within 24 trials)	ROR 0.99 (0.95–1.03)	Little or no bias (Moderate[h])	
PP vs. ITT/mITT	1 (within 157 trials)	ROR 0.98 (0.96–1.00)	Little or no bias (High)	
PP vs. ITT/mITT (positive trials: significant difference between intervention and control)	1 (within 52 trials)	ROR 0.93 (0.90–0.98)	Overestimation (High)	

(*Continued*)

Table 2. Continued

Comparison	Number of ME studies (between- or within-trial comparisons)	Average bias (95% CI)	Direction of bias (certainty of evidence[a])	Overall inference for the item
PP vs. ITT/mITT (negative trials: no significant difference between intervention and control)	1 (within 105 trials)	ROR 0.99 (0.97–1.01)	Little or no bias (High)	
Early termination				**Overestimation (Moderate certainty)**
Stop early for benefit vs. not	1 (between-trial)	RRR 0.71 (0.65–0.77)	Overestimation (Moderate)	

Abbreviations: ME, meta-epidemiological studies; ROR, ratio of odds ratios; CI, confidence interval; ITT, intention-to-treat analysis; PP, per-protocol analysis; mITT, modified intention-to-treat analysis.

Bold indicates the overall inference for each potential risk of bias elements. Italics are used to distinguish different types of outcomes.

[a] Studies basing on between-trial comparisons started from moderate certainty; studies basing on with-trial comparisons started from high certainty.

[b] Rated down twice for imprecision.

[c] Rated down three levels for imprecision.

[d] Rated down once for indirectness (i.e., 10–20 comparisons).

[e] Rated down twice for indirectness (i.e., less than 10 comparisons).

[f] Rated down because the result was inconsistent with hypothesized direction (risk of bias leads to effect overestimation).

[g] Subgroup analysis suggested moderate credibility subgroup effect.

[h] Rated down once for imprecision.

multiple studies. Thus, our conduct of meta-analysis of results represents a major strength.

We developed a carefully thought-out, detailed analysis plan. We separately conducted analyses for studies based on between- and within-trial comparisons. To utilize as much data as possible, we separately conducted analyses for different comparisons. To reduce confounding, we selected adjusted results in preference of unadjusted, and developed explicit, detailed rules for selection of results into analyses (Appendix 4).

We are the first to differentiate meta-epidemiological studies using within- and between-trial comparisons, a crucial issue in determining credibility of subgroup effects based on risk of bias.

We are the first to assess the certainty of evidence from meta-epidemiological studies. Our approach applies well-established principles of GRADE [19] and ICEMAN [9].

This systematic survey has limitations. Overlap of meta-analyses or trials across meta-epidemiological studies leads to possible double counting in our survey, which may influence our results, possibly by resulting in narrower CIs than would otherwise be the case. We did not quantify the extent of overlap in this survey; however, the BRANDO project, which combined data from seven meta-epidemiological studies, identified that 21% meta-analyses (16% trials) were included in more than one meta-epidemiological study [40].

Conducting separate analyses for different comparisons avoids possibly inappropriate pooling, the decision comes at the expense of a possible loss of precision. However, our approach to making overall inference has taken into account the situation in which all comparisons reported results supporting the same inference, in which we rated certainty for overall inference on the basis of the highest certainty comparison.

A fundamental limitation of all meta-epidemiological studies is that they do not address whether particular methodological limitations cause bias, but rather whether they consistently cause bias in the same direction. Were it the case that a particular limitation resulted in overestimates of effects approximately half the time and underestimates the other half, a meta-epidemiological study would conclude little or no bias. The conclusions of meta-epidemiological studies are always limited in this regard.

Another limitation is that the approach we used for assessing certainty of evidence is not developed specifically or validated for meta-epidemiological studies, but rather it was an adaptation of existing approaches which are developed for other contexts.

Studying heterogeneity of treatment effect estimates is one approach to address the above-described limitation of examining average biases. Such heterogeneity would emanate from biases vacillating from effect overestimation to underestimation. However, most meta-epidemiological studies have not studied heterogeneity.

The definition of double blinding and intention-to-treat analysis differs in meta-epidemiological studies. Although the term double-blinding is very widely used, the varying definitions limit the value of this comparison. For missing outcome data, the comparisons used arbitrary missing proportion thresholds, and no meta-epidemiological study distinguished differential vs. nondifferential missing data. Another limitation is we did not register this systematic survey in advance.

Four previous surveys of meta-epidemiological studies have investigated impact of risk of bias elements

[10–13]. Our results are consistent with them in that all concluded evidence supports overestimation as a result of inadequate random sequence generation or allocation concealment. Our results are also similar to the one survey that generated a pooled estimate [11] in the conclusion of possible effect overestimation with lack of or unclear blinding of patients or outcome assessors for subjective outcomes. Our results, however, also suggested possible effect overestimation in trials with imbalanced cointerventions and trials stopped early for benefit, and possible effect underestimation in trials with missing outcome data. Moreover, we provided an explicit rating of certainty of inferences.

Evidence from meta-epidemiological studies has established the likely average impact of random sequence generation and allocation concealment in reducing bias. Blinding of patients is clearly important for patient-reported outcomes and blinding of outcome assessors is crucial for subjective outcomes. Impact of blinding of patients or outcome assessors on other outcomes, and impact of blinding of other groups (health-care providers, data collectors, data analysts), remain uncertain. Trials that stopped early for benefit, and trials with imbalanced cointerventions likely exaggerate treatment effects. Missing outcome data may lead to effect underestimation. Impact of baseline imbalance, compliance, selective reporting, and intention-to-treat analysis remains uncertain.

Our results have clear and important implications for the conduct and interpretation of clinical trials. All trials can and should appropriately generate random sequence and ensure concealment through central randomization wherever possible and numbered, sealed envelopes when impossible. When outcomes are subjective, investigators should ensure wherever possible patients and outcome assessors are blinded, and acknowledge in their study limitations that overestimation of effects is likely. Trialists should document cointervention, and acknowledge the likelihood of bias if cointerventions are not balanced. Trials should not stop for benefit until a large number of events, ideally more than 500, have accrued [43].

Our empirical evidence, together with theoretical considerations, trial reporting practice, and the ease with which reviewers can make the required judgements, can inform the choice of items for risk of bias instruments.

CRediT authorship contribution statement

YW and GHG conceived the study. YW and RC designed the search strategy, and RC conducted the literature search. YW, RC, and QW screened studies for eligibility, and extracted data. SP and YW performed the statistical analysis. YW and GHG assessed the certainty of evidence. YW and GHG drafted the first version of the manuscript and all other authors revised it. All authors approved the final version of the manuscript.

Data availability

Data will be made available on request.

Declaration of competing interest

SA, DB, PR, KFS, DZ, and GHG are coauthors of one or more meta-epidemiological studies that this systematic survey included. This work was supported by the Einstein Foundation Berlin as part of the Einstein Foundation Award for Promoting Quality in Research. The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, the Einstein Foundation or the award jury. Authors do not have any personal financial interests or professional relationships related to the subject matter but not directly to this manuscript. There are no patents or copyrights licensed to the author(s) that are relevant to the work submitted for publication. No additional relationships or activities to declare.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.11.001.

References

[1] Collins R, Bowman L, Landray M, Peto R. The magic of randomization versus the myth of real-world evidence. N Engl J Med 2020;382:674–8.

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

[3] National Institute for Health and Care Excellence (NICE). Methodology checklist: randomised controlled trials. Available at https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-2549703709/chapter/appendix-c-methodology-checklist-randomised-controlled-trials. Accessed November 26, 2023.

[4] Sterne JAC, Savovic 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.

[5] Critical Appraisal Skills Programme (CASP). CASP randomised controlled trial standard checklist. Available at https://casp-uk.net/casp-tools-checklists/. Accessed November 26, 2023.

[6] Wang Y, Ghadimi M, Wang Q, Hou L, Zeraatkar D, Iqbal A, et al. Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues. J Clin Epidemiol 2022;152:218–25.

[7] Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 2002;21:1513–24.

[8] Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Control Clin Trials 1990;11:339–52.

[9] Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the instrument to assess the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ 2020;192:E901–6.

[10] Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol 2016;77:24–37.

[11] Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One 2016;11:e0159267.

[12] Wang H, Song J, Lin Y, Dai W, Gao Y, Qin L, et al. Trial-level characteristics associate with treatment effect estimates: a systematic review of meta-epidemiological studies. BMC Med Res Methodol 2022;22:171.

[13] Berkman ND, Santaguida PL, Viswanathan M, Morton SC. The empirical evidence of bias in trials measuring treatment differences. Rockville (MD): Agency for Healthcare Research and Quality; 2014.

[14] Amer MA, Herbison GP, Grainger SH, Khoo CH, Smith MD, McCall JL. A meta-epidemiological study of bias in randomized clinical trials of open and laparoscopic surgery. Br J Surg 2021;108:477–83.

[15] Armijo-Olivo S, da Costa BR, Ha C, Saltaji H, Cummings GG, Fuentes J. Are biases related to attrition, missing data, and the use of intention to treat related to the magnitude of treatment effects in physical therapy trials?: a meta-epidemiological study. Am J Phys Med Rehabil 2022;101:520–9.

[16] Haring R, Ghannad M, Bertizzolo L, Page MJ. No evidence found for an association between trial characteristics and treatment effects in randomized trials of testosterone therapy in men: a meta-epidemiological study. J Clin Epidemiol 2020;122:12–9.

[17] Moustgaard H, Clayton GL, Jones HE, Boutron I, Jorgensen L, Laursen DLT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ 2020;368:l6802.

[18] Wang Z, Alahdab F, Farah M, Seisa M, Firwana M, Rajjoub R, et al. Association of study design features and treatment effects in trials of chronic medical conditions: a meta-epidemiological study. BMJ Evid Based Med 2022;27:104–8.

[19] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.

[20] Sierma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. Stat Med 2007;26:2745–58.

[21] Welton NJ, Ades AE, Carlin JB, Altman D, Sterne J. Models for potentially biased evidence in meta-analysis using empirically based priors. J R Stat Soc Ser A Stat Soc 2009;172:119–36.

[22] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–12.

[23] Rhodes KM, Mawdsley D, Turner RM, Jones HE, Savovic J, Higgins JPT. Label-invariant models for the analysis of meta-epidemiological data. Stat Med 2018;37:60–70.

[24] Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med 2000;19:3127–31.

[25] Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Ann Intern Med 2014;160:267–70.

[26] Abraha I, Cherubini A, Cozzolino F, De Florio R, Luchetta ML, Rimland JM, et al. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ 2015;350:h2445.

[27] Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in physical therapy trials and its association with treatment effects: a meta-epidemiological study. Am J Phys Med Rehabil 2017;96:34–44.

[28] Armijo-Olivo S, Saltaji H, da Costa BR, Fuentes J, Ha C, Cummings GG. What is the influence of randomisation sequence generation and allocation concealment on treatment effects of physical therapy trials? A meta-epidemiological study. BMJ Open 2015;5:e008562.

[29] Bolvig J, Juhl CB, Boutron I, Tugwell P, Ghogomu EAT, Pardo JP, et al. Some Cochrane risk-of-bias items are not important in osteoarthritis trials: a meta-epidemiological study based on Cochrane reviews. J Clin Epidemiol 2018;95:128–36.

[30] Chaimani A, Vasiliadis HS, Pandis N, Schmid CH, Welton NJ, Salanti G. Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study. Int J Epidemiol 2013;42:1120–31.

[31] Hartling L, Hamm MP, Fernandes RM, Dryden DM, Vandermeer B. Quantifying bias in randomized controlled trials in child health: a meta-epidemiological study. PLoS One 2014;9.

[32] Herbison P, Hay-Smith J, Gillespie WJ. Different methods of allocation to groups in randomized trials are associated with different levels of bias. A meta-epidemiological study. J Clin Epidemiol 2011;64:1070–5.

[33] Martin GL, Trioux T, Gaudry S, Tubach F, Hajage D, Dechartres A. Association between lack of blinding and mortality results in critical care randomized controlled trials: a meta-epidemiological study. Crit Care Med 2021;49:1800–11.

[34] Mostazir M, Taylor G, Henley WE, Watkins ER, Taylor RS. Per-Protocol analyses produced larger treatment effect sizes than intention to treat: a meta-epidemiological study. J Clin Epidemiol 2021;138:12–21.

[35] Nuesch E, Reichenbach S, Trelle S, Rutjes AWS, Liewald K, Sterchi R, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. Arthritis Rheum 2009;61:1633–41.

[36] Nuesch E, Trelle S, Reichenbach S, Rutjes AW, Burgi E, Scherer M, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ 2009;339:b3244.

[37] Papageorgiou SN, Xavier GM, Cobourne MT. Basic study design influences the results of orthodontic clinical investigations. J Clin Epidemiol 2015;68:1512–22.

[38] Saltaji H, Armijo-Olivo S, Cummings GG, Amin M, da Costa BR, Flores-Mir C. Influence of blinding on treatment effect size estimate in randomized controlled trials of oral health interventions. BMC Med Res Methodol 2018;18:42.

[39] Saltaji H, Armijo-Olivo S, Cummings GG, Amin M, da Costa BR, Flores-Mir C. Impact of selection bias on treatment effect size estimates in randomized trials of oral health interventions: a meta-epidemiological study. J Dent Res 2018;97:5–13.

[40] Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess 2012;16:1–82.

[41] Savovic J, Turner RM, Mawdsley D, Jones HE, Beynon R, Higgins JPT, et al. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: the ROBES meta-epidemiologic study. Am J Epidemiol 2018;187:1113–22.

[42] Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973–82.

[43] Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 2010;303:1180–7.

[44] Bialy L, Vandermeer B, Lacaze-Masmonteil T, Dryden DM, Hartling L. A meta-epidemiological study to examine the association between bias and treatment effects in neonatal trials. Evid Based Child Health 2014;9:1052–9.

[45] Dechartres A, Altman DG, Trinquart L, Boutron I, Ravaud P. Association between analytic strategy and estimates of treatment outcomes in meta-analyses. JAMA 2014;312:623–30.
[46] Dello Russo C, Cappoli N, Navarra P. A comparison between the assessments of progression-free survival by local investigators versus blinded independent central reviews in phase III oncology trials. Eur J Clin Pharmacol 2020;76:1083–92.
[47] Hempel S, Miles J, Suttorp MJ, Wang Z, Johnsen B, Morton S, et al. Detection of associations between trial quality and effect sizes. Rockville (MD): Agency for Healthcare Research and Quality; 2012.
[48] Hopewell S. Impact of grey literature on systematic reviews of randomized trials [PhD thesis] 2004.
[49] Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol 2014;43:1272–83.
[50] Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ 2012;344:e1119.
[51] Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ 2013;185:E201–11.
[52] Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Rasmussen JV, Hilden J, et al. Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. Int J Epidemiol 2014;43:937–48.
[53] Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. Fertil Steril 1996;65:939–45.
[54] Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, et al. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. Health Technol Assess 1999;3:1–98. i-iv.
[55] Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609–13.
[56] Ndounga Diakou LA, Trinquart L, Hrobjartsson A, Barnes C, Yavchitz A, Ravaud P, et al. Comparison of central adjudication of outcomes and onsite outcome assessment on treatment effect estimates. Cochrane Database Syst Rev 2016;3:MR000043.
[57] Unverzagt S, Prondzinsky R, Peinemann F. Single-center trials tend to provide larger treatment effects than multicenter trials: a systematic review. J Clin Epidemiol 2013;66:1271–80.
[58] Zeraatkar D, Pitre T, Diaz-Martinez JP, Chu D, Rochwerg B, Lamontagne F, et al. Effects of allocation concealment and blinding in trials addressing treatments for COVID-19: A methods study. Am J Epidemiol 2023;192:1678–87.
[59] Niederer D, Weippert M, Behrens M. What modifies the effect of an exercise treatment for chronic low back pain? A meta-epidemiologic regression analysis of risk of bias and comparative effectiveness. J Orthop Sports Phys Ther 2022;52:792–802.
[60] Stadelmaier J, Roux I, Petropoulou M, Schwingshackl L. Empirical evidence of study design biases in nutrition randomised controlled trials: a meta-epidemiological study. BMC Med 2022;20:330.
[61] Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336:601–5.

Appendix 1. PRISMA Checklist

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	P1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	P3-4
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	P5-6
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	P6
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	P7-9
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	P7
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	P7, appendix 1
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	P7
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	P9-10
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	P9-10
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	P9-10
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	NA
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	P10
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	P10-11
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	P10-11
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	P10-11
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	P10-11
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	P11-12
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	NA
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	NA
Certainty	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	P12-14

PRISMA 2020 Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
assessment			
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	P14, Figure 1
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	P14
Study characteristics	17	Cite each included study and present its characteristics.	P15, Appendix 5
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	NA
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	P15-20
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	P15-20
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	P15-20
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	P15-20
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	P15-20
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	NA
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	P15-20, Table 2
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	P21
	23b	Discuss any limitations of the evidence included in the review.	P23-24
	23c	Discuss any limitations of the review processes used.	P23-24
	23d	Discuss implications of the results for practice, policy, and future research.	P25-26
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	P24
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	NA
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	NA
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	P36
Competing interests	26	Declare any competing interests of review authors.	P36
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	P36

From: 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. doi:

Appendix 2. Protocol

Empirical Evidence of Potential Risk of Bias Issues in Randomized Trials: Protocol for a Systematic Survey of Meta-Epidemiological Studies

1. Introduction

Researchers generally agree that randomized controlled trials (RCTs) represent the best study design for assessing effects of health care interventions [1]. Nevertheless, flaws in design and execution of RCTs may lead to overestimation or underestimation of the treatment effects; that is, biased [2].

Empirical evidence of how risk of bias influences effect estimates in RCTs comes primarily from meta-epidemiological studies [3]. Meta-epidemiological studies typically collect large numbers of meta-analyses of RCTs and, within meta-analyses, compare effect estimates between trials with and without a methodological safeguard against risk of bias [4,5]. Another type of meta-epidemiological study collects a large number of RCTs, and within trials, compares effect estimates between sub-studies without the characteristic and sub-studies with the characteristic [6]. The latter type using within-trial comparisons provides more credible evidence than the former type using between-trial comparisons. An important concern with between-trial comparisons is study-level confounding: an association observed between a risk of bias issue and effect estimate may be confounded by other differences between trials [7].

Meta-epidemiological studies can be collected and pooled together in a systematic survey which, if conducted optimally, could provide best available empirical evidence regarding risk of bias. Investigators have conducted three systematic surveys of meta-epidemiological studies: one published in 2014 [8] and two in 2016 [3,9], two of which did not pool results from meta-epidemiological studies [8,9]. The study that generated a pooled estimate from meta-epidemiological studies selected the comparison high/unclear versus low risk of bias ahead of other comparisons (e.g., high versus low risk of bias), and thus failed to utilize all the existing evidence; it also failed to

differentiate within-trial studies and between-trial studies comparisons [3]. Furthermore, several new meta-epidemiological studies have been published in recent years [10-14]. Thus, our aim is to update the systematic survey of meta-epidemiological studies related to potential risk of bias issues in RCTs.

2. Methods

2.1. Data sources and searches

We will review the meta-epidemiological studies included in the 3 existing surveys [3,8,9] and assess their eligibility, and in addition, conduct an electronic literature search of Medline, Embase, Web of Science, and the Cochrane Database of Systematic Reviews from 2015 until the present (the two latest existing systematic surveys searched up to 2015). Our study team will review the reference lists of all eligible meta-epidemiological studies to identify other potential eligible studies.

2.2. Inclusion criteria

Two reviewers will independently screen abstracts followed by full texts in Covidence. They will resolve disagreement by consensus or involving a third reviewer.

We will include meta-epidemiological studies investigating the influence of potential risk of bias issues on effect estimates in RCTs of any health care topic. Meta-epidemiological studies can be based on a collection of systematic reviews and meta-analyses of RCTs (i.e., between-trial comparisons), or based on a collection of RCTs (i.e., within-trial comparisons). We will include only meta-epidemiological studies that preserved the clustering of RCTs within meta-analyses (for between-trial comparisons) or the clustering of sub-studies within trials (for within-trial comparisons). For meta-epidemiological studies preserving the clustering of RCTs within meta-analyses, the comparisons are only made between RCTs within meta-analyses which have same/similar population, comparison and outcome. For meta-epidemiological studies preserving the clustering of sub-studies within trials, the comparisons are only made between sub-studies within trials which have same population, comparison and

outcome.

Eligible meta-epidemiological studies should analyze the influence of risk of bias issues quantitatively as a difference in effect estimates and combine across the collection to obtain a single result. Most commonly used effect measures (i.e., difference in effect estimates or average bias) include:

- Ratio of odds ratios (ROR) = $OR_{methodologically\ inferior} / OR_{superior}$
- Ratio of hazard ratios (RHR) = $HR_{methodologically\ inferior} / HR_{superior}$
- Ratio of risk ratios (RRR) = $RR_{methodologically\ inferior} / RR_{superior}$
- Difference in standardized mean differences (dSMD) = $SMD_{methodologically\ inferior} - SMD_{superior}$

We collected potential risk of bias issues from three resources: items of existing RCT risk of bias instruments (a systematic survey), items regarded as "other bias" in systematic reviews [15], and comments on Cochrane risk of bias tool 1.0 [16]. Through a survey of experts, we classified the items into three categories: items that the majority thought are addressing risk of bias (category 1); items that the majority thought are not addressing risk of bias (category 2); items with substantial disagreement regarding whether they are addressing risk of bias (category 3).

Items in category 1 and category 3 are the potential risk of bias issues on which we will focus in this systematic survey:

- Random sequence generation (category 1)
- Allocation concealment (category 1)
- Baseline imbalance (category 3)
- Blinding of health care providers (category 1)
- Blinding of participants (category 1)
- Blinding of data collectors (category 1)
- Blinding of outcome assessors (category 1)

- Blinding of data analysts (category 1)
- Double blinding (including some typically unspecified combination of the blinding categories above) (category 1)
- Imbalance in co-interventions (category 3)
- Failure in implementing the intervention as intended or non-adherence to the assigned intervention (category 3)
- Outcome measurement differs between groups (category 1)
- Follow-up time differs between groups (category 1)
- Missing outcome data (category 1)
- Selective outcome reporting (category 1 and 3)
- Intention-to-treat analysis (category 1 and 3)
- Early termination: stop early for benefit, stop early for harm, stop early for futility) (category 3)

2.3. Exclusion criteria

We will exclude:

- Single systematic reviews and meta-analyses that presented meta-regression, subgroup or sensitivity analyses based on risk of bias issues;
- Meta-epidemiological studies that did not preserve the clustering of RCTs within meta-analyses or the clustering of sub-studies within trials (e.g., we excluded the studies that collected a large number of RCTs and compared effect estimates between trials with a characteristic and trials without a characteristic. These studies provide much less credible results because of the substantial difference between trials, compared with studies preserving clustering);
- Meta-epidemiological studies examining only the association between quality or risk of bias scores or overall risk of bias and effect estimates;
- Meta-epidemiological studies examining association between non-risk of bias characteristics and effect estimates;
- Meta-epidemiological studies comparing effect estimates between RCTs and observational studies;

- Meta-epidemiological studies addressing non-health care topics (e.g. animal experimental studies);
- Protocols or conference abstracts.

2.4. Data extraction

Paired reviewers will independently extract data using a standardized, pre-designed data extraction form. They will resolve disagreement by consensus or by involving a third reviewer.

We will extract the following information from each meta-epidemiological study:

- Based on meta-analyses (between-trial comparisons) or trials (within-trial comparisons)
- Number of included systematic reviews, meta-analyses, trials, and participants
- Identification of systematic reviews/trials: source (database and year), eligibility criteria, topic area, sampling frame
- Choice of meta-analyses within systematic reviews (a systematic review may include several meta-analyses with different interventions or outcomes, how authors chose including which meta-analysis in their analysis)
- Statistical approach/model (common models that preserve clustering include: meta-meta-analytic approach (i.e. two-step approach) [4], multivariable multilevel model [17], Bayesian hierarchical model [18], linear or logistic regression model preserving clustering [19], label-invariant model [20])
- Focusing on only undesirable outcomes (e.g. death), desirable outcomes (e.g. survival), or both
- Approach to distinguishing between intervention and control groups
- Approach to dealing with overlap of RCTs across meta-analyses
- Risk of bias issues and assessment: risk of bias instrument used, source (assess themselves, or use assessment from existing systematic reviews)
- Type of comparison (e.g., high/unclear versus low risk of bias, high versus low, unclear versus low, high versus low/unclear, definitely/probably no versus

definitely/probably yes, etc.) and definition of each category

- Type of outcome, control, intervention
- Adjustment or no adjustment for other risk of bias issues or confounders
- Difference in effect estimates (i.e., average bias): ROR, RHR, RRR, or dSMD, and 95% confidence or credible interval

2.5. Data synthesis and analysis

We will analyze our data such that a ROR/RHR/RRR <1 or dSMD <0 indicates methodological inferior trials yielded a larger effect estimate (more beneficial or less harmful effect of the experimental intervention) compared to methodological superior trials. For meta-epidemiological studies that present effects in the opposite direction (the original study compared methodological superior trials versus inferior trials, or the original study focused on desirable outcomes), we will recalculate the effect measures: inverse of the reported ROR/RHR/RRR, or 1 - reported dSMD.

To synthesize the difference in effect estimates (i.e., average bias) for binary and continuous outcomes together, we will convert dSMDs to log(RORs) by multiplying by $\pi/\sqrt{3} = 1.814$ [21]. We will combine the RORs from different meta-epidemiological studies using the DerSimonian and Laird with Hartung-Knapp adjustment random-effects model [22]. To assess inconsistency, we will inspect forest plots and calculate the I^2 statistic. We will conduct analyses for meta-epidemiological studies based on a collection of meta-analyses and meta-epidemiological studies based on a collection of RCTs separately.

Meta-epidemiological studies may report different comparisons based on how they dealt with trials with unclear risk of bias or unclear reporting: high versus low risk of bias, high/unclear versus low, high versus low/unclear, etc. We will conduct analyses for different comparisons separately. We will use the following rules for the primary analyses:

1) Type of outcome

- If a study reported results separately for objective, subjective outcome and combined both (comprehensive), we will include the comprehensive outcome;
- If a study reported results separately for objective and subjective outcomes, or for binary and continuous outcomes, but did not report results combining these subcategories, and the two types of outcome came from different systematic reviews (i.e., one systematic review only contributed once), we will include them as separate studies; if the two types of outcome came from different meta-analyses within the same systematic reviews (i.e., lack of independence), we will include the one with largest number of trials;
- We will include studies that did not specify the type of outcome;
- We will include studies that reported only objective or subjective outcome.

2) Type of control

- If a study reported results separately for active, inactive control and combined both (comprehensive), we will include the comprehensive control;
- If a study reported results separately for active and inactive controls, but did not report results combining both, and the two types of control came from different systematic reviews, we will include them as separate studies; if the two types of control came from different meta-analyses within the same systematic reviews, we will include the one with largest number of trials;
- We will include studies that did not specify the type of control;
- We will include studies that reported only active or inactive control.

3) Type of intervention

- If a study reported results separately for pharmacological, non-pharmacological intervention and combined both (comprehensive), we will include the comprehensive intervention;
- If a study reported results separately for pharmacological and non-pharmacological interventions, but did not report results combining both, and the

two types of intervention came from different systematic reviews, we will include them as separate studies; if the two types of intervention came from different meta-analyses within the same systematic review, we will include the one with largest number of trials;

- We will include studies that did not specify the type of intervention;
- We will include studies that reported only pharmacological or non-pharmacological intervention.

4) Adjusted for other risk of bias issues (consider this rule only after considering the prior three)

- If a study reported both adjusted and unadjusted results, we will include the adjusted results;
- If a study reported results adjusted for more risk of bias issues and results adjusted for less risk of bias issues, we will include the results adjusted for more;
- We will include studies that reported only unadjusted result;
- If a study, rather than adjusting for multiple issues simultaneously in a multivariable analysis, reported results adjusted for a number of issues separately, we will use the rules that follow to choose the analysis (another file provides more details):
  - For studies examining the effect of blinding on effect estimates of RCTs, we will choose the result adjusting for allocation concealment in preference to results adjusting for other issues;
  - For studies examining the effect of allocation concealment, we will choose the result adjusting for blinding of patients in preference to results adjusting for other issues;
  - For studies examining the effect of missing outcome data, we will choose the result adjusting for blinding in preference to results adjusting for other issues.

We will conduct the following subgroup analyses: objective versus subjective outcome, active versus inactive control, and adjusted versus unadjusted result. If applicable, we

will separately conduct both within-study comparisons (e.g., if a meta-epidemiological study reported results for both objective and subjective outcome, we can compare them within the study) and between-study comparisons subgroup analyses [7]. We will use Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) to assess the credibility of subgroup effect [7].

For blinding issues (blinding of health care providers, blinding of participants, blinding of data collectors, blinding of outcome assessors, blinding of data analysts, double blinding), meta-epidemiological studies may compare trials based on risk of bias judgement (e.g., high versus low risk of bias), or compare trials based on what happened, i.e., whether blinding has been implemented or not (unblinded versus blinded). We will first conduct subgroup analyses. If the results suggest no difference between subgroups or the credibility of subgroup effect is low or very low, we will combine them together. In that case, if a study reported both, we will include the comparison based on whether blinding has been implemented or not.

For missing outcome data, some meta-epidemiological studies may compare trials based on the proportion of missing (e.g., ≥20% versus <20%). We will conduct subgroup analysis first. If the results suggest no difference between subgroups or the credibility of subgroup effect is low or very low, we will combine them together; otherwise, we will report them separately.

For intention-to-treat analysis, meta-epidemiological studies may compare trials based on the analytic approach, e.g., per-protocol analysis versus intention-to-treat, deviation from intention-to-treat versus intention-to-treat, deviation from intention-to-treat/unclear versus intention-to-treat, etc. We will conduct subgroup analysis first. If the results suggest no difference between subgroups or the credibility of subgroup effect is low or very low, we will combine them together; otherwise, we will report them separately.

References

1. Verhagen AP, de Vet HC, de Bie RA, Boers M, van den Brandt PA. The art of quality assessment of RCTs included in systematic reviews. J Clin Epidemiol 2001;54:651-4.
2. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343:d5928.
3. Page MJ, Higgins JP, Clayton G, Sterne JA, Hrobjartsson A, Savovic J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLoS One 2016;11:e0159267.
4. Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Statistics in medicine 2002;21:1513-24.
5. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Controlled clinical trials 1990;11:339-52.
6. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ 2012;344:e1119.
7. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ 2020;192:E901-E6.
8. Berkman ND, Santaguida PL, Viswanathan M, Morton SC. *The Empirical Evidence of Bias in Trials Measuring Treatment Differences*. Rockville (MD)2014.
9. Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol 2016;77:24-37.
10. Armijo-Olivo S, Dennett L, Arienti C, Dahchi M, Arokoski J, Heinemann AW, et al. Blinding in Rehabilitation Research: Empirical Evidence on the Association Between Blinding and Treatment Effect Estimates. American journal of physical medicine & rehabilitation 2020;99:198-209.
11. Mostazir M, Taylor G, Henley WE, Watkins ER, Taylor RS. Per-Protocol analyses produced larger treatment effect sizes than intention to treat: a meta-epidemiological study. J Clin Epidemiol 2021;138:12-21.
12. Moustgaard H, Clayton GL, Jones HE, Boutron I, Jorgensen L, Laursen DRT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ 2020;368:l6802.
13. Martin GL, Trioux T, Gaudry S, Tubach F, Hajage D, Dechartres A. Association Between Lack of Blinding and Mortality Results in Critical Care Randomized Controlled Trials: A Meta-Epidemiological Study. Crit Care Med 2021;49:1800-11.
14. Haring R, Ghannad M, Bertizzolo L, Page MJ. No evidence found for an association between trial characteristics and treatment effects in randomized trials of testosterone therapy in men: a meta-epidemiological study. J. Clin. Epidemiol. 2020;122:12-9.
15. Babic A, Pijuk A, Brazdilova L, Georgieva Y, Raposo Pereira MA, Poklepovic Pericic T, et al. The judgement of biases included in the category "other bias" in Cochrane systematic reviews of interventions: a systematic survey. BMC medical research methodology

2019;19:77.

16. Jorgensen L, Paludan-Muller AS, Laursen DR, Savovic J, Boutron I, Sterne JA, 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. Systematic reviews 2016;5:80.
17. Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. Statistics in medicine 2007;26:2745-58.
18. Welton NJ, Ades AE, Carlin JB, Altman D, Sterne J. Models for potentially biased evidence in meta-analysis using empirically based priors. J.R. Statist. Soc. A 2009;172:119-36.
19. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Jama 1995;273:408-12.
20. Rhodes KM, Mawdsley D, Turner RM, Jones HE, Savovic J, Higgins JPT. Label-invariant models for the analysis of meta-epidemiological data. Statistics in medicine 2018;37:60-70.
21. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Statistics in medicine 2000;19:3127-31.
22. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Annals of internal medicine 2014;160:267-70.

Appendix 3 Search strategy

MEDLINE Database: OVID Medline Epub Ahead of Print, In-Process & Other Non-Indexed Citations, Ovid MEDLINE(R) Daily and Ovid MEDLINE(R) 1946 to Present
Search Strategy:
---

1 meta-epidemiolog*.ti,ab.
2 metaepidemiolog*.ti,ab.
3 1 or 2 (321)
4 ((comparison* or impact* or influence*) and (intervention effect* or treatment effect* or "study characteristics" or conclusion*)).ti.
5 (investigat* and bias).ti.
6 or/1-5
7 (meta-meta-anal$ or meta-review$ or meta-epidemiologic$ or metaepidemiologic$).ti,ab.
8 6 or 7
9 limit 8 to yr="2015-2023"

EMBASE (OVID)
Search Strategy:
---

1 meta-epidemiolog*.ti,ab.
2 metaepidemiolog*.ti,ab.
3 1 or 2
4 ((comparison* or impact* or influence*) and (intervention effect* or treatment effect* or "study characteristics" or conclusion*)).ti.
5 (investigat* and bias).ti.
6 or/1-5
7 (meta-meta-anal$ or meta-review$ or meta-epidemiologic$ or metaepidemiologic$).ti,ab.
8 6 or 7
9 limit 8 to yr="2015 -Current"

Web of Science
4 #1 OR #2 OR #3
3 ALL=(meta-meta-anal$ or meta-review$ or meta-epidemiologic$ or metaepidemiologic$)
2 ALL=(metaepidemiolog*)
1 meta-epidemiolog* (All Fields)

Cochrane Library
ID Search Hits

#1	(meta-epidemiol*):ti,ab,kw (Word variations have been searched)
#2	(metaepidemiol*):ti,ab,kw (Word variations have been searched)
#3	(meta-meta-anal*):ti,ab,kw (Word variations have been searched)
#4	(meta-review*):ti,ab,kw (Word variations have been searched)
#5	#1 or #2 or #3 or #4

Appendix 4 Rules for selection of results into primary analyses

1) Type of outcome

- If a study reported results separately for objective, subjective outcome and combined both (comprehensive), we included the comprehensive outcome;
- If a study reported results separately for objective and subjective outcomes, or for binary and continuous outcomes, but did not report results combining these subcategories, and the two types of outcome came from different systematic reviews (i.e., one systematic review only contributed once), we included them as separate studies; if the two types of outcome came from different meta-analyses within the same systematic reviews (i.e., lack of independence), we included the one with largest number of trials;
- We included studies that did not specify the type of outcome;
- We included studies that reported only objective or subjective outcome.

2) Type of control

- If a study reported results separately for active, inactive control and combined both (comprehensive), we included the comprehensive control;
- If a study reported results separately for active and inactive controls, but did not report results combining both, and the two types of control came from different systematic reviews, we included them as separate studies; if the two types of control came from different meta-analyses within the same systematic reviews, we included the one with largest number of trials;
- We included studies that did not specify the type of control;
- We included studies that reported only active or inactive control.

3) Type of intervention

- If a study reported results separately for pharmacological, non-pharmacological

intervention and combined both (comprehensive), we included the comprehensive intervention;

- If a study reported results separately for pharmacological and non-pharmacological interventions, but did not report results combining both, and the two types of intervention came from different systematic reviews, we included them as separate studies; if the two types of intervention came from different meta-analyses within the same systematic review, we included the one with largest number of trials;
- We included studies that did not specify the type of intervention;
- We included studies that reported only pharmacological or non-pharmacological intervention.

4) Adjusted for other risk of bias elements (consider this rule only after considering the prior three rules)

- If a study reported both adjusted and unadjusted results, we included the adjusted results;
- If a study reported results adjusted for more risk of bias elements and results adjusted for less risk of bias elements, we included the results adjusted for more;
- We included studies that reported only unadjusted result;
- If a study, rather than adjusting for multiple elements simultaneously in a multivariable analysis, reported results adjusted for a number of elements separately, we used the below rules to choose results:
  - For studies examining the influence of blinding on effect estimates of randomized trials, we chose the result adjusting for allocation concealment in preference to results adjusting for other issues. Logically, we anticipate a stronger association between blinding and allocation concealment than the association between blinding and other issues because blinded drug trials with medication prepared by the pharmacy are always concealed. This is consistent with the pattern of association in most meta-epidemiological studies.
  - For studies examining the influence of allocation concealment, we chose the result

adjusting for blinding of patients in preference to results adjusting for other issues. Logically, we anticipate a stronger association between allocation concealment and blinding of patients or health care providers than the association between concealment and other issues because blinded drug trials with medication prepared by the pharmacy are always concealed. Only one meta-epidemiological study reported effect of concealment adjusting for blinding of patients and intention-to-treat analysis separately. The data suggests that the association between concealment and blinding of patients ($p = 0.008$) may be stronger than the association between concealment and intention-to-treat analysis ($p = 0.07$), but we cannot exclude the probability of chance.

- For studies examining the influence of missing outcome data, we chose the result adjusting for blinding in preference to results adjusting for other issues. Logically we anticipate a stronger association between missing outcome data and blinding than the association between missing outcome data and other issues. Only one meta-epidemiological study reported effect of missing outcome data adjusting for blinding of patients and allocation concealment separately. Although the data suggests a possible stronger association between missing outcome data and concealment ($p = 0.07$) than the association between missing outcome data and blinding of patients ($p = 0.43$), we cannot exclude the probability of chance.

Appendix 5 Methods of subgroup analyses

We performed the following subgroup analyses:

1) Subgroup analyses based on type of outcome. For blinding of healthcare providers, we tested whether its influence differs on outcomes determined by the healthcare providers versus others (e.g., patient- or observer-reported outcomes); for blinding of patients, whether its influence differs on patient-reported outcomes versus others (e.g., observer-reported outcomes); for other potential risk of bias elements, objective versus subjective outcomes.

2) Subgroup analyses based on type of control (active versus inactive control). For the type of outcome and control, we followed the classification in the original meta-epidemiological studies. Most meta-epidemiological studies classified placebo and no intervention, some also classified standard care, as inactive control.

3) Subgroup analyses based on adjustment (adjusted versus unadjusted results).

4) For blinding elements, we also performed subgroup analysis of risk of bias judgement versus blinding status.

If applicable, we separately conducted within-meta-epidemiological study subgroup analyses (e.g., a meta-epidemiological study reported results for both objective and subjective outcomes) and between-meta-epidemiological study subgroup analyses (e.g., some studies reported objective outcomes and some others reported subjective outcomes). When multiple studies included within-study subgroup analyses, to obtain a single interaction coefficient, we used a multilevel meta-regression model that accounts for the clustering of subgroups within each meta-epidemiological study.

Appendix 6 Characteristics of included meta-epidemiological studies

Table 1 Characteristics of meta-epidemiological studies based on between-trial comparisons

Study ID	Number of included SRs/MAs	Number of included trials	Number of patients	Source of SRs	Eligibility criteria of SRs	Topic area of SRs	Sampling frame 1=All SRs 2=Randomly sampled 3=Others	Choice of MAs within SRs	Statistical approach	Type of outcomes 1=Binary 2=Continuous 3=Both - reported separately 4=Both - combined together	Focusing on 1=Undesirable outcomes 2=Desirable outcomes	Characteristics assessed	Approach to distinguishing between intervention and control groups	Approach to dealing with overlap of trials across MAs	Type of outcomes (%)	Type of controls (%)	Type of interventions (%)
Abraha 2015	43/50	310	NR	PubMed (2006-2010)	SRs of therapeutic or preventive interventions, with ≥1 MAs with categorical data, each with ≥2 RCTs	Any	2	MAs with primary outcome	Weighted linear regression model with multivariable multilevel analysis, Meta-meta-analytic approach	1	1	Intention-to-treat analysis	Excluded if unclear	NR	38 (objective), 62 (subjective)	68 (placebo), 32 (non-placebo)	100 (pharmacological)
Amer 2021	9 procedures in 50 SRs	318	NR	CDSR (inception-Mar 2015)	SRs of RCTs comparing a laparoscopic and open approach to an abdominal general surgical procedure	Abdominal surgical procedures	1	Analyzed for different outcomes separately	Meta-meta-analytic approach	3	1	Random sequence generation, allocation concealment, double blinding, blinding of participants, blinding of health care providers, blinding of data collectors, blinding of outcome assessors	Control: open approach	No overlap	Reported for different outcomes separately	100 (open approach)	100 (laparoscopic)
Armijo-Olivo 2015	43/43	393	44622	CDSR (Jan 2005-May 2011)	MAs including ≥3 RCTs comparing ≥2 interventions, with ≥1 interventions being part of physical therapy and	Physical therapy	2	MAs with main outcome of interest or MAs with largest number of trials	Meta-meta-analytic approach	2	NR	Random sequence generation, allocation concealment	Followed the classification in SRs	Trials were only included once in the MA with fewer number of trials	Mixed	7.89 (inactive), 92.11 (active)	100 (physical therapy)

					main outcome or outcome with largest number of RCTs was continuous												
Armijo-Olivo 2017	43/43	393	44622	CDSR (Jan 2005-May 2011)	MAs including ≥3 RCTs comparing ≥2 interventions, with ≥1 interventions being part of physical therapy and main outcome or outcome with largest number of RCTs was continuous	Physical therapy	2	MAs with main outcome of interest or MAs with largest number of trials	Meta-meta-analytic approach	2	NR	Blinding of participants, blinding of outcome assessors, blinding of data analysts	Followed the classification in SRs	Trials were only included once in the MA with fewer number of trials	Mixed	7.89 (inactive), 92.11 (active)	100 (physical therapy)
Armijo-Olivo 2022	43/43	393	44622	CDSR (Jan 2005-May 2011)	MAs including ≥3 RCTs comparing ≥2 interventions, with ≥1 interventions being part of physical therapy and main outcome or outcome with largest number of RCTs was continuous	Physical therapy	2	MAs with main outcome of interest or MAs with largest number of trials	Meta-meta-analytic approach	2	NR	Missing outcome data, intention-to-treat analysis	Followed the classification in SRs	Trials were only included once in the MA with fewer number of trials	Mixed	7.89 (inactive), 92.11 (active)	100 (physical therapy)
Balk 2002	26/26	276	NR	Cardiovascular MAs in a previous analysis; MAs for other areas: Medline (1996-2000), CDSR (issue 4, 2000)	MAs including ≥6 RCTs with dichotomous outcomes, and had significant between-study heterogeneit	Cardiovascular disease, infectious disease, pediatrics, surgery	1	MAs with largest number of trials or that were most clearly defined	Bayesian hierarchical model (model 3) with non-informative prior distribution	1	1	Blinding of participants, blinding of health care providers, blinding of outcome assessors, intention-to-treat	NR	NR	NR	14.86 (placebo)	75.72 (pharmacological), 24.28 (surgical)

					y in the OR scale							analysis, baseline imbalance, missing outcome data						
Bassler 2010 (STOPIT-2)	63 questions	515	NR	Truncated RCTs: Medline, Embase, Current Contents, full-text journal content databases (inception-Jan 2007), hand search; SRs: CDSR, Database of Abstracts of Reviews of Effects, Medline (inception-Jan 2008)	Truncated RCTs: RCTs reported as having stopped earlier than initially planned owing to interim results in favor of the intervention SRs: excluded matching SRs without a methods section and without a literature search (at least Medline); if a SR was published prior to the matching truncated RCT, updated it; eligible nontruncated RCTs addressed the outcome that led to the early termination of the truncated RCT	Any	1	Eligible nontruncated RCTs were those addressed the outcome that led to the early termination of the truncated RCT and stated clearly that allocation was randomized	Meta-meta-analytic approach	4	NR	Stop early for benefit	NR	No overlap	NR	64.66 (placebo), 11.84 (active medication), NR (23.50)	86.80 (pharmacological), 11.84 (non-pharmacological therapeutic), 1.36 (nontherapeutic)
Bialy 2014	25/25	208	NR	Neonatal Review Group in CDSR (inception-issue 2, 2009)	SRs pertaining to infants in treatment areas	Neonatal (surfactant, corticosteroids, indomethacin, ibuprofen, nitric oxide	1	MAs with primary outcome or the outcome listed first in methods section	Meta-meta-analytic approach	1	1	Random sequence generation, allocation concealment, blinding of participants,	NR	Removed duplicate trials from the MA with largest number of trials	60 (mortality)	61.54 (inactive), 29.81 (active), 8.65 (active/inactive)	NR

						and head/total body cooling)						blinding of health care providers, blinding of outcome assessors, missing outcome data, selective reporting					
Bolvig 2018	20/20	126	19052	CDSR (inception-Jan 2014)	Cochrane SRs with MAs of RCTs in patients with Osteoarthritis, reported results from a patient reported pain outcome comparing with sham, placebo, or no intervention control	Osteoarthritis	1	NR	Mixed-effects restricted maximum likelihood model	2	2	Blinding of participants	Control: sham, placebo, or no intervention	NR	100 (patient reported pain)	100 (sham, placebo, or no intervention control)	50 (pharmacological), 40 (non-pharmacological), 10 (surgical)
Chaimani 2013	22/22	545	NR	PubMed (inception-Mar 2011)	Star-shaped NMAs	Any	1	NR	similar to Meta-meta-analytic approach	1	both	Random sequence generation, allocation concealment, blinding of participants, blinding of outcome assessors	Included star-shaped NMAs which have common comparators	No overlap	NR	Mixed	NR
Dechartres 2014	163/163	1240	NR	Three collections of MAs assessing therapeutic interventions: 1. Ten leading journals (Jul 2008-Jan 2009, Jan-Jun 2010, CDSR (Issue 4, 2008); 2. CDSR (Jan-Jul	MAs of RCTs assessing therapeutic interventions with binary outcomes	Any	1	Collection 1: Binary outcome (preference for objective outcome). Collection 2: Binary primary outcome that first reported. Collection 3: Binary primary	Meta-meta-analytic approach	1	1	Random sequence generation, allocation concealment, double blinding, missing outcome data	NR	NR	43.56 (objective), 56.44 (subjective)	NR	NR

				2011); 3. CDSR (Apr 2012-Mar 2013)				outcome with largest number of trials								
Haring 2020	19/19	132	10725	PubMed (2008-2018)	MAs of RCTs with binary/continuous outcomes	Testosterone therapy in adult men	1	Largest MAs	Meta-meta-analytic approach	4	1	Random sequence generation, allocation concealment, double blinding, blinding of outcome assessors, missing outcome data	Control: placebo	Removed duplicate trials starting with the MAs with smallest number of trials	63.16 (objective), 36.84 (patient-reported outcomes)	100 (placebo)	100 (pharmacological - testosterone replacement therapy)
Hartling 2014	17/17	287	NR	CDSR	MAs with ≥5 and ≤40 superiority parallel RCTs with pediatric patients; addressed a question of therapeutic effectiveness	Child health	3 (SRs with largest numbers of studies)	MAs with primary outcome	Meta-meta-analytic approach	4	1	Random sequence generation, allocation concealment, double blinding, blinding of outcome assessors, missing outcome data, selective reporting, baseline imbalance	NR	Trials were retained in the MA that was randomly selected	64.71 (objective), 35.29 (subjective)	52.94 (placebo or no intervention), 11.76 (active), 35.29 (mixed)	41.18 (pharmacological), 58.82 (non-pharmacological)
Hempel 2012 (dataset 4)	13/13	149	NR	Eight cardiovascular MAs and five pediatric MAs from Balk 2002	This dataset was obtained by replicating a selection used by Balk 2002	Cardiovascular disease, pediatric	NA	MAs with primary outcome, or the outcome with largest number of trials	Meta-regressions correcting for clustering within MAs	1	1	Co-intervention imbalance, compliance	Most compared against placebo. If active versus active, guided by input from experts or followed classification in the original meta-epidemiological dataset	NR	NR	Most placebo	100 (pharmacological)
Herbison 2011	18/65	389	NR	CDSR (issue 1, 2001)	SRs with binary outcomes	Any	1	MAs of the RORs was bootstrappe	Meta-meta-analytic approach	1	NR	Allocation concealment	NR	NR	NR	NR	77.78 (pharmacological)

					with ≥10 RCTs with ≥1 had > 500 people randomized to each arm.			d to get bootstrap CIs with one MA being chosen at random from each SR									
Hopewell 2004	25/25	311	452448	CDSR (issue 3, 2003)	SRs in cancer with ≥1 MAs of which ≥1 reports from grey literature and ≥1 reports from published literature	Cancer	1	MAs with primary outcome with largest number of trials	Meta-meta-analytic approach	1	1	Random sequence generation, allocation concealment, double blinding	NR	NR	NR	NR	NR
Khan 1996	9/9	34	NR	Report of Royal Commission on New Reproductive Technologies and two other publications	MAs of RCTs in infertility (included trials with both crossover and parallel group designs)	Infertility	1	NR	Logistic regression model	1	2	Random sequence generation, allocation concealment, double blinding, missing outcome data	NR	NR	100 (pregnancy)	NR	NR
Martin 2021	36/36	467	NR	Five general medical and six critical care journals, CDSR) (Jan 2009-Mar 2019)	MAs of RCTs of critical care that assessed an intervention against placebo or standard care and including mortality as an outcome	Critical care	1	MAs of short-term (≤ 31 d) mortality with largest number of trials	Meta-meta-analytic approach	1	1	Double blinding, blinding of participants and personnel	Control: placebo or standard care	Removed MAs sharing ≥3 RCTs and kept the MA with highest number of trials	100 (mortality)	100 (placebo or standard care)	97.22 (pharmacological), 2.78 (non-pharmacological)
Moher 1998, 1999	11/11	127	10492	Own database (ten journals or CDSR) (1988-1995)	Randomly selected 12 MAs from their larger database of 491 MAs of RCTs	Digestive diseases, circulatory diseases, mental health, stroke, pregnancy and childbirth	2	MAs with largest number of trials	Logistic regression model	1	1	Random sequence generation, allocation concealment, double blinding	NR	NR	68.18 (objective), 31.82 (subjective)	NR	NR
Moustgaard 2020 (MetaBLIND)	142/142	1153	NR	CDSR (Feb 2013-Feb 2014)	MAs with ≥1 RCTs with blinding of patients, healthcare	Any	1 (blinding of patients or healthcare providers), 2 (blinding of	NR	Bayesian hierarchical model (model 3) with vague	4	1	Allocation concealment, blinding of participants, blinding of	Based on descriptions in trial reports, except when	Removed duplicate trials randomly until no	Mixed	40.14 (placebo or no treatment), 26.76	66.90 (pharmacological), 2.11 (surgical), 11.97

					providers, or outcome assessors and ≥1 RCTs without blinding		outcome assessors)		prior distribution			health care providers, blinding of outcome assessors, double blinding, allocation concealment	SRs clearly labelled the comparator as placebo, control, standard care, or treatment as usual	overlap		(standard care), 33.10 (active)	(psychosocial, behavioral, or educational), 19.01 (other)
Niederer 2022	26/26	264	NR	PubMed, CENTRAL, Embase, CINAHL (inception-Jan 2021)	MAs of randomized controlled exercise studies intended to reduce pain in patients with chronic nonspecific low back pain	Exercise treatment for chronic low back pain	1	MAs of chronic low back pain	Multilevel meta-regression	2	1	Random sequence generation, allocation concealment, missing outcome data, selective reporting	NR	NR	100 (subjective - chronic nonspecific low back pain)	25.50 (inactive), 13.47 (passive treatments), 0.57 (medication), 52.44 (active exercise)	100 (exercise treatment)
Nuesch 2009a	14/14	167	41170	CDSR, Medline, Embase, CINAHL (inception-Nov 2007)	MAs of RCTs or quasi-RCTs in patients with osteoarthritis of the knee or hip and assessed patient reported pain comparing intervention with placebo, sham, or a non-intervention control	Osteoarthritis	1	MAs of patient reported pain	Meta-meta-analytic approach	2	1	Missing outcome data	Control: placebo, sham, or no intervention	Inflated standard errors to avoid double counting	100 (patient reported pain)	100 (placebo, sham, or no intervention)	57.14 (pharmacological), 42.86 (non-pharmacological)
Nuesch 2009b	16/16	175	41142	CDSR, Medline, Embase, CINAHL (inception-Nov 2007)	MAs of RCTs or quasi-RCTs in patients with osteoarthritis of the knee or hip and assessed patient reported pain comparing intervention with placebo, sham, or a	Osteoarthritis	1	MAs of pain intensity	Meta-meta-analytic approach	2	1	Allocation concealment, blinding of participants	Control: placebo, sham, or no intervention	NR	100 (patient reported pain)	100 (placebo, sham, or no intervention)	43.75 (pharmacological), 56.25 (non-pharmacological)

					non-intervention control											
Papageorgiou 2015	25/25	75	NR	Medline, Scopus, CDSR, Thomson Reuters Web of Knowledge, Bibliografia Brasileira de Odontologia, ADA Center for Evidence-Based Dentistry, PROSPERO, Digital Dissertations (inception-Sep 2014)	SRs in orthodontics with ≥1 MAs of interventional studies	Orthodontic	1	Multiple MAs were extracted from a SR only when the component trials or their outcomes differed	Meta-meta-analytic approach	4	1	Random sequence generation	NR	Removed duplicate until no overlap	96 (objective), 4 (subjective)	NR	NR
Saltaji 2018a	64/64	540	137957	PubMed, Embase, Medline, ISI Web of Science, CDSR, HealthSTAR (inception-May 2014)	MAs in the field of dental, oral, or craniofacial research, evaluated a therapeutic intervention related to dental specialties, examined ≥1 continuous outcome and included ≥5 RCTs	Oral health	1	MAs with primary outcome or the continuous outcome with largest number of trials	Meta-meta-analytic approach	2	NR	Random sequence generation, allocation concealment, baseline imbalance	Followed classification in SRs	Duplicate trials were only included once in the MA with fewest number of trials	10.94 (objective), 89.06 (subjective)	most inactive, 37.78% (placebo)	33.52 (pharmacological), 31.48 (surgical)
Saltaji 2018b	64/64	540	137957	PubMed, Embase, Medline, ISI Web of Science, CDSR, HealthSTAR (inception-May 2014)	MAs in the field of dental, oral, or craniofacial research, evaluated a therapeutic intervention related to dental specialties, examined	Oral health	1	MAs with primary outcome or the continuous outcome with largest number of trials	Meta-meta-analytic approach	2	NR	Blinding of participants, blinding of health care providers, blinding of outcome assessors, double blinding	Followed classification in SRs	Duplicate trials were only included once in the MA with fewest number of trials	10.94 (objective), 89.06 (subjective)	most inactive, 37.78% (placebo)	33.52 (pharmacological), 31.48 (surgical)

					≥1 continuous outcome and included												
Savovic 2012 (BRANDO)	352/363	3474	NR	Seven meta-epidemiological studies	NA	Any	NA	NA	Bayesian hierarchical model (model 3) with vague prior distribution	1	1	Random sequence generation, allocation concealment, double blinding, missing outcome data	At least two study collaborators made a consensus; excluded if unclear	Considered each set of overlapping MAs in turn; excluded MAs from each set until minimal overlap	18.80 (all-cause mortality), 15.38 (other objectively assessed), 17.95 (objectively measured with judgement), 41.88 (subjectively assessed), 6.00 (mixed)	73.50 (placebo or no treatment), 6.84 (other inactive), 18.80 (active), 0.85 (mixed)	69.23 (pharmacological), 5.98 (surgical), 5.56 (psychosocial/behavioral/educational), 19.23 (others)
Savovic 2018 (ROBES)	228/228	2443	NR	CDSR (issue 4, Apr 2011)	MAs with ≥5 RCTs, each with ≥1 event across the 2 trial arms with all 5 risk of bias domains having been assessed; compared an active intervention with a control or "older" intervention	Any	1	MAs with primary outcome with largest number of trials with largest number of participants	Bayesian hierarchical model (model 3)	1	1	Random sequence generation, allocation concealment, double blinding, missing outcome data	Excluded if unclear; the newer or more recently introduced intervention was experimental intervention and the older or standard intervention was control	No overlap	18.42 (all-cause mortality), 8.77 (other objectively assessed), 16.23 (objectively measured with judgement), 55.70 (subjectively assessed), 0.88 (mixed)	22.37 (placebo), 10.96 (no treatment), 25.44 (placebo/no treatment), 14.04 (standard care), 10.53 (standard care/placebo/no treatment), 16.67 (active)	66.23 (pharmacological), 6.14 (provision of care), 5.26 (surgical intervention or procedure), 4.82 (psychosocial and behavioral), 17.54 (others)
Schulz 1995	33/33	250	62091	Cochrane Pregnancy and Childbirth Database	First, identified an initial subset of 82 MAs with ≥5 RCTs with a total of ≥25 outcome events among the control groups. Second, identified MAs to which component trials had	Pregnancy and childbirth	3 (subset)	MAs with most homogeneous group of interventions	Logistic regression model	1	1	Random sequence generation, allocation concealment, double blinding, missing outcome data	NR	Duplicate trials were only included in the MA with the most homogeneous grouping of interventions	NR	NR	NR

					contributed only the MA with most homogeneous grouping of interventions for inclusion												
Siersma 2007	41/48	523	NR	CDSR (issue 2, 2001)	MAs with ≥5 RCTs	Any	2	MAs with the most clinically important binary outcome	Weighted linear regression model with multivariable multilevel analysis	1	1	ITT	NR	NR	NR	17.02 (no treatment), 42.83 (placebo), 40.15 (active)	70.75 (pharmacological), 19.12 (surgery/procedural), 10.13 (behavioral)
Stadelmaier 2022	27/77	927	NR	CDSR (Jan 2010-Dec 2019)	SRs of nutritional interventions on patient-relevant outcomes	Nutrition	1	MAs with largest number of trials	Meta-meta-analytic approach	3	1	Random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessors, missing outcome data, selective reporting, compliance	NR	NR	Mixed	NR	Micronutrients, fatty acids, dietary approach, food groups, fibre, and food
Unverzagt 2013	6/12	82	24657	CDSR (issue 11, 2011)	SRs on critically ill patients with indications of cardiogenic shock, sepsis, severe sepsis and septic shock	Critical care	1	NR	Logistic regression model with multivariable multilevel analysis	1	1	Random sequence generation, allocation concealment, double blinding, missing outcome data, selective reporting, baseline imbalance	NR	NR	100 (all-cause mortality)	66.67 (inactive), 33.33 (active)	91.67 (pharmacological), 8.33 (surgical)
Wang 2021	86/86	1098	NR	10 general medical journals (Jan 2007-Jun 2019)	MAs published between 1 Jan 2007 and 10 Jun 2019; evaluated chronic	Chronic medical conditions	1	MAs with the most clinically important outcome with largest number of	Mixed-effects random intercept linear regression model	1	2	Random sequence generation, allocation concealment, double blinding,	NR	Conducted sensitivity analyses by excluding MAs with overlapped trials	99.18 (objective), 0.82 (subjective)	29.78 (placebo), 33.33 (standard care), 8.38 (no treatment),	77.05 (pharmacological), 10.38 (device), 12.57 (medical or surgical

					medical conditions; compared different medications, procedures or devices; included only RCTs with ≥5 RCTs.			trials				blinding of outcome assessors, missing outcome data				28.51 (active)	procedure)
Zeraatkar 2022	19 comparisons	436	NR	A living SR and NMA of therapeutics for COVID-19	Comparisons of unique treatment vs. placebo/standard care	COVID-19 therapeutics	1	Analyzed for different outcomes separately	Meta-meta-analytic approach	3	1	Allocation concealment, double blinding	Control: placebo or standard care	Small overlap (three or four-arm trials)	100 (mortality)	100 (placebo or standard care)	100 (pharmacological)

Table 2. Characteristics of meta-epidemiological studies based on within-trial comparisons

Study ID	Number of included trials	Number of patients	Source of trials	Eligibility criteria of trials	Topic area of trials	Choice of outcome included in analysis	Statistical approach	Type of outcomes	Focusing on	Characteristics assessed	Approach to distinguishing between intervention and control groups	Type of outcomes (%)	Type of comparator (%)	Type of interventions (%)
Dello Russo 2020	20	11445	clinicatrials.gov (inception-Jul 2019)	Phase III oncology open-label RCTs reporting the results of both independently assessed and investigator-assessed progression-free survival	Oncology	Progression-free survival	Random effects meta-analysis	Time-to-event outcomes	Undesirable outcomes	Blinding of outcome assessors	NR	Progression-free survival	NR	100 (pharmacological)
Hróbjartsson 2012	21	4391	PubMed, Embase, PsycINFO, CINAHL, Cochrane Central Register of Controlled Trials, HighWire Press, Google Scholar (inception-Jan 2010)	RCTs with blinded and non-blinded assessment of same binary outcome	Any	Most clinically important primary outcome with the first time point after treatment	Random effects meta-analysis	Binary outcomes	Undesirable outcomes	Blinding of outcome assessors	Excluded if unclear	0 (objective), 23.81 (moderately subjective), 76.19 (clearly subjective)	14.29 (placebo/no treatment), 85.71 (standard care)	66.67 (surgical or procedural), 23.81 (pharmacological)
Hróbjartsson 2013	16	2854	PubMed, EMBASE, PsycINFO, CINAHL, Cochrane Central Register of Controlled Trials, HighWire Press, Google Scholar (inception-Jan 2010)	RCTs with blinded and non-blinded assessment of same measurement scale outcome	Any	Primary outcome with the first time point after treatment	Random effects meta-analysis	Measurement scale outcomes	NR	Blinding of outcome assessors	Excluded if unclear	0 (objective), 18.75 (moderately subjective), 81.25 (clearly subjective)	25 (placebo/no treatment), 75 (standard care)	68.75 (surgical or procedural), 31.25 (pharmacological)
Hróbjartsson 2014a	11	1969	PubMed, EMBASE, PsycINFO, CINAHL, Cochrane Central Register of Controlled Trials, HighWire Press, Google Scholar (inception-Sep	RCTs with both blinded and non-blinded assessors of same time-to-event outcome	Any	NR	Random effects meta-analysis	Time-to-event outcomes	Undesirable outcomes	Blinding of outcome assessors	Excluded if unclear	100 (clearly subjective)	16.67 (placebo/no treatment), 83.33 (usual care/active control)	22.22 (surgical or procedural), 77.78 (pharmacological)

			2013)											
Hróbjartsson 2014b	12	3869	Medline, EMBASE, Cochrane Methodology Register, and a systematic review	Parallel group four-armed trials that randomized patients to a blinded sub-study (experimental vs control) and an identical non-blind sub-study (experimental vs control); three-armed trials with experimental and no-treatment groups and a placebo group portrayed to patients as another experimental group	Any	Hierarchy: (i) primary outcome; (ii) for symptom-specific outcomes, visual analogue scales; (iii) for global outcomes, global assessments before quality of life measures	Random effects meta-analysis	Measurement scale outcomes	NR	Blinding of participants	NR	100 (patient-reported measurement scale outcomes), 2 trials (16.67%) also assessed observer-reported measurement scale outcomes	100 (placebo/no treatment)	91.67 (non-pharmacological), 8.33 (pharmacological)
Ndounga Diakou 2016	2	11358	CENTRAL, PubMed, EMBASE, PsycINFO, CINAHL, Google Scholar (inception-Aug 2015)	RCTs reporting same subjective outcome event assessed by both an onsite assessor and an adjudication committee	Any	Most clinically important primary efficacy outcome with most events with the first time point after treatment	Random effects meta-analysis	Binary outcomes	Undesirable outcomes	Blinding of outcome assessors	NR	100 (subjective)	NR	NR
Mostazir 2021	156	NR	Five high impact general medical journals (four-armed trials: inception-Mar 2013; three-armed trials: 2008-2012)	RCTs reporting both an estimate of treatment effect based on intention-to-treat or and an estimate based on per-protocol analysis	Any	Primary outcome; first outcome reported in the abstract	Random effects meta-analysis	Both binary and continuous outcomes - combined together	NR	Intention-to-treat analysis	NR	NR	22.44 (placebo), 77.56 (active control /usual treatment)	57.05 (pharmacological), 42.95 (surgical/medical device/others)

Table 3. Definition of Type of Outcomes

Study ID	Definition
Abraha 2015	NR
Amer 2021	Objective outcome: duration of postoperative hospital stay; Subjective outcome: postoperative pain measured by ordinal, visual analogue or composite scales
Armijo-Olivo 2015	The definition of objective and subjective outcomes was based on the extent to which outcome assessment could be influenced by investigators' judgment. Objectively assessed outcomes included all-cause mortality, measures based on a recognized laboratory procedure, other objective measures, and surgical or instrumental outcomes. Subjectively assessed outcome measures included patient reported outcomes, physician assessed disease outcomes, measures combined from several outcomes, and withdrawals or study dropouts. When different methods of outcome assessment were used in different trials in the same meta-analysis we classified the review according to the most subjective method.
Armijo-Olivo 2017	The definition of objective and subjective outcomes was based on the extent to which outcome assessment could be influenced by investigators' judgment. Objectively assessed outcomes included all-cause mortality, measures based on a recognized laboratory procedure, other objective measures, and surgical or instrumental outcomes. Subjectively assessed outcome measures included patient reported outcomes, physician assessed disease outcomes, measures combined from several outcomes, and withdrawals or study dropouts. When different methods of outcome assessment were used in different trials in the same meta-analysis we classified the review according to the most subjective method.
Armijo-Olivo 2022	The definition of objective and subjective outcomes was based on the extent to which outcome assessment could be influenced by investigators' judgment. Objectively assessed outcomes included all-cause mortality, measures based on a recognized laboratory procedure, other objective measures, and surgical or instrumental outcomes. Subjectively assessed outcome measures included patient reported outcomes, physician assessed disease outcomes, measures combined from several outcomes, and withdrawals or study dropouts. When different methods of outcome assessment were used in different trials in the same meta-analysis we classified the review according to the most subjective method.
Balk 2002	NA
Bassler 2010 (STOPIT-2)	NA
Bialy 2014	NA
Bolvig 2018	Subjective outcome: patient reported pain outcome
Chaimani 2013	NA
Dechartres 2014	We considered objective outcomes as all-cause mortality, other objectively assessed outcomes (ie, pregnancy, live births, laboratory outcomes), or outcomes objectively measured but potentially influenced by clinician or patient judgment (e.g., hospitalizations, total dropouts or withdrawals, cesarean delivery, assisted delivery, additional treatments administered). We considered subjective outcomes as all other outcomes (i.e., patient-reported outcomes, clinician-assessed outcomes, cause-specific mortality).
Haring 2020	Objectively measured versus patient-reported subjective outcomes
Hartling 2014	The definition of objective and subjective outcomes was based on the extent to which outcome assessment could be influenced by investigators' judgment. Objectively assessed outcomes included all-cause mortality, measures based on a recognized laboratory procedure, other objective measures, and surgical or instrumental outcomes. Subjectively assessed outcome measures included patient reported outcomes, physician assessed disease outcomes, measures combined from several outcomes, and withdrawals or study dropouts. When different methods of outcome assessment were used in different trials in the same meta-analysis we classified the review according to the most subjective method.
Hempel 2012 (dataset 4)	NA
Herbison 2011	NA
Hopewell 2004	NA
Khan 1996	Objective outcome: pregnancy
Martin 2021	Objective outcome: mortality
Moher 1998, 1999	NA
Moustgaard 2020 (MetaBLIND)	Outcome measures were classified as observer reported, patient reported (via interviewer or directly recorded by patients), healthcare provider decision outcomes, or mixed (in instances where the outcome was a mixture of more than one category—e.g., both patient and observer reported elements). Observer reported outcomes were subdivided into four outcomes: objective—all cause mortality, objective—other than total mortality (e.g., automatized non-repeatable laboratory tests), subjective—pure observation (e.g., assessment of radiographs), and subjective—interactive (e.g., assessment of clinical status). Subjective observer reported outcomes were scored 1-3 according to the degree of subjectivity (that is, the extent to which determination of the outcome depended on the judgment of the observer, with 1 indicating a low degree of subjectivity).
Niederer 2022	Subjective outcome: chronic nonspecific low back pain
Nuesch 2009a	Subjective outcome: patient reported pain
Nuesch 2009b	Subjective outcome: patient reported pain
Papageorgiou 2015	Objective outcome: NR; Subjective outcomes: self-reported pain intensity and eating or speaking difficulty.
Saltaji 2018a	The definition of objective and subjective outcomes was based on the extent to which outcome assessment could be influenced by investigators' judgment. Objectively assessed outcomes included all-cause mortality, measures based on a recognized laboratory procedure, other objective measures, and surgical or instrumental outcomes. Subjectively assessed outcome measures included patient reported outcomes, physician assessed disease outcomes, measures combined from several outcomes, and withdrawals or study dropouts. When different methods of outcome assessment were used in different trials in the same meta-analysis we classified the review according to the most subjective method.
Saltaji 2018b	The definition of objective and subjective outcomes was based on the extent to which outcome assessment could be influenced by investigators' judgment. Objectively assessed outcomes included all-cause mortality, measures based on a recognized laboratory procedure, other objective measures, and surgical or instrumental outcomes. Subjectively assessed outcome measures included patient reported outcomes, physician assessed disease outcomes, measures combined from several outcomes, and withdrawals or study dropouts. When different methods of outcome assessment were used in different trials in the same meta-analysis we classified the review according to the most subjective method.
Savovic 2012 (BRANDO)	Objective outcomes: all-cause mortality, other objectively assessed outcomes; Subjective outcomes: NR.
Savovic 2018 (ROBES)	Objective outcomes: all-cause mortality, other objectively assessed outcomes; Subjective outcomes: NR.
Schulz 1995	NA
Siersma 2007	NA
Stadelmaier 2022	Objective outcomes: all-cause mortality, other objectively assessed outcomes (pregnancy outcomes); Subjective outcomes: NR.
Unverzagt 2013	Objective outcome: all-cause mortality
Wang 2021	NA
Zeraatkar 2022	Objective outcome: mortality

Dello Russo 2020	Objective outcomes: progression-free survival
Hróbjartsson 2012	Degree of outcome subjectivity (that is, the degree of assessor judgment, high in assessment of global improvement and low in reading a laboratory sheet) on a 1 to 5 scale. Objective outcomes: scores 1; Clearly subjective outcomes: scores 4-5; Moderately subjective outcomes: scores 2-3.
Hróbjartsson 2013	Degree of outcome subjectivity (that is, the degree of assessor judgment, high in assessment of global improvement and low in reading a laboratory sheet) on a 1 to 5 scale. Objective outcomes: scores 1; Clearly subjective outcomes: scores 4-5; Moderately subjective outcomes: scores 2-3.
Hróbjartsson 2014a	Degree of outcome subjectivity (that is, the degree of assessor judgment, high in assessment of global improvement and low in reading a laboratory sheet) on a 1 to 5 scale. Objective outcomes: scores 1; Clearly subjective outcomes: scores 4-5; Moderately subjective outcomes: scores 2-3.
Hróbjartsson 2014b	Blinded observer-reported measurement scale outcomes versus patient-reported measurement scale outcomes
Ndounga Diakou 2016	An outcome was considered "subjective" if it was based on an observer exercising judgment while assessing an event or state and could consequently be influenced by the assessor's knowledge of the allocated treatment.
Mostazir 2021	NA

Appendix 7 Definition of potential risk of bias elements

Table 1 Random sequence generation

Study ID	Comparisons	Definition
Amer 2021	High versus low RoB	Cochrane risk of bias tool 1.0
Armijo-Olivo 2015	High/unclear versus low RoB Unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Category 2 included trials that had adequate randomization (e.g., use of a computer software, random number table and minimization); Category 3 included trials using acceptable methods of randomization, but less efficient than the previous category (e.g., drawing lots, envelopes, shuffling cards, throwing a dice). High RoB: Category 4 involved trials using inappropriate methods of sequence generation (e.g., date of birth, day of admission, hospital record number). Unclear RoB: Category 1 included trials where random sequence generation was unclear or not reported.
Bialy 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Chaimani 2013	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Methods that suggested an adequate sequence generation where; use of a random number table, a computer random number generator, coin tossing, throwing dice, restricted randomization methods such as random permuted blocks, minimization technique or similar. Unclear RoB: When we had insufficient information about the random sequence generation to permit judgement of 'Low risk' or 'High risk', then we judged as unclear.
Dechartres 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Random number table, computer random number generator, coin tossing, shuffling cards or envelopes, minimization. High RoB: Sequence generated by odd or date of birth or date of admission. Unclear RoB: Method used to generate sequence of randomization not reported.
Haring 2020	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Hartling 2014	High versus low RoB High/unclear versus low RoB High versus low/unclear RoB	Cochrane risk of bias tool 1.0
Hopewell 2004	High/unclear versus low RoB	Low RoB: Examples of adequate methods of generation of the allocation sequence included computer randomization, random number tables and the drawing of lots or envelopes. High RoB: Inadequate methods of generation of the allocation sequence included allocation by date of birth, clinic group, or by day of the week.
Khan 1996	High/unclear versus low RoB	Low RoB: Methods of sequence generation that were considered adequate included random numbers generated by computer, random number tables, coin tossing, or card shuffling. High RoB: Trials in which allocation was performed using case record numbers, social insurance numbers, or birth dates were considered to have inadequate randomization sequence generation. Unclear RoB: Trials in which the authors did not report their method of sequence generation.

Moher 1998, 1999	High versus low RoB	Jadad scale Low RoB: Clinical trials that reported the following methods for generation of their allocation sequence were considered adequate: computer, random number table, shuffled cards or tossed coins, and minimization. High RoB: Inadequate methods included alternate assignment and assignment by odd/even birth date or hospital number.
Niederer 2022	High versus low RoB	NR
Papageorgiou 2015	High versus low RoB High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Adequate random sequence generation method clearly described and adequate according to Cochrane Collaboration criteria. High RoB: Inadequate random sequence generation. Unclear RoB: Unclear random sequence generation.
Saltaji 2018a	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Adequate random sequence generation. High RoB: Inadequate random sequence generation. Unclear RoB: Unclear random sequence generation.
Savovic 2012	High versus low RoB High/unclear versus low RoB Unclear versus low RoB	As assessed in the original meta-epidemiological studies. Low RoB: Adequate random sequence generation. High RoB: Inadequate random sequence generation. Unclear RoB: Unclear random sequence generation.
Savovic 2018	High/unclear versus low RoB High versus low/unclear RoB	NR
Schulz 1995	High versus low RoB (in adequately concealed trials)	Low RoB: Adequate sequence generation (reported using random-number table, computer random-number generator, coin tossing, or shuffling). High RoB: Did not report one of the adequate approaches, those with inadequate sequence generation.
Stadelmaier 2022	High/unclear versus low RoB	NR
Unverzagt 2013	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Wang 2021	High versus low RoB High/unclear versus low RoB	NR

Table 2 Allocation concealment

Study ID	Comparisons	Definition
Amer 2021	High versus low RoB	Cochrane risk of bias tool 1.0

Armijo-Olivo 2015	High versus low RoB High/unclear versus low RoB Unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Category 1 comprised trials that used any type of central randomization (e.g., a remote telephone service or a central office); Category 2 comprised trials that used sequentially numbered, opaque and sealed envelopes. High RoB: Category 4 for comparison high versus low RoB; Category 3 and category 4 for comparison high/unclear versus low RoB. (Category 3 comprised trials that used sealed envelopes without reporting any further details; Category 4 comprised trials where allocation was clearly not hidden (e.g., being based on an open list, odd or even days of the week, participant's birth date or the team on duty at enrolment) Unclear RoB: Category 5 comprised trials were concealment of allocation was not reported or unclear.
Bialy 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Chaimani 2013	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Participants and investigators enrolling participants could not foresee assignment because one of the following, or an equivalent method, was used to conceal allocation; central allocation, sequentially numbered drug containers of identical appearance or sequentially numbered, opaque, sealed envelopes were used. Unclear RoB: Insufficient information about the allocation concealment to permit judgement of 'Low risk' or 'High risk'.
Dechartres 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Central allocation, sequentially numbered drug containers of identical appearance, sequentially numbered opaque sealed envelopes. High RoB: Predictable assignment (date of birth, alternation, open random allocation schedule, unsealed envelopes). Unclear RoB: Method to maintain allocation concealment not reported.
Haring 2020	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Hartling 2014	High versus low RoB High/unclear versus low RoB High versus low/unclear RoB	Cochrane risk of bias tool 1.0
Herbison 2011	High versus low RoB	Low RoB: Trials that used some form of central randomization that clearly should hide the allocation, such as a remote telephone service or randomization by a pharmacy. High RoB: Trials where the allocation was clearly not hidden, for example, being based on an open list, odd or even days of the week, participant's birth date, or the team on duty at enrollment.
	High/unclear versus low RoB	Low RoB: Category 1 comprised trials that used some form of central randomization that clearly should hide the allocation, such as a remote telephone service or randomization by a pharmacy; Category 2 comprised trials that used sealed envelopes with some form of security enhancement, such as ensuring that envelopes were opaque and numbered. High/unclear RoB: Category 3 comprised trials that used sealed envelopes without any further details; Category 4 comprised trials that were reported as randomized without details, and also as "double blind"; Category 5 comprised trials that simply said they were randomized with no further details; Category 6 comprised trials where the allocation was clearly not hidden, for example, being based on an open list, odd or even days of the week, participant's birth date, or the team on duty at enrollment.
Khan 1996	High/unclear versus low RoB	Low RoB: Concealment of allocation was considered adequate if randomization was performed at a site remote from the treatment site, or if coded bottles,

		serially numbered drug containers, opaque sealed envelopes, or other such methods were used. High RoB: Concealment was considered inadequate if randomization was based on open methods such as reference to case record numbers or birth dates. Unclear RoB: In the absence of any information about concealment, the trial was categorized as being concealed unclearly.
Moher 1998, 1999	High versus low RoB	Jadad scale Low RoB: Trials in which concealment up to the point of treatment (e.g. central randomization) was reported. High RoB: Trials reporting allocation concealment inadequately.
Moustgaard 2020	High/unclear versus low RoB	Low RoB: Adequate allocation concealment. High RoB: Inadequate allocation concealment. Unclear RoB: Unclear allocation concealment.
Niederer 2022	High versus low RoB	NR
Nuesch 2009b	High/unclear versus low RoB	Low RoB: Investigators responsible for patient selection and inclusion were unable to know before allocation which treatment was next, e.g., central randomization; the use of sequentially numbered, sealed, and opaque assignment envelopes; or coded drug packs. Unclear RoB: Concealment of allocation of trials, which lacked a specific statement, was classified as unclear.
Saltaji 2018a	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Adequate allocation concealment. High RoB: Inadequate allocation concealment. Unclear RoB: Unclear allocation concealment.
Savovic 2012	High versus low RoB High/unclear versus low RoB Unclear versus low RoB High/unclear versus low RoB (in double-blinded trials)	As assessed in the original meta-epidemiological studies. Low RoB: Adequate allocation concealment. High RoB: Inadequate allocation concealment. Unclear RoB: Unclear allocation concealment.
Savovic 2018	High/unclear versus low RoB High versus low/unclear RoB High/unclear versus low RoB (in adequately generated random sequence and blinded trials)	NR
Schulz 1995	High versus low RoB Unclear versus low RoB	Low RoB: Adequately concealed trials, the referent group, that were deemed to have taken adequate measures to conceal allocation (ie, central randomization; numbered or coded bottles or containers; drugs prepared by the pharmacy; serially numbered, opaque, sealed envelopes; or other description that contained elements convincing of concealment). High RoB: Inadequately concealed trials, in which concealment was inadequate (such as alternation or reference to case record numbers or to dates of

		birth. Unclear RoB: Unclearly concealed trials, in which the authors either did not report an allocation concealment approach at all or reported an approach that did not fall into one of the categories just named.
Stadelmaier 2022	High/unclear versus low RoB	NR
Unverzagt 2013	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Wang 2021	High versus low RoB High/unclear versus low RoB	NR
Zeraatkar 2022	Definitely/probably unconcealed versus definitely/probably concealed High versus low RoB (Definitely unconcealed versus definitely concealed)	Definitely concealed: If the trial report described central randomization either via a computer or telephone system, pharmacy-controlled randomization, or sequentially-numbered opaque sealed envelopes. Probably concealed: We assumed that trials that omitted any description of allocation concealment and blinding did not implement allocation concealment and blinding with one exception. We assumed that trials in which healthcare providers were blinded implemented allocation concealment because it is unlikely that healthcare providers could be adequately blinded without allocation concealment.

Table 3 Baseline imbalance

Study ID	Comparisons	Definition
Balk 2002	Imbalance/unclear versus balance	Balance: Treatment and control groups were similar in the characteristics reported. Imbalance: Not similar. Unclear: Unreported.
Hartling 2014	Imbalance versus balance Imbalance/unclear versus balance Imbalance versus balance/unclear	Cochrane risk of bias tool 1.0
Saltaji 2018a	Imbalance/unclear versus balance	Cochrane risk of bias tool 1.0 Balance: Authors state that the groups were comparable or had an equal prognostic factor baseline. They analyzed this by comparing groups through a statistical test in all variables of interest. Or the authors state that groups are not comparable and they adjusted statistically. Groups must be comparable with regard to (for example) pain, global perceived effect, participation in daily activities; at least one of the main outcomes must be described, age; sex; and pre-existing participation problems.

		Imbalance: Authors state that groups are not equal at baseline and they did not adjust for any difference. Unclear: Unknown.
Unverzagt 2013	Imbalance/unclear versus balance	Cochrane risk of bias tool 1.0

Table 4 Blinding of healthcare providers

Study ID	Comparisons	Definition
Amer 2021	High versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Healthcare staff were unaware of the patient's assigned treatment.
Balk 2002	High/unclear versus low RoB	Caregivers included physicians, nurses, and other health care practitioners in direct patient care or parents (or equivalent) of outpatient infants.
Bialy 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Moustgaard 2020	Definitely/probably unblinded versus definitely/probably blinded Definitely/probably unblinded or unclear versus definitely/probably blinded	A modified algorithm of Akl 2011 (Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. J Clin Epidemiol 2012;65:262-7. doi:10.1016/j.jclinepi.2011.04.015)
Saltaji 2018b	High/unclear versus low RoB	Low RoB: The study describes in the title, abstract, or text that the therapists/care-providers were blinded. The blinding was appropriate. High RoB: The study describes in the title, abstract, or text that the therapists/care-providers were not blinded, or because of the nature of the intervention (e.g., exercise prescription or supervision, etc.), the therapist could not be blinded. Unclear RoB: Unclear or not reported.

Table 5 Blinding of patients

Study ID	Between- or within-trial comparisons	Comparisons	Definition
Amer 2021	Between-trial	High versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Patients were unaware of the assigned treatment.
Armijo-Olivo 2017	Between-trial	High/unclear versus low RoB	A 3-point scale (yes, no, unclear) Yes: No blinding or incomplete blinding, but the review authors judge that the outcome is not likely to be influenced by lack of blinding (automated outcome or administrative); Blinding of participants and key study personnel ensured, and unlikely that the blinding could have

			been broken; Objectives automatized outcomes coming from databases or hospital register office. No: No blinding or incomplete blinding, and the outcome is likely to be influenced by lack of blinding; Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken, and the outcome is likely to be influenced by lack of blinding. Unclear: Insufficient information to permit judgement of low risk or high risk; or the study did not address the issue of blinding.
Balk 2002	Between-trial	High/unclear versus low RoB	Were patients reported to have been blinded? Low RoB: Yes. If not stated explicitly, infants and patients receiving identical-appearing treatments (active or placebo) were considered to have been blinded. Unclear RoB: Not reported.
Bialy 2014	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Bolvig 2018	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Chaimani 2013	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: When the authors described the study as double dummy and used identical containers, identical pills etc., we judged blinding of participants as adequate. Unclear RoB: When the authors stated that the study was double-blind but there was no adequate description in the text, we classified the study as unclear.
Hróbjartsson 2014b	Within-trial	High versus low RoB (in adequately concealed trials)	Low RoB: Patients were regarded as blinded when this was explicitly reported or when blinding was indicated by use of a placebo treatment (and if there was no indication of unblinding of patients). High RoB: Non-blinded patients.
Moustgaard 2020	Between-trial	Definitely/probably unblinded versus definitely/probably blinded Definitely/probably unblinded or unclear versus definitely/probably blinded	A modified algorithm of Akl 2011 (Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. J Clin Epidemiol 2012;65:262-7. doi:10.1016/j.jclinepi.2011.04.015)
Nuesch 2009b	Between-trial	High/unclear versus low RoB	Low RoB: Patient blinding was considered adequate if a placebo or sham control intervention was used and experimental and control interventions were described as indistinguishable or the use of a double dummy technique was reported.
Saltaji 2018b	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: No blinding or incomplete blinding, but the review authors judge that the outcome is not likely to be influenced by lack of blinding (automated outcome or administrative); Blinding of participants and key study personnel ensured, and unlikely that the blinding could have

			been broken; Objectives automatized outcomes coming from databases or hospital register office. High RoB: No blinding or incomplete blinding, and the outcome is likely to be influenced by lack of blinding; Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken, and the outcome is likely to be influenced by lack of blinding. Unclear RoB: Insufficient information to permit judgement of low risk or high risk; or the study did not address the issue of blinding.

Table 6 Blinding of data collectors

Study ID	Comparisons	Definition
Amer 2021	High versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Data collectors were unaware of the patient's assigned treatment.

Table 7 Blinding of outcome assessors

Study ID	Between- or within-trial comparisons	Comparisons	Definition
Amer 2021	Between-trial	High versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: Outcome assessors were unaware of the patient's assigned treatment.
Armijo-Olivo 2017	Between-trial	High/unclear versus low RoB	A 3-point scale (yes, no, unclear) Yes: No blinding of outcome assessment, but the review authors judge that the outcome measurement is not likely to be influenced by lack of blinding; Blinding of outcome assessment ensured, and unlikely that the blinding could have been broken. No: No blinding of outcome assessment, and the outcome measurement is likely to be influenced by lack of blinding; Blinding of outcome assessment, but likely that the blinding could have been broken and the outcome measurement is likely to be influenced by lack of blinding. Unclear: Insufficient information to permit judgement of low risk or high risk; The study did not address the issue of blinding.
Balk 2002	Between-trial	High/unclear versus low RoB	Outcome assessors included physicians or other health care practitioners or researchers who evaluated either patients, their records, or their laboratory or radiology tests to determine study outcomes.
Bialy 2014	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Chaimani 2013	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: For hard outcomes (e.g. death) we evaluated outcome assessors as low. When outcomes were not hard, outcome assessment was judged according to the details reported. We considered lab outcomes as objective and blinding of outcome assessors was judged as adequate. Unclear RoB: When the authors stated that the study was double-blind but there was no adequate description in the text, we classified the

			study as unclear.
Dello Russo 2020	Within-trial	High versus low RoB	Low RoB: Outcome assessed by blinded independent central reviews. High RoB: Outcome assessed by unblinded investigator.
Diakou 2016	Within-trial	High versus low RoB	Low RoB: Adjudication committee assessed events identified independent of unblinded onsite assessors (85.1% blinded). High RoB: Unblinded onsite assessors.
Haring 2020	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Hartling 2014	Between-trial	High versus low RoB High/unclear versus low RoB High versus low/unclear RoB	Cochrane risk of bias tool 1.0
Hróbjartsson 2012	Within-trial	High versus low RoB	Low RoB: Blinded outcome assessor. High RoB: Non-blinded outcome assessor.
Hróbjartsson 2013	Within-trial	High versus low RoB	Low RoB: Blinded outcome assessor. High RoB: Non-blinded outcome assessor.
Hróbjartsson 2014a	Within-trial	High versus low RoB	Low RoB: Blinded outcome assessor. High RoB: Non-blinded outcome assessor.
Moustgaard 2020	Between-trial	Definitely/probably unblinded versus definitely/probably blinded Definitely/probably unblinded or unclear versus definitely/probably blinded	A modified algorithm of Akl 2011 (Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. J Clin Epidemiol 2012;65:262-7. doi:10.1016/j.jclinepi.2011.04.015)
Saltaji 2018b	Between-trial	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: No blinding of outcome assessment, but the review authors judge that the outcome measurement is not likely to be influenced by lack of blinding; Blinding of outcome assessment ensured, and unlikely that the blinding could have been broken. High RoB: No blinding of outcome assessment, and the outcome measurement is likely to be influenced by lack of blinding; Blinding of outcome assessment, but likely that the blinding could have been broken and the outcome measurement is likely to be influenced by lack of blinding. Unclear RoB: Insufficient information to permit judgement of low risk or high risk; The study did not address the issue of blinding.
Stadelmaier 2022	Between-trial	High/unclear versus low RoB	NR
Wang 2021	Between-trial	High versus low RoB High/unclear versus low RoB	NR

Table 8 Blinding of data analysts

Study ID	Comparisons	Definition
Armijo-Olivo 2017	High/unclear versus low RoB	A 3-point scale (yes, no, unclear) Yes: If the study described in the title, abstract or text that the statistician is blinded, and the blinding is appropriate. No: If the study described in the title, abstract or text that the statistician was not blinded. Unclear: Insufficient information to permit judgment of 'Yes' or 'No'.

Table 9 Double blinding

Study ID	Comparisons	Definition
Amer 2021	High versus low RoB	Cochrane risk of bias tool 1.0 **Blinding of patients, healthcare providers, data collectors, outcome assessors and/or data analysts**
Dechartres 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: No blinding but objective outcome (i.e., mortality, biological tests); **Blinding of participants, personnel and outcome assessor** and unlikely that the blinding could have been broken; Either participants or some key personnel were not blinded but outcome assessment was blinded and the non-blinding of others unlikely to introduce bias. High RoB: No blinding or incomplete blinding and the outcome measurement is likely to be influenced by lack of blinding (i.e., subjective outcome); Blinding of participants and personnel attempted but likely that the blinding could have been broken; Either participants or personnel were not blinded, and the non-blinding likely to introduce bias. Unclear RoB: Insufficient information regarding blinding.
Haring 2020	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 **Blinding of participants and personnel**
Hartling 2014	High versus low RoB High/unclear versus low RoB High versus low/unclear RoB	Cochrane risk of bias tool 1.0 **Blinding of participants or personnel**
Hopewell 2004	High versus low RoB	**"Double blind"**
Khan 1996	High/unclear versus low RoB	Low RoB: Blinding was considered adequate if **both the physician and the patients** were blinded and the placebos used were identical in taste and/or appearance to the active treatment. High RoB: Inadequate if it was clear that the physician or the study participants were able to identify the intervention being provided, e.g., oral versus parenteral administration of treatment.

Martin 2021	High/unclear versus low RoB	Low RoB: **Described as "double blind"** (defined as unknown intervention by both patients and personnel) and/or mentioned the use of adequate methods for double-blinding (e.g., matched placebo). High RoB: Described as a single-blind or open-label trial, had distinguishable interventions, or no information was reported.
Moher 1998, 1999	High versus low RoB	Jadad scale Low RoB: **Described as "double blind".** High RoB: Trials reporting double-blinding inadequately.
Moustgaard 2020	High/unclear versus low RoB	Low RoB: **Described as "double blind" (or "triple blind").** High/unclear RoB: Those not so described or unclear.
Saltaji 2018b	High/unclear versus low RoB	**Described as "double blind"**
		Cochrane risk of bias tool 1.0 **Blinding of both patients and outcome assessors**
		Cochrane risk of bias tool 1.0 **Blinding of patients, assessors, and care-providers**
Savovic 2012	High versus low RoB High/unclear versus low RoB Unclear versus low RoB High/unclear versus low RoB (in adequately concealed trials)	As assessed in the original meta-epidemiological studies. **"Double blind"**
Savovic 2018	High/unclear versus low RoB High versus low/unclear RoB High/unclear versus low RoB (in adequately generated random sequence and concealed trials)	**"Blinding"**
Schulz 1995	High versus low RoB (in adequately or unclear concealed trials)	Low RoB: A referent group of trials that reported having been **double-blinded.** High RoB: A second group that did not report as such, deemed not double-blinded.
Stadelmaier 2022	High/unclear versus low RoB	**Blinding of participants and personnel**
Unverzagt 2013	High/unclear versus low RoB	**"Double blind"**
Wang 2021	High versus low RoB	**Blinding of participants and personnel**

	High/unclear versus low RoB	
Zeraatkar 2022	Definitely/probably unblinded versus definitely/probably blinded High versus low RoB (Definitely unblinded versus definitely blinded)	**Blinding of patients and health care providers** Blinded: We considered a trial blinded if both patients and healthcare providers are described as being unaware of the intervention to which patients were assigned and described adequate blinding methods (i.e., matching placebo). Unblinded: Neither patients and healthcare providers were blinded or only one party was blinded. We considered a trial open-label if patients and healthcare providers were described as being aware of the intervention to which patients were assigned.

Table 10 Imbalance in co-interventions

Study ID	Comparisons	Definition
Hempel 2012	Imbalance/unclear versus balance	Cochrane Back Review Group Criteria Balance: This item should be scored "yes" if there were no co-interventions or they were similar between the index and control groups.

Table 11 Compliance/adherence with intervention

Study ID	Comparisons	Definition
Hempel 2012	Unacceptable/unclear non-compliance versus no/acceptable non-compliance	Cochrane Back Review Group Criteria The reviewer determines if the compliance with the interventions is acceptable, based on the reported intensity, duration, number and frequency of sessions for both the index intervention and control intervention(s). For example, physiotherapy treatment is usually administered over several sessions; therefore it is necessary to assess how many sessions each patient attended. For single-session interventions (for ex: surgery), this item is irrelevant.
Stadelmaier 2022	Unacceptable/unclear non-compliance versus no/acceptable non-compliance	NR

Table 12 Missing outcome data

Study ID	Comparisons	Definition
Armijo-Olivo 2022	Missing ≥ 5% versus < 5%	NA
	Missing ≥ 10% versus < 10%	NA
	Missing ≥ 15% versus < 15%	NA
	Missing ≥ 20% versus < 20%	NA

	Missing ≥ 25% versus < 25%	NA
	Missing ≥ 30% versus < 30%	NA
	Missing ≥ 35% versus < 35%	NA
	Missing ≥ 40% versus < 40%	NA
	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Balk 2002	Per 1% point increase in missing	NA
Bialy 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Dechartres 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 Low RoB: No missing outcome data; Missing data have been imputed using appropriate methods (worst-case analysis); Missing data balanced in numbers across intervention groups with similar reasons for missing data across groups; The proportion of missing outcomes compared with observed event risk not enough to have a clinically relevant impact on the intervention effect estimate (< 10% of the number of patients randomized or < the number of outcomes). High RoB: Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups; The proportion of missing outcomes compared with observed event risk enough to induce clinically relevant bias in intervention effect estimate (≥ 10% of patients randomized or ≥ the number of outcomes); As-treated analysis done with substantial departure of the intervention received from that assigned at randomization (≥ 10% of patients randomized or ≥ the number of outcomes). Unclear RoB: Insufficient reporting of attrition/exclusion (i.e., number of participants randomized and analyzed not stated, no reason for missing data provided).
Haring 2020	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Hartling 2014	High versus low RoB High/unclear versus low RoB High versus low/unclear RoB	Cochrane risk of bias tool 1.0
Khan 1996	Missing ≥ 10% versus < 10%	NA
Niederer 2022	High versus low RoB	NR
Nuesch 2009a	Missing versus no missing	Excluding patients from the analysis. No missing: Trials were classified to have had no exclusions of patients from the analysis if there was an explicit statement that all randomized patients were included in the analysis of the outcome we extracted or if the reported numbers of patients randomized and analyzed on this outcome were identical. Missing: We classified trials to have had exclusions if they explicitly reported exclusions from the analysis, if the number of patients analyzed was lower than the number of patients randomized, or if it was unclear whether exclusions from the analysis had occurred.
Savovic 2012	Missing ≥ 20% versus < 20%	NA
Savovic 2018	High/unclear versus low RoB	NR

	High versus low/unclear RoB	
Schulz 1995	Missing versus no missing	Excluding patients from the analysis. No missing: Trials that reported, or gave the impression, that no exclusions had taken place. Missing: Trials that reported having made exclusions. The reasons for exclusions (when given) included protocol deviations, withdrawals, dropouts, and losses to follow-up.
Stadelmaier 2022	High/unclear versus low RoB	NR
Unverzagt 2013	High/unclear versus low RoB	Cochrane risk of bias tool 1.0
Wang 2021	High versus low RoB High/unclear versus low RoB	NR

Table 13 Selective reporting

Study ID	Comparisons	Definition
Bialy 2014	High/unclear versus low RoB	Cochrane risk of bias tool 1.0 High RoB: If the primary outcome, as stated in each RCT, was not included in the results section, the domain was rated as high RoB.
Hartling 2014	High versus low RoB High/unclear versus low RoB High versus low/unclear RoB	Cochrane risk of bias tool 1.0
Niederer 2022	High versus low RoB	NR
Stadelmaier 2022	High/unclear versus low RoB	NR
Unverzagt 2013	High/unclear versus low RoB	Cochrane risk of bias tool 1.0

Table 14 Intention-to-treat analysis

Study ID	Between- or within-trial comparisons	Comparisons	Definition
Abraha	Between-	Not ITT versus ITT	ITT: Trials reported the phrase "intention to treat" with no apparent deviation in the description or trials that correctly described the intention

2015	trial	Not ITT versus unclear Not ITT versus ITT/unclear	to treat principle (**analyzing patients according to their original allocation**). If a trial did not use the phrase but intended to analyze the patient data according to the original allocation of the patients, then it was classified in this category. Not ITT: Used the term "modified intention to treat" or reported a deviation from the intention to treat approach. The number and type of deviations were retrieved and deviations were classified as treatment related deviation, baseline assessment related deviation, target condition related deviation, and post-baseline assessment related deviation. Unclear: Trials did not refer any intention to treat approach and did not fall into the previous two categories.
Armijo-Olivo 2022	Between-trial	Not ITT/unclear versus ITT	ITT: ITT is a strategy used to analyze the results of an RCT that **considers the subjects in the way they were randomized at the beginning of the trial** regardless of whether they completed the intervention given to their group, their compliance with the entry criteria, the treatment they actually received, or whether they withdrew from treatment or deviated from the experimental protocol. The "original concept and ideal ITT" require a complete set of data, which means that **all data from all randomized subjects should be included in the final analysis regardless of whether they completed the trial according to protocol or not**. Thus, all patients should be followed, and their data should be obtained regardless of any protocol deviation (i.e., compliance, adverse events, or migration between groups).
Balk 2002	Between-trial	Not ITT/unclear versus ITT	ITT: **All patients were analyzed in the group to which they were originally allocated. Dropouts were allowable** so long as the reasons for withdrawal were not related to the group to which they were assigned (bias).
Mostazir 2021	Within-trial	PP versus ITT PP versus mITT PP versus ITT/mITT	ITT: A strategy for analyzing data in which all participants are included in the group to which they were assigned, whether or not they completed the intervention given to the group (when not all patients as randomized were available for analysis, the analysis included only those patients for whom an outcome measure was available is also considered as ITT analysis). mITT: Where patients were excluded if a certain minimum dose of intervention was not received. PP: An analytical strategy restricted to only participants who fulfil the protocol in terms of eligibility, interventions, and outcome assessment.
Siersma 2007	Between-trial	Not ITT/unclear versus ITT	ITT: **All randomized participants were included in the analysis in the group to which they originally were assigned.** Not ITT: Some participants were excluded from the analysis. Unclear: Not described.

Table 15 Stop early for benefit

Study ID	Comparisons	Definition
Bassler 2010	Stop early for benefit versus not	NA

Appendix 8 Forest plots

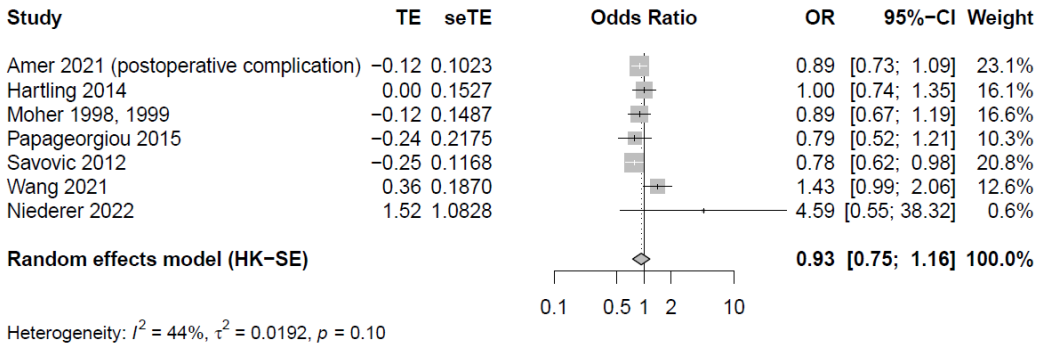

Figure 1 Random sequence generation - High versus low risk of bias

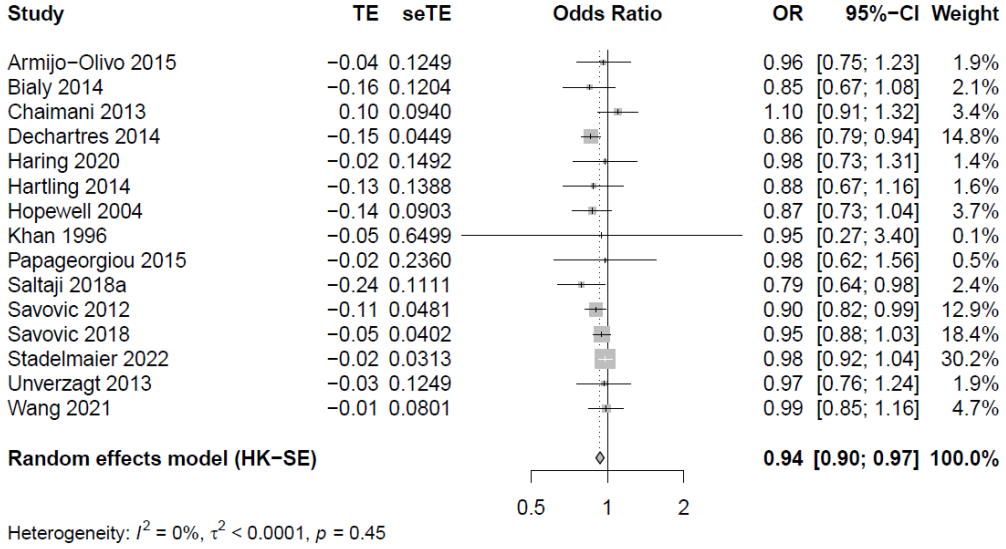

Figure 2 Random sequence generation – High/unclear versus low risk of bias

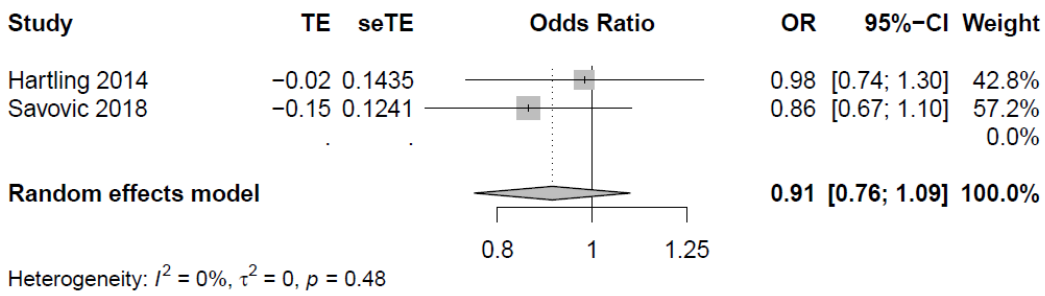

Figure 3 Random sequence generation – High versus low/unclear risk of bias

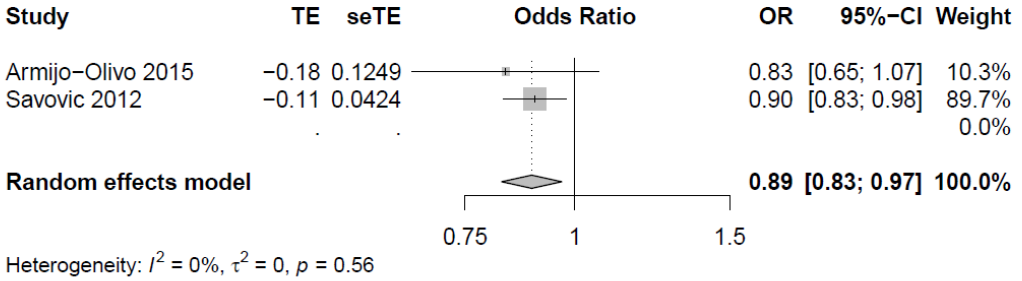

Figure 4 Random sequence generation – Unclear versus low risk of bias

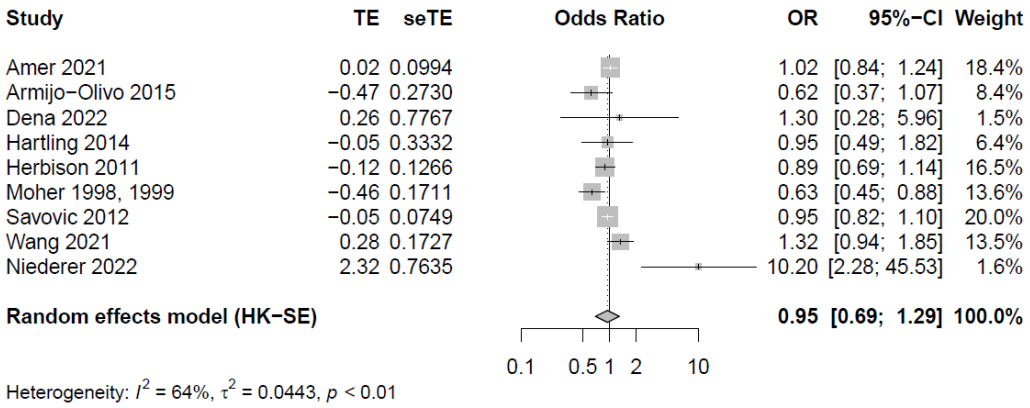

Figure 5 Allocation concealment - High versus low risk of bias

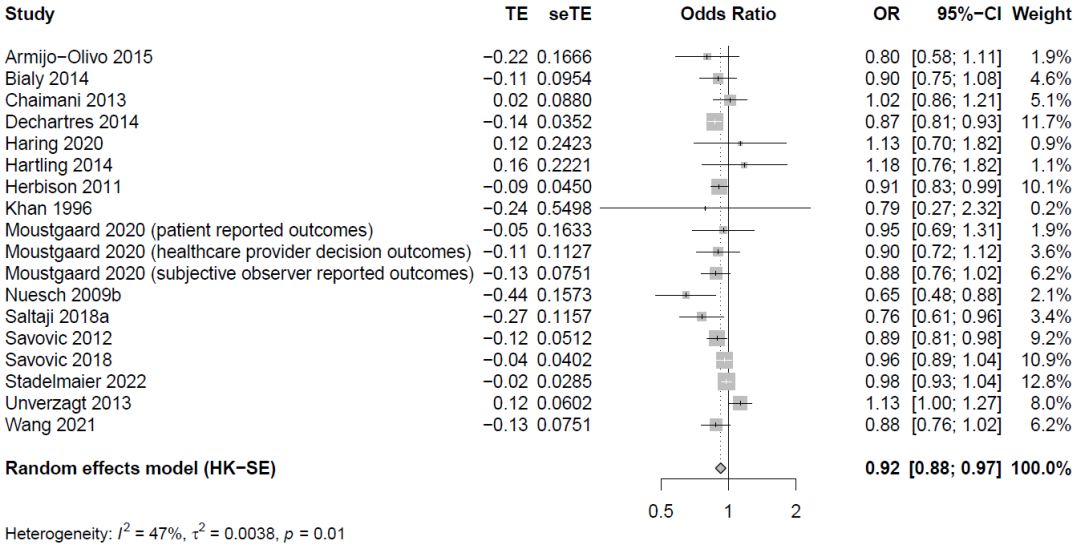

Figure 6 Allocation concealment - High/unclear versus low risk of bias

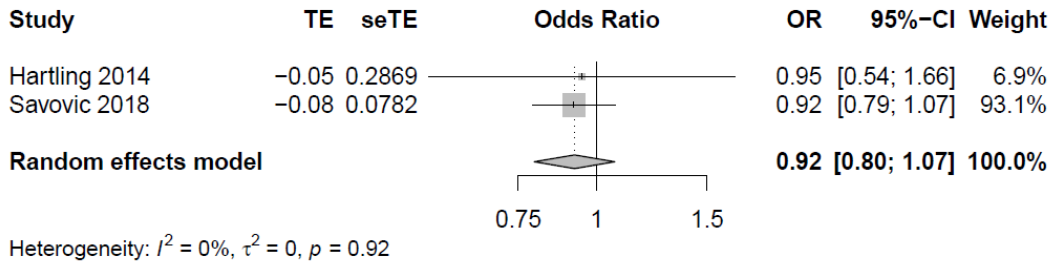

Figure 7 Allocation concealment - High versus low/unclear risk of bias

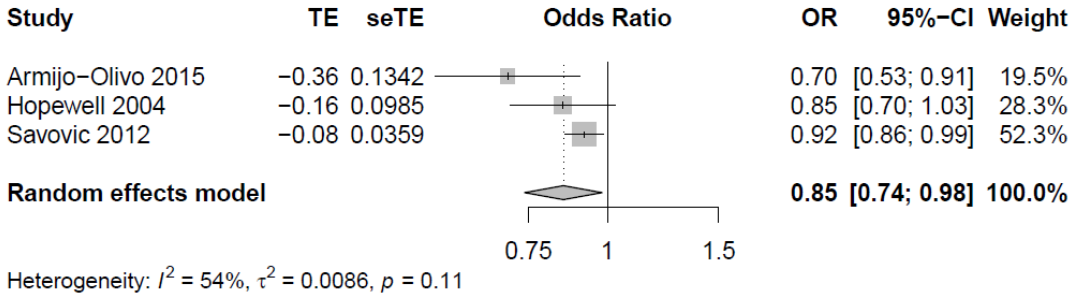

Figure 8 Allocation concealment - Unclear versus low risk of bias

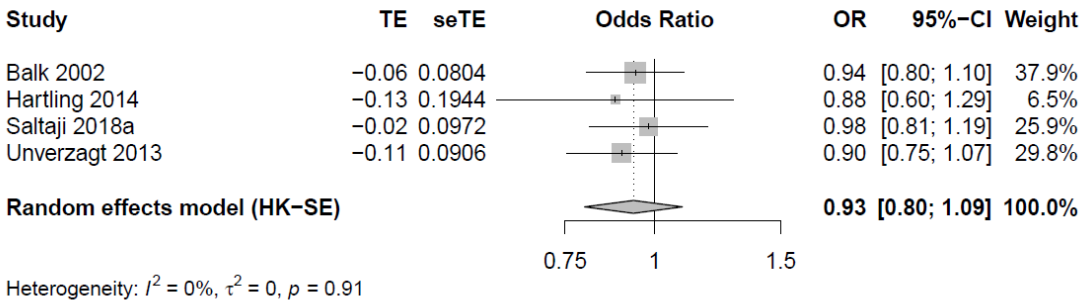

Figure 9 Baseline imbalance - Imbalance/unclear versus balance

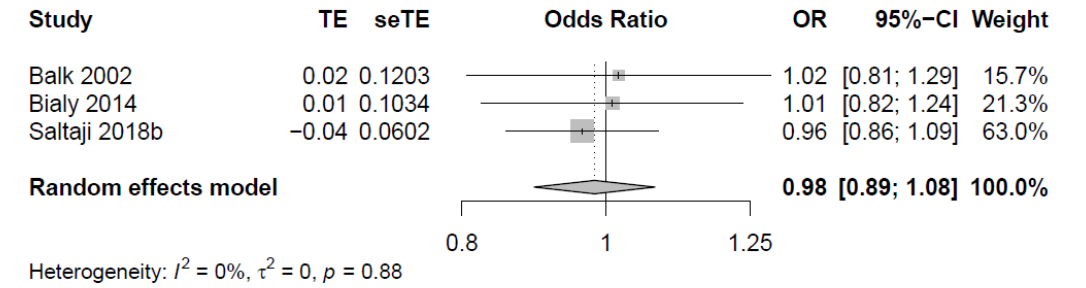

Figure 10 Blinding of healthcare providers - High/unclear versus low risk of bias

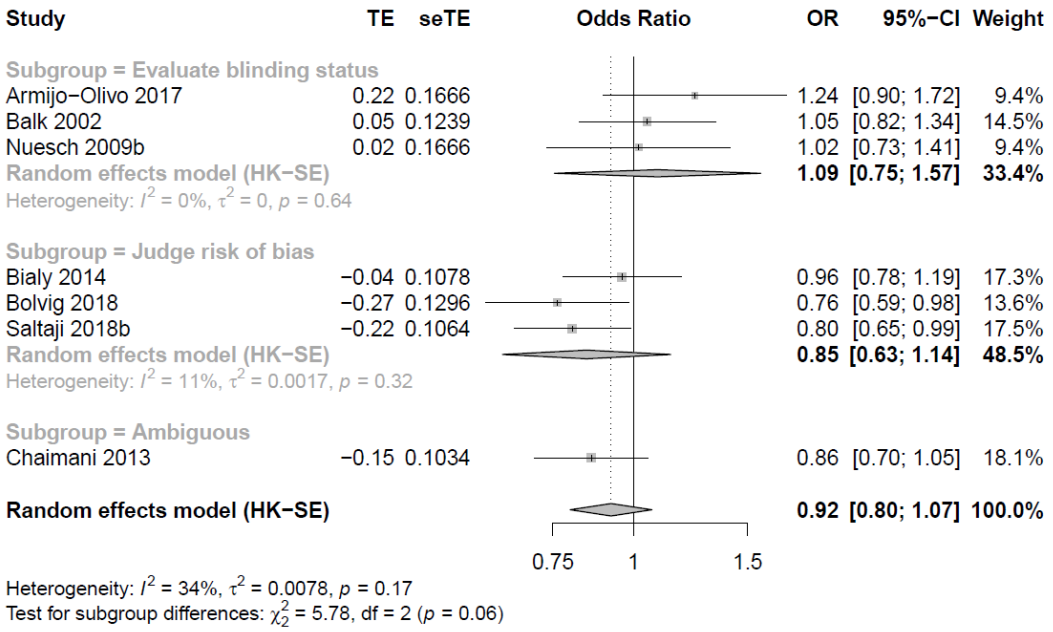

Figure 11 Blinding of patients - High/unclear versus low risk of bias

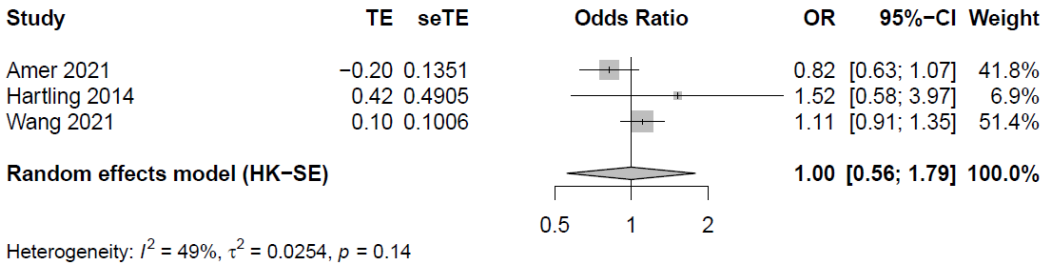

Figure 12 Blinding of outcome assessors - High versus low risk of bias

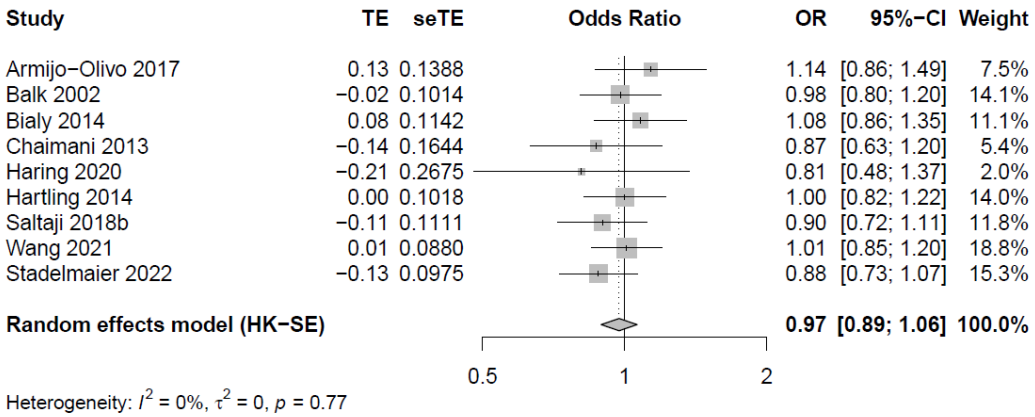

Figure 13 Blinding of outcome assessors - High/unclear versus low risk of bias

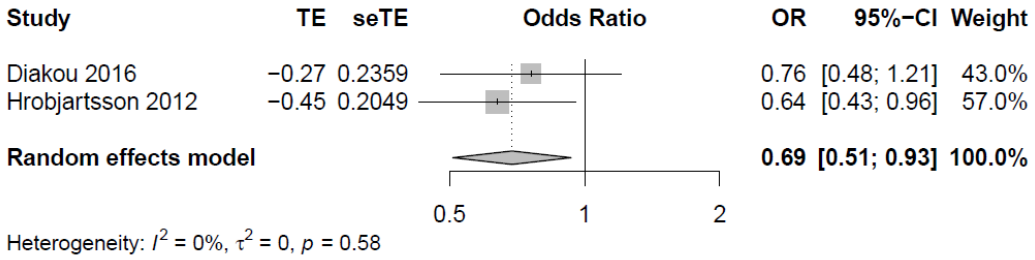

Figure 14 Blinding of outcome assessors - High versus low risk of bias (binary outcomes)

for within-trial comparisons – Subjective outcomes

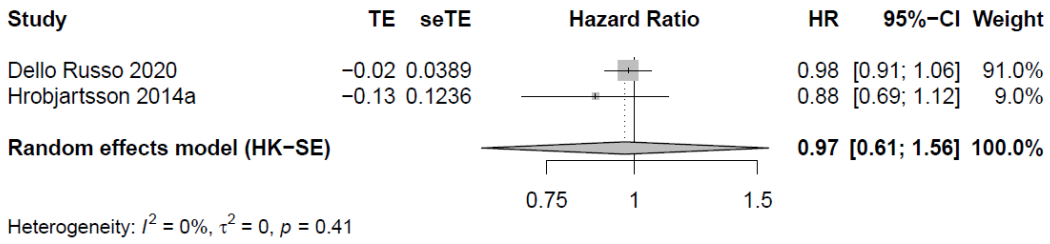

Figure 15 Blinding of outcome assessors - High versus low risk of bias (time-to-event outcomes) for within-trial comparisons

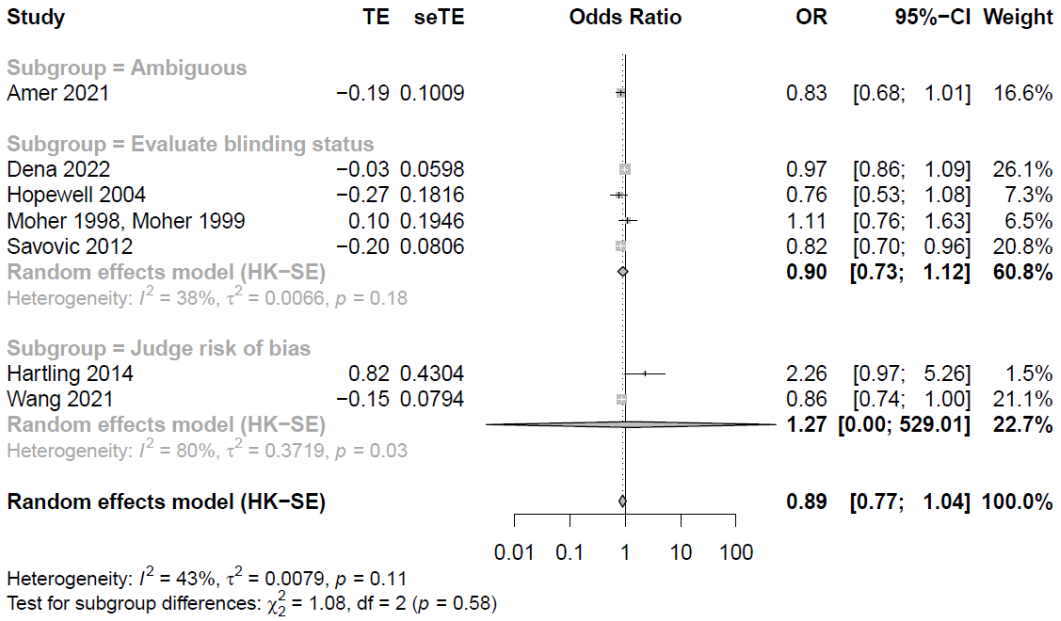

Figure 16 Double blinding – High versus low risk of bias - Any outcomes

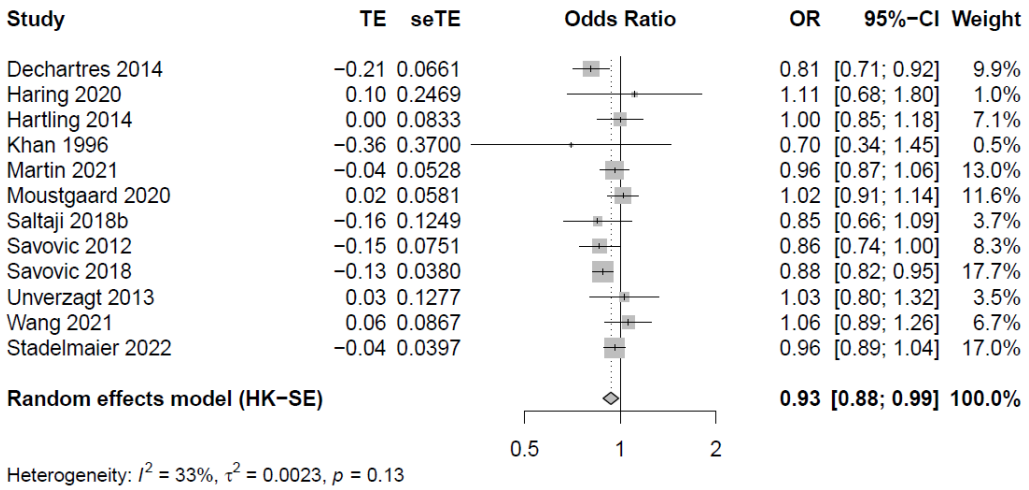

Figure 17 Double blinding - High/unclear versus low risk of bias – Any outcome

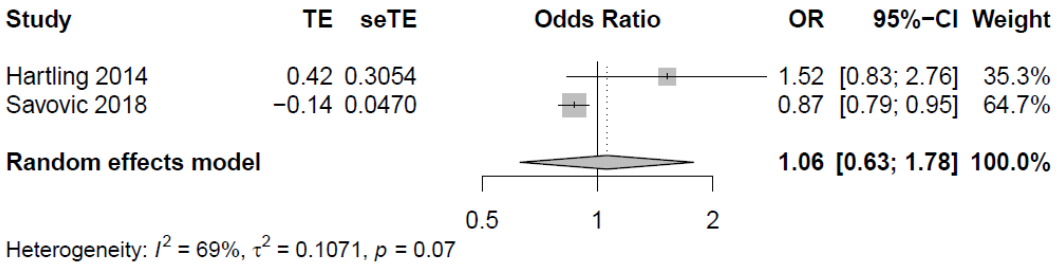

Figure 18 Double blinding - High versus low/unclear risk of bias

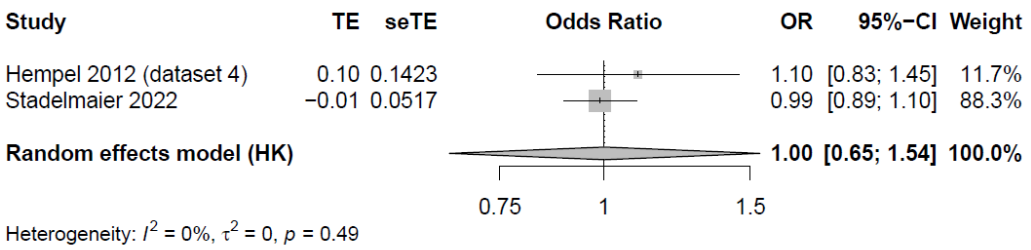

Figure 19 Compliance with intervention - Unacceptable/unclear non-compliance versus no/acceptable non-compliance

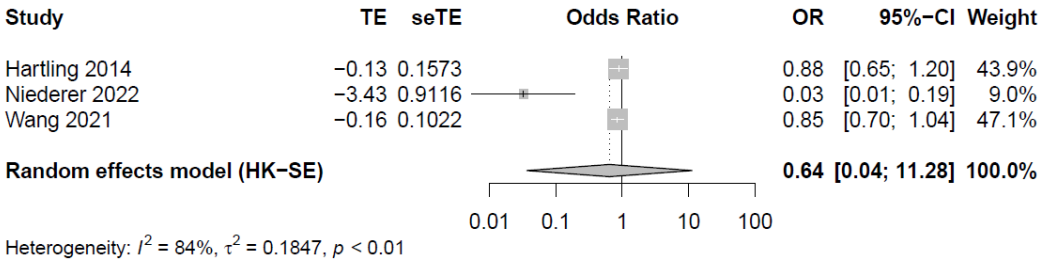

Figure 20 Missing outcome data - High versus low risk of bias

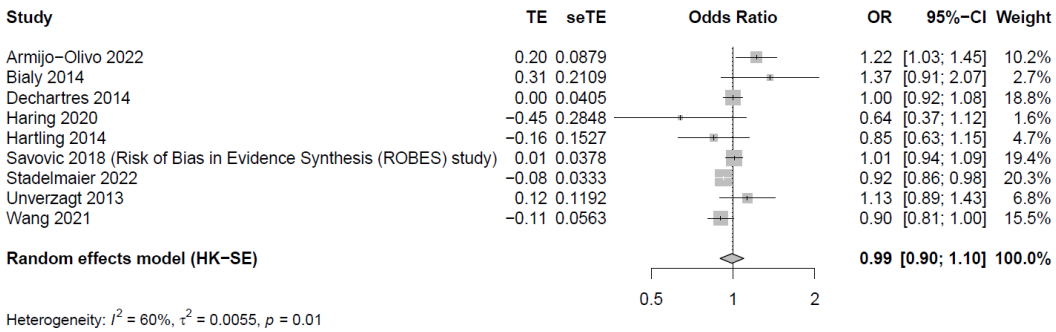

Figure 21 Missing outcome data - High/unclear versus low risk of bias

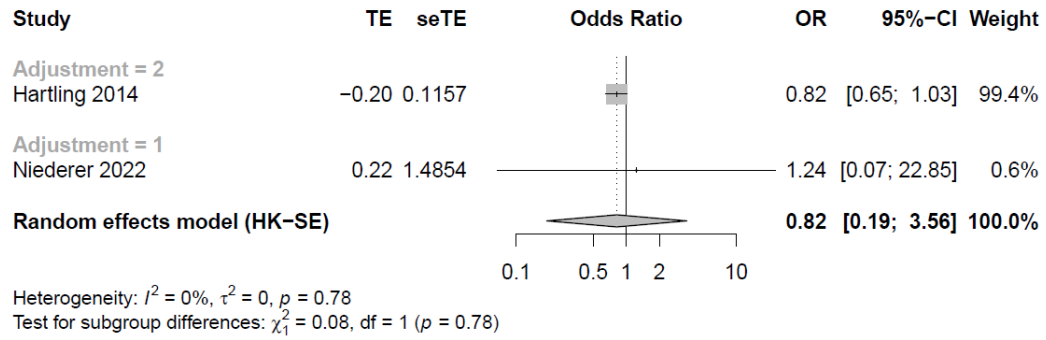

Figure 22 Selective reporting - High versus low risk of bias

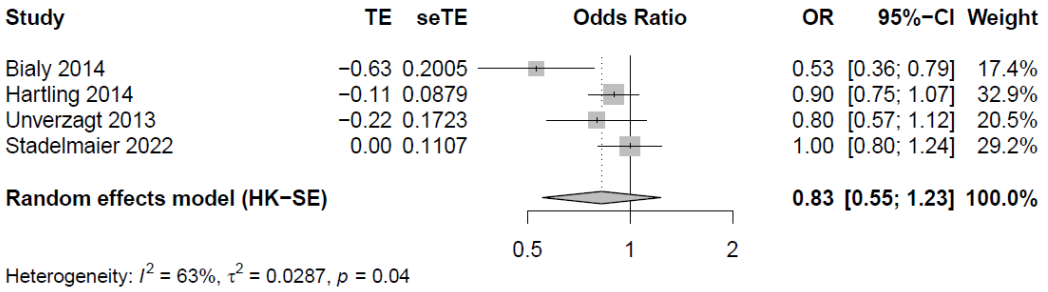

Figure 23 Selective reporting - High/unclear versus low risk of bias

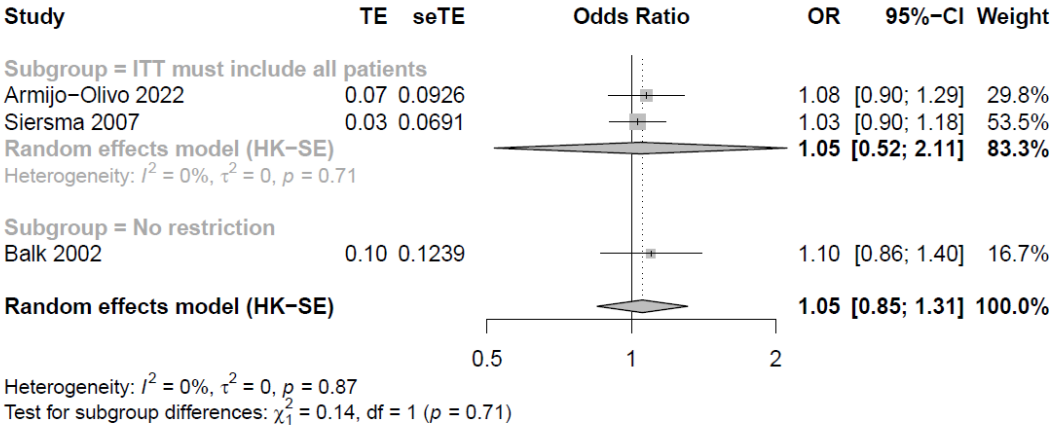

Figure 24 Intention-to-treat analysis - Not ITT/unclear versus ITT

Appendix 9 Results for other comparisons

Comparison	Number of ME studies (between- or within-trial comparisons)	Average bias (95% CI)	Direction of bias (Certainty of evidence*)
Random sequence generation			
High versus low risk of bias (in adequately concealed trials)	1 (between-trial)	ROR 0.75 (0.55-1.02)	Overestimation (Low†)
High/unclear versus low risk of bias (in adequately concealed and blinded trials)	1 (between-trial)	ROR 0.98 (0.84-1.11)	Little or no bias (Very low‡)
Allocation concealment			
High/unclear versus low risk of bias (in double-blinded trials)	1 (between-trial)	ROR 0.90 (0.83-0.98)	Overestimation (Moderate)
High/unclear versus low risk of bias (in adequately generated random sequence and blinded trials)	1 (between-trial)	ROR 0.91 (0.81-1.03)	Overestimation (Low†)
Allocation concealment approach			
Enhanced envelopes versus central randomization	1 (between-trial)	ROR 1.02 (0.85-1.22)	Little or no bias (Very low‡§)
Sealed envelopes with no further elaboration versus central randomization	1 (between-trial)	ROR 0.87 (0.76-1.00)	Overestimation (Low†)
Sealed envelopes with no further elaboration versus enhanced envelopes	1 (between-trial)	ROR 0.87 (0.73-1.05)	Overestimation (Very low‡)
Described as "randomized" with no further details versus central randomization	1 (between-trial)	ROR 0.76 (0.66-0.87)	Overestimation (Moderate)
Double blinding			
High versus low risk of bias (in adequately or unclear concealed trials)	1 (between-trial)	Any outcomes: ROR 0.83 (0.71-0.96)	Overestimation (Moderate)
High/unclear versus low risk of bias (in adequately concealed trials)	1 (between-trial)	Any outcomes: ROR 0.84 (0.73-0.95)	Overestimation (Moderate)
		Objective outcomes: ROR 0.82 (0.69-0.99)	Overestimation (Moderate)
		Subjective outcomes: ROR 0.85 (0.67-1.05)	Overestimation (Very low‡)
High/unclear versus low risk of bias (in adequately generated random sequence and concealed trials)	1 (between-trial)	Any outcomes: ROR 0.92 (0.84-1.03)	Overestimation (Low†)
		Objective outcomes: ROR 0.95 (0.81-1.10)	Little or no bias (Very low‡)
		Subjective outcomes: ROR 0.89 (0.74-1.03)	Overestimation (Low†)
Intention-to-treat analysis			
PP versus ITT/mITT (in adequately generated random sequence trials)	1 (within 107 trial)	ROR 0.98 (0.95-1.00)	Little or no bias (Moderate†)
PP versus ITT/mITT (in adequately concealed trials)	1 (within 91 trial)	ROR 0.99 (0.96-1.01)	Little or no bias (High)

ME=meta-epidemiological studies; ROR=ratio of odds ratios; CI=confidence interval; ITT=intention-to-treat analysis; PP=per-protocol analysis; mITT=modified intention-to-treat analysis.

*Studies basing on between-trial comparisons started from moderate certainty; studies basing on with-trial comparisons started from high certainty.

†Rated down once for imprecision.

‡Rated down twice for imprecision.

§Rated down once for indirectness (i.e., 10-20 comparisons).

Appendix 10 Results for subgroup analyses

Table 1 Results for subgroup analyses

Subgroup type	Is the subgroup analysis based on comparison within or between meta-epidemiological studies?	Interaction p-value	Interaction p-value <0.10?	If interaction p-value <0.10, what is the credibility of subgroup effect?
Random sequence generation				
High versus low risk of bias				
Outcome (objective, subjective)	Completely within (2 studies)	0.94	No	NA
	Both between and within	0.88	No	NA
Adjustment (adjusted, unadjusted)	Completely between (7 studies)	0.14	No	NA
High/unclear versus low risk of bias				
Outcome (objective, subjective)	Completely within (9 studies)	0.67	No	NA
	Both between and within	0.62	No	NA
Control (active, inactive)	Completely within (1 study)	**0.05**	Yes	Low
	Both between and within	0.22	No	NA
Adjustment (adjusted, unadjusted)	Completely within (2 studies)	0.53	No	NA
	Both between and within	0.57	No	NA
High/unclear versus low risk of bias (in adequately concealed and blinded trials)				
Outcome (objective, subjective)	Completely within (1 study)	0.28	No	NA
Unclear versus low risk of bias				
Outcome (objective, subjective)	Completely within (1 study)	0.17	No	NA
Allocation concealment				
High versus low risk of bias				
Outcome (objective, subjective)	Completely within (2 studies)	0.27	No	NA
	Both between and within	0.43	No	NA
Adjustment (adjusted, unadjusted)	Completely between (10 studies)	0.54	No	NA
High/unclear versus low risk of bias				
Outcome (objective, subjective)	Completely within (9 studies)	0.28	No	NA
	Both between and within	**0.06**	Yes	Low
Control (active, inactive)	Completely within (1 study)	0.34	No	NA
	Both between and within	0.39	No	NA
Adjustment (adjusted, unadjusted)	Completely within (3 studies)	0.95	No	NA
	Both between and within	0.47	No	NA
High/unclear versus low risk of bias (in double-blind trials)				
Outcome (objective, subjective)	Completely within (1 study)	0.34	No	NA
High/unclear versus low risk of bias (in adequately generated random sequence and blinded trials)				
Outcome (objective, subjective)	Completely within (1 study)	0.66	No	NA
Unclear versus low risk of bias				
Outcome (objective, subjective)	Completely within (1 study)	**0.04**	Yes	Low
Adjustment (adjusted, unadjusted)	Completely between (4 studies)	**0.04**	Yes	Low
Definitely/probably no versus definitely/probably yes				

Adjustment (adjusted, unadjusted)	Completely within (1 study)	0.95	No	NA
Baseline imbalance				
Imbalance/unclear versus balance				
Outcome (objective, subjective)	Completely within (2 studies)	0.94	No	NA
Control (active, inactive)	Completely within (1 study)	0.42	No	NA
Adjustment (adjusted, unadjusted)	Completely between (4 studies)	0.62	No	NA
Blinding of healthcare providers				
High/unclear versus low risk of bias				
Evaluate what happened or judge risk of bias	Completely between (3 studies)	0.76	No	NA
Definitely/probably no versus definitely/probably yes				
Outcome (healthcare provider decision outcomes, blinded patient- or observer-reported outcomes)	Completely within (1 study)	0.96	No	NA
Definitely/probably no or unclear versus definitely/probably yes				
Outcome (healthcare provider decision outcomes, blinded patient- or observer-reported outcomes)	Completely within (1 study)	1.00	No	NA
Adjustment (adjusted, unadjusted)	Completely within (1 study)	0.87	No	NA
Blinding of patients				
High/unclear versus low risk of bias (between-trial comparisons)				
Outcome (objective, subjective)	Completely within (2 studies)	0.72	No	NA
	Both between and within	0.80	No	NA
Adjustment (adjusted, unadjusted)	Completely within (1 study)	0.30	No	NA
	Both between and within	0.52	No	NA
Evaluate what happened or judge risk of bias	Completely between (7 studies)	0.06	Yes	Very low
Definitely/probably no versus definitely/probably yes (between-trial comparisons)				
Outcome (patient-reported outcomes, blinded observer-reported outcomes)	Completely within (1 study)	0.74	No	NA
Definitely/probably no or unclear versus definitely/probably yes (between-trial comparisons)				
Outcome (patient-reported outcomes, blinded observer-reported outcomes)	Completely within (1 study)	0.56	No	NA
Adjustment (adjusted, unadjusted)	Completely within (1 study)	0.77	No	NA
High versus low risk of bias (in adequately concealed trials) (within-trial comparisons)				
Outcome (patient-reported outcomes, blinded observer-reported outcomes)	Completely within (1 study)	<0.0001	Yes	Moderate
Blinding of data collectors				
High versus low risk of bias				
Outcome (objective, subjective)	Completely within (1 study)	0.96	No	NA
Blinding of outcome assessors				
High versus low risk of bias (between-trial comparisons)				
Outcome (objective, subjective)	Completely within (1 study)	<0.01	Yes	Moderate
Evaluate what happened or judge risk of bias	Completely between (3 studies)	0.06	Yes	Very low
High/unclear versus low risk of bias (between-trial comparisons)				

Outcome (objective, subjective)	Completely within (4 studies)	0.53	No	NA
Control (active, inactive)	Completely within (1 study)	0.48	No	NA
	Both between and within	0.61	No	NA
Evaluate what happened or judge risk of bias	Completely between (9 studies)	0.42	No	NA
Definitely/probably no or unclear versus definitely/probably yes (between-trial comparisons)				
Outcome (objective, subjective)	Completely within (1 study)	0.62	No	NA
Control (active, inactive)	Completely within (1 study)	1.00	No	NA
High versus low risk of bias (within-trial comparisons): binary outcomes				
Outcome (clearly subjective, moderately subjective)	Completely within (1 study)	0.17	No	NA
High versus low risk of bias (within-trial comparisons): measurement scale outcome				
Outcome (clearly subjective, moderately subjective)	Completely within (1 study)	0.17	No	NA
High versus low risk of bias (within-trial comparisons): time-to-event outcomes				
Outcome (objective, subjective)	Completely within (1 study)	0.41	No	NA
Double blinding				
High versus low risk of bias				
Outcome (objective, subjective)	Completely within (1 study)	0.12	No	NA
Control (active, inactive)	Completely within (1 study)	0.72	No	NA
	Both between and within	0.72	No	NA
Evaluate what happened or judge risk of bias	Completely between (7 studies)	0.58	No	NA
High/unclear versus low risk of bias				
Outcome (objective, subjective)	Completely within (8 studies)	0.60	No	NA
	Both between and within	0.39	No	NA
Control (active, inactive)	Completely within (1 study)	0.70	No	NA
	Both between and within	0.68	No	NA
Adjustment (adjusted, unadjusted)	Completely within (2 studies)	0.90	No	NA
	Both between and within	0.99	No	NA
Evaluate what happened or judge risk of bias	Completely within (1 study)	0.30	No	NA
	Both between and within	0.59	No	NA
High/unclear versus low risk of bias (in adequately concealed trials)				
Outcome (objective, subjective)	Completely within (1 study)	0.82	No	NA
High/unclear versus low risk of bias (in adequately generated random sequence and concealed trials)				
Outcome (objective, subjective)	Completely within (1 study)	0.60	No	NA
Unclear versus low risk of bias				
Outcome (objective, subjective)	Completely within (1 study)	0.17	No	NA
Compliance with intervention				
Unacceptable/unclear non-compliance versus no/acceptable non-compliance				
Outcome (objective, subjective)	Completely within (1 study)	0.88	No	NA
Missing outcome data				
High versus low risk of bias				
Adjustment (adjusted, unadjusted)	Completely between (3 studies)	<0.01	Yes	Low
High/unclear versus low risk of bias				

Outcome (objective, subjective)	Completely within (2 studies)	0.97	No	NA
	Both between and within	0.67	No	NA
Control (active, inactive)	Completely within (1 study)	0.36	No	NA
	Both between and within	0.42	No	NA
Adjustment (adjusted, unadjusted)	Completely within (1 study)	0.74	No	NA
	Both between and within	0.39	No	NA
Selective reporting				
High versus low risk of bias				
Adjustment (adjusted, unadjusted)	Completely between (2 studies)	0.78	No	NA
High/unclear versus low risk of bias				
Outcome (objective, subjective)	Completely within (2 studies)	0.63	No	NA
Control (active, inactive)	Completely within (1 study)	0.28	No	NA
Adjustment (adjusted, unadjusted)	Completely between (4 studies)	0.91	No	NA
Intention-to-treat analysis				
Not ITT versus ITT (based on meta-analyses)				
Outcome (objective, subjective)	Completely within (1 study)	0.78	No	NA
Control (placebo, not placebo)	Completely within (1 study)	0.67	No	NA
Adjustment (adjusted, unadjusted)	Completely within (1 study)	0.76	No	NA
Not ITT or unclear versus ITT (based on meta-analyses)				
Outcome (objective, subjective)	Completely within (1 study)	1.00	No	NA
Adjustment (adjusted, unadjusted)	Completely within (1 study)	0.61	No	NA
ITT must include all patients or not	Completely between (3 studies)	0.71	No	NA
PP versus ITT or mITT (based on trials)				
Control (active, inactive)	Completely within (1 study)	0.68	No	NA
Positive trials (statistical superiority of treatment versus control group), negative trials	Completely within (1 study)	**<0.01**	Yes	Low

Table 2 Credibility of subgroup effect

Subgroup type	Subgroup results	Is the subgroup effect based on comparison within rather than between meta-epidemiological studies?	If two or more within-study comparisons are available, is the effect modification similar from study to study?	For between-study comparisons, is the number of studies large?	Was the direction of subgroup effect correctly hypothesized a priori?	Does a test for interaction suggest that chance is an unlikely explanation of the apparent subgroup effect?	Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?	Did the authors use a random effects model?	If the effect modifier is a continuous variable, were arbitrary cut points avoided?	Are there any additional considerations that may increase or decrease credibility?	Overall credibility of subgroup effect
Random sequence generation											
High/unclear versus low risk of bias											
Control (active, inactive)	Active: ROR 0.51 (0.25-1.03) Inactive: ROR 1.06 (0.85-1.31)	Completely within	NA (1 study)	NA	Unclear (vague hypothesis)	Chance a likely explanation or unclear (p = 0.05)	Probably yes (3 effect modifiers tested)	Definitely yes	NA	Subgroup analysis based on both between and within studies showed no significant subgroup effect (p = 0.22)	Low
Allocation concealment											
High/unclear versus low risk of bias											
Outcome (objective, subjective)	Objective: ROR 0.98 (0.91-1.05) Subjective: ROR 0.88 (0.81-0.95)	Mostly within	Yes	NA	Unclear (vague hypothesis)	Chance a very likely explanation (p = 0.06)	Probably yes (3 effect modifiers tested)	Definitely yes	NA	Subgroup analysis based on completely within studies showed no significant subgroup effect (p = 0.28)	Low
Unclear versus low risk of bias											
Outcome (objective, subjective)	Objective: ROR 0.97 (0.55-1.72) Subjective: ROR 0.84 (0.76-0.93)	Completely within	NA (1 study)	NA	Unclear (vague hypothesis)	Chance a likely explanation or unclear (p = 0.04)	Probably yes (3 effect modifiers tested)	Definitely yes	NA	No	Low
Adjustment (adjusted, unadjusted)	Adjusted: ROR 0.70 (0.62-0.79) Unadjusted: ROR 0.85 (0.62-1.17)	Completely between	NA	Very small (1 in smallest subgroup)	Unclear (vague hypothesis)	Chance a likely explanation or unclear (p = 0.04)	Probably yes (3 effect modifiers tested)	Definitely yes	NA	No	Low
Blinding of patients											
High/unclear versus low risk of bias (between-trial comparisons)											
Evaluate what happened or	Evaluate blinding status: ROR 1.09	Completely between	NA	Very small (1 in smallest	Unclear (vague hypothesis)	Chance a very likely explanation	Probably no or unclear (4 effect	Definitely yes	NA	No	Very low

judge risk of bias	(0.75-1.57) Judge risk of bias: ROR 0.85 (0.63-1.14) Ambiguous: ROR 0.86 (0.70-1.05)			subgroup)		(p = 0.06)	modifiers tested)				
High versus low risk of bias (in adequately concealed trials) (within-trial comparisons)											
Outcome (patient-reported measurement scale outcomes, blinded observer-reported measurement scale outcomes)	Patient-reported outcomes: ROR 0.36 (0.28-0.48) Blinded observer-reported outcomes: ROR 0.96 (0.67-1.39)	Completely within	NA (1 study)	NA	Probably yes (our hypothesis is blinding of participants is more important for patient-reported outcomes)	Chance an unlikely explanation (p <0.0001)	Probably no or unclear (4 effect modifiers tested)	Definitely yes	NA	No	Moderate
Blinding of outcome assessors											
High versus low risk of bias (between-trial comparisons)											
Outcome (objective, subjective)	Objective outcomes: ROR 0.82 (0.63-1.07) Subjective outcomes: ROR 0.07 (0.02-0.33)	Completely within	NA (1 study)	NA	Probably yes (our hypothesis is blinding of outcome assessors is more important for subjective outcomes)	Chance may not explain (p <0.01)	Probably no or unclear (4 effect modifiers tested)	Definitely yes	NA	No	Moderate
Evaluate what happened or judge risk of bias	Judge risk of bias: ROR 1.12 (0.32-3.93) Ambiguous: ROR 0.82 (0.63-1.07)	Completely between	NA	Very small (1 in smallest subgroup)	Unclear (vague hypothesis)	Chance a very likely explanation (p = 0.06)	Probably no or unclear (4 effect modifiers tested)	Definitely yes	NA	No	Very low
Missing outcome data											
High versus low risk of bias											
Adjustment (adjusted, unadjusted)	Adjusted: ROR 0.03 (0.01-0.19) Unadjusted: ROR 0.86 (0.29-2.55)	Completely between	NA	Very small (1 in smallest subgroup)	Unclear (vague hypothesis)	Chance may not explain (p <0.01)	Probably yes (3 effect modifiers tested)	Definitely yes	NA	No	Low
Intention-to-treat analysis											
PP versus ITT/mITT (within-trial comparisons)											
Positive trials (statistical superiority of	Positive trials: ROR 0.93 (0.90-0.98)	Completely within	NA (1 study)	NA	Unclear (vague hypothesis)	Chance may not explain (p <0.01)	Probably no or unclear (5 effect modifiers tested)	Definitely yes	NA	No	Low

treatment versus control group), negative trials	Negative trials: ROR 0.99 (0.97-1.01)

Chapter 4: Development of the Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT)

This chapter is under review at *The British Medical Journal*.

Development of the Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT)

Ying Wang,[1] Sheri Keitz,[2] Matthias Briel,[1,3] Paul Glasziou,[4] Romina Brignardello-Petersen,[1] Reed AC Siemieniuk,[1] Dena Zeraatkar,[1,5] Elie A Akl,[1,6] Susan Armijo-Olivo,[7,8] Dirk Bassler,[9] Carrol Gamble,[10] Lise Lotte Gluud,[11] Jane Luise Hutton,[12] Luz M Letelier,[13] Philippe Ravaud,[14] Kenneth F Schulz,[15] David J Torgerson,[16] Gordon H Guyatt[1,17,18]

[1]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada

[2]Department of Medicine, Lahey Hospital & Medical Center, Burlington, MA, USA

[3]Department of Clinical Research, CLEAR-Methods Centre, University Hospital Basel and University of Basel, Basel, Switzerland

[4]Institute for Evidence-Based Healthcare, Bond University, Gold Coast, Queensland, Australia

[5]Department of Anesthesia, McMaster University, Hamilton, ON, Canada

[6]Department of Internal Medicine, American University of Beirut, Beirut, Lebanon

[7]University of Applied Sciences, Faculty of Business and Social Sciences, Osnabrück, Germany

[8]Department of Physical Therapy, Faculty of Rehabilitation Medicine, University of Alberta, Edmonton Canada

[9]Department of Neonatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

[10]Liverpool Clinical Trials Clinical Trials Research Centre, Department of Biostatistics, University of Liverpool, Liverpool, UK

[11]Gastro Unit, Copenhagen University Hospital Hvidovre, Copenhagen, Denmark

[12]Department of Statistics, The University of Warwick, Coventry, UK

[13]Department of Internal Medicine, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile

[14]Epidemiology and Statistics Sorbonne Paris Cité Research Center (CRESS), INSERM, Université Paris Descartes, Paris, France

[15]School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[16]York Trials Unit, Department of Health Sciences, University of York, York, UK

[17]Department of Medicine, McMaster University, Hamilton, Ontario, Canada

[18]MAGIC Evidence Ecosystem Foundation [www.magicevidence.org], Lovisenberggata 17C, 0456 Oslo, Norway

Correspondence to: Y Wang yingwwy@163.com

Standfirst

Recent innovations in evidence-based medicine methods, in particular instruments addressing risk of bias in randomized trials, have focused on methodological rigor at the expense of simplicity and practicability. Results of this focus could include challenges in application and loss of reliability. To address this issue, we have developed the Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT), a rigorously developed, simply structured and user-friendly instrument for assessing risk of bias of randomized controlled trials in the context of systematic reviews. This paper describes the development of ROBUST-RCT, and provides the ROBUST-RCT instrument documents and the associated manual.

Summary points

- ROBUST-RCT (Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials) is a rigorously developed, simply structured and user-friendly instrument for assessing risk of bias of randomized controlled trials in systematic reviews.
- ROBUST-RCT aims to achieve an optimal balance between simplicity and methodological rigor.
- Systematic review teams with different levels of expertise can use ROBUST-RCT when undertaking risk of bias assessments.

▪ Introduction

Although systematic reviews of randomized controlled trials (RCTs) provide the best evidence regarding the impact of health care interventions,[1] flaws in trial design and execution may result in biased estimates of effects, and hence misleading conclusions.[2] As a result, risk of bias assessment of RCTs has become an essential step in systematic reviews. Further, risk of bias represents one domain in the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) system of rating certainty of evidence, and trial limitations resulting in risk of bias may lead systematic review authors to rate down the certainty of evidence.[3 4]

Although many instruments for assessing risk of bias in RCTs exist,[5] most suffer from important limitations. A systematic survey found that existing instruments frequently include items that do not address risk of bias.[5] Risk of bias instruments, to be suitable for use in systematic reviews, should include only items addressing risk of bias issues rather than other GRADE domains.[3]

The most popular and rigorously developed instruments include those offered by the Cochrane Collaboration. The first Cochrane risk of bias instrument[6] includes an "unclear" response option that fails to take advantage of reasonable inferences regarding presence or absence of risk of bias.[7] Users reported problems with assessing the incomplete outcome data and selective reporting domains of the first Cochrane instrument.[8]

The revised Cochrane instrument (RoB 2),[9] intended to replace the first, introduced non-intuitively labelled domains and a less than straightforward series of signaling questions and algorithms for assessing each domain. The sophisticated algorithms (up to 7 signaling questions),[10] and the difficulty in understanding new terminologies (e.g., "deviations from the intended intervention that arose because of the trial context") raise challenges for systematic reviewers.[11]

Possibly as a consequence of these limitations, uptake of the RoB 2 is relatively low in non-Cochrane reviews and misapplication is frequent.[12 13] Previous published studies have documented the low inter-rater reliability of Cochrane RoB 2 and documented its challenges in implementation, even when used by systematic reviewers with substantial expertise.[14 15] Less experienced systematic reviewers - those who often assess risk of bias of individual RCTs in systematic review teams - may in particular experience daunting challenges in application of RoB 2.[11]

In considering the possibility of developing a new instrument that addresses the limitations of Cochrane RoB 2, we contacted 9 international experts very well published in the area of risk of bias assessment in RCTs. These individuals agreed regarding the limitations of RoB 2 related to its complexity, and shared the experience of the challenges their less experienced systematic review team members face in applying the instrument.

We have argued that the clinical epidemiology and evidence-based medicine (EBM) movements have lost sight of the optimal balance between simplicity and methodological rigor, and that RoB 2 represents one example of the phenomenon.[11] This perspective motivated us to use rigorous methodology, while still attending to desirable simplicity, to develop a new instrument. Our goal is an instrument that serves the needs of systematic review teams that include less experienced members assessing risk of bias. This paper describes the development of the Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT).

- **Methods**

Structure of the instrument development team

Operations Committee

Operations committee members (GG, YW, RBP, RAS, DZ) identified the need for a new instrument, developed a protocol (appendix 1), recruited the panel, organized materials, presented proposals to the panel, and created drafts of the instrument and associated materials.

Panel

The Operations Committee identified experts in risk of bias assessment from the author lists of methodological papers that explicitly stated or implicitly indicated that they addressed risk of bias issues. By screening the references of existing RCT risk of bias instruments and their guidance documents (by conducting a systematic survey),[5] as well as eligible papers suggested by operations committee members, we identified the first round of eligible papers. Then we screened the references of these papers to identify additional eligible papers. We identified 295 eligible papers in total. Individuals eligible for panel membership had participated as first, last or corresponding authors of at least one eligible paper and as co-author of at least two other papers. From a total of 63 eligible experts, stratified by region and gender, we randomly selected 10 and invited them to join the panel; 9 agreed. The panel included 2 more methodological experts (MB, PG) that committee members knew and thought could make substantial contributions.

In addition, operations committee members suggested a list of 22 internationally recognized experts in EBM education, from whom we randomly selected 2, stratified by gender, to join the panel; both agreed. The panel included a third individual (SK) known to committee members as an exceptionally astute EBM educator. The three EBM educators came from different regions.

The panel included 19 members: 5 from the operations committee and 14 additional members. Sixteen members have expertise in the methodology of risk of bias assessment (GG, YW, RBP, RAS, DZ, MB, PG, EAA, SAO, DB, CG, LLG, JLH, PR, KFS, DJT) and 3 have expertise in EBM education (SK, RJ, LML). This international collaboration

included 10 men and 9 women, 7 from North America, 5 from Europe, 3 from the UK, 2 from Oceania, 1 from South America and 1 from Asia.

Ground rules for instrument development

The following ground rules, developed by the operations committee and endorsed by panelists, guided the instrument development process.

- The instrument aims to assess risk of bias of RCTs in the context of systematic reviews.
- The objective is to develop a user-friendly instrument: item presentation will be simple and straightforward; making judgements not overly complex or difficult.
- We define bias as a systematic error or systematic deviation from the truth.
- We assume that systematic reviewers will use the GRADE approach to assess certainty of evidence.
- Decisions should be consistent with the GRADE system in distinguishing risk of bias from imprecision (random error), indirectness (applicability), and publication bias. Reporting quality represents another issue to distinguish from risk of bias.
- The instrument currently addresses only risk of bias assessment of individually-randomized parallel-group trials. We leave the risk of bias assessment of cluster-trials and crossover trials for future consideration.
- This instrument will not include items that address the detection of fraud.

Candidate items collection

To collect candidate items, we systematically surveyed the 17 RCT risk of bias instruments published from 2010 to October 2021 for their included items (see details in a separate publication).[5] We extracted additional candidate items from two studies: one collected items that Cochrane reviews regarded as "other bias" when they used the Cochrane's first risk of bias instrument;[16] another summarized the published comments on the Cochrane's first instrument.[17]

Through an item classification survey in which 13 panelists participated and judged what issue the items address (risk of bias, imprecision, indirectness, reporting quality, or none of the above), we classified the items into three categories (box 1):[5]

Box 1: Categories of items according to an item classification survey

Category 1: items that the majority judged as addressing risk of bias

Category 2: items that the majority judged as not addressing risk of bias

Category 3: participants have substantial disagreement about whether the items address risk of bias

To generate an organized item list for efficient discussion by the panel, the operations committee combined the highly related items (e.g., items addressing different aspects of missing outcome data). We removed items that specifically address issues relevant to cluster or crossover trials.

Empirical evidence from meta-epidemiological studies

To provide empirical evidence for item selection for our instrument, we conducted a systematic survey of meta-epidemiological studies examining the impact of potential risk of bias issues (items in category 1 and 3) on effect estimates in RCTs.[18] A separate paper presents the methods and results.[18]

Panel process

The operations committee presented issues and proposals to the panel. Panel meetings, co-chaired by YW and GG, used an open discussion format in which panelists first spoke freely after which GG guided the panel toward consensus. After each meeting, YW constructed minutes including the panel's tentative decisions and the discussion involved. Panelists revisited controversial issues in subsequent meetings. Through 16 1.5-hour panel meetings and associated email conversations from February to October 2023, the panel achieved consensus on item selection, instructions for included items, and format of the instrument.

The operations committee presented the organized item list to the panel. The panel discussed each item in category 2 (majority of panelists judged as not addressing risk of bias), then category 1 (majority judged as addressing risk of bias), and finally category 3 (substantial disagreement regarding addressing risk of bias or not).

The panel used 6 item selection criteria (box 2) developed by the operations committee and endorsed by the panel to help decisions regarding items in category 1 and 3. No single criterion, or group of criteria were deemed essential. The more criteria meet, the more likely the items are suitable to be selected as an item in the instrument.

Box 2: Six criteria for item selection

- Clearly risk of bias issue rather than imprecision, indirectness, publication bias, or reporting quality
- Theoretical or logical argument for why the item is important
- Information required to make judgement on the item is commonly reported in trials
- Non-expert systematic reviewers can make the judgment easily
- Problem occurs more often than rarely
- Empirical evidence supports item influence on effect estimates

The panel chose core items for the instrument, and also items of potential importance that although rejected as core items were ultimately chosen as optional items that systematic review teams might or might not consider.

The operations committee drafted instructions for core items and considerations regarding optional items. The panel discussed the draft, revised it, and approved the final version. To support the use of the instrument, we developed a manual.

User testing exercises

To identify challenges experienced by junior systematic reviewers in comprehending and applying the instrument, we conducted user tests. We enrolled 15 people who: (i) had assessed risk of bias in RCTs for at least one systematic review and (ii) had never led a systematic review of RCTs. Participants varied with respect to gender, country, clinical background, student status, and the number of systematic reviews of RCTs in which they had assessed risk of bias (appendix 2). We identified eligible individuals through suggestions from panel members. We discontinued recruitment when we achieved saturation regarding comments on the instrument.

For user testing, panelists suggested RCTs with challenges in risk of bias assessment. Two committee members (YW and GG) assessed risk of bias in these trials, and selected 5 trials (appendix 2) in which systematic reviewers would face challenges in assessing as many items in the instrument as possible. We ensured that the trials presented challenges in each item.

Each participant received one trial, the draft of the instrument and the manual. YW conducted a think-aloud interview of approximately one hour with each participant. During the interviews, participants applied the instrument to the trial, and for each item articulated the thought process that led to their assessment. YW compared the participants assessment with the assessment made and agreed on by YW and GG; when mistakes or problems occurred, she explored the reasons. Participants expressed their overall experience in applying the instrument.

To identify concerns or questions systematic review experts may have about the instrument, we conducted a second user testing exercise. We searched the Cochrane Library, randomly selected Cochrane systematic reviews published in the last five years and identified their first or last or corresponding authors. If the authors had led at least five systematic reviews of RCTs (not limited to Cochrane reviews), we invited

them to participate in the user testing. The 8 participants varied with respect to gender, country, and clinical background (appendix 2).

Before the interviews, review experts received the instrument and the manual. To explore their concerns and suggestions, YW followed a semi-structured interview guide, interviewing each participant for approximately one hour.

YW recorded and transcribed the interviews from both user testing groups and extracted people's feedback, comments, and suggestions. GG and YW reviewed the results after completing interviews for each 5 junior systematic reviewers and after 4, 6 and 8 review experts. Together, they identified concerns and solutions and presented these to the panel in email communications, ultimately deciding on modifications to the instrument and the manual. Appendix 2 summarizes feedback and resulting changes. After each revision, subsequent user testing presented participants with the updated version. The Hamilton Integrated Research Ethics Board approved the user testing study.

- **Results**

Panel's initial decisions

The item list from which the panel selected items contained 29 items: 10 items in category 1 (majority judged as addressing risk of bias in a survey), 9 in category 2 (majority judged as not addressing risk of bias), and 10 in category 3 (substantial disagreement) (appendix 3).

The panel initially selected 7 core items (6 from category 1 and 1 from category 3) and 7 optional items (2 from category 1 and 5 from category 3). Table 1 presents the extent to which these items met the item selection criteria. Appendix 3 summarizes the panel's decisions and rationale for all items.

The panel initially developed two instrument versions (tables 4 and 5 in appendix 3). Version A asks the systematic reviewers assessing individual trials to evaluate what happened in each trial for each item (e.g., item 3, judge if participants were blinded). Response options include definitely yes, probably yes, probably no, and definitely no. Version B asks the systematic reviewers assessing individual trials to decide the extent to which any deficits in instituting methodological safeguards actually result in risk of bias (e.g., item 3, judge if failure to blind participants resulted in risk of bias). Response options include definitely low, probably low, probably high, and definitely high risk of bias.

Revision based on user testing

User testing with junior systematic reviewers revealed a serious problem with the initial core item related to intention-to-treat analysis: 4 out of the first 5 reviewers, when applying the instrument to different trials, made incorrect assessments for this item (appendix 2 presents details). This problem led the panel to drop the intention to treat item from the core items list and modify an existing core item related to missing data to address the issue of participants whose outcome data were not included in the analysis for whatever reason (missing outcome data or per-protocol analysis), which became the ultimate core item 6 (table 2). In addition, the panel added the failure to avoid as-treated analysis as optional item 6 (table 3). After we made the revision, junior systematic reviewers in the subsequent user tests consistently assessed the core items correctly.

For presentation of the instrument, user testing with systematic review experts revealed that the instrument version A (evaluate what happened) may not work well in practice: review experts questioned the rationale for only using version A. One expert suggested combining the two instrument versions into a single instrument with two steps for assessing risk of bias: first step is evaluating what happened, and second step is judging risk of bias based on what happened. Regarding the two options for instrument presentation (two instrument versions or the two-step

approach), 5 review experts expressed their preferences: 2 opted for two instrument versions and 3 opted for the two-step approach. The panel ultimately decided to adopt the single instrument with the two-step approach while providing the option that systematic reviewers assessing individual trials can choose to complete only step 1 (see details below) – this option incorporates the flexibility and advantage of the two versions approach.↵

↵

ROBUST-RCT↵

Appendixes 4 and 5 present the PDF version and Word version of the ROBUST-RCT. Appendix 6 provides an Excel sheet in which systematic reviewers can enter their risk of bias assessment and thus generate a risk of bias assessment table for all trials in the systematic review. Appendix 7 presents the manual with instructions to help systematic review leaders coordinate the risk of bias assessment, and explanations and examples for each item to assist the systematic reviewers to complete the instrument.↵

↵

Core items↵

Ultimately, the ROBUST-RCT includes 6 core items (table 2, appendix 4, appendix 5). For each core item, there are two steps for assessing risk of bias. The first step is evaluating what happened, that is, whether the methodological safeguard has been implemented (e.g., step 1 of item 3 is judging if participants were blinded). For all but the last item, response options include definitely yes, probably yes, probably no, and definitely no. The second step is judging risk of bias based on what happened (e.g., step 2 of item 3 is judging risk of bias related to blinding of participants). It requires systematic review team members to decide the extent to which any deficits in instituting methodological safeguards actually resulted in risk of bias. Response options include definitely low, probably low, probably high, and definitely high risk of bias.↵

↵

Systematic reviewers assessing individual trials (i.e., risk of bias assessors) can

complete both steps. However, for core items 1-5, if the risk of bias assessors that the systematic review team recruits are less experienced and may face difficulty in judging risk of bias, review leaders may ask the risk of bias assessors to complete only step 1 and leave step 2 to the systematic reviewers with more experience.

For core item 6 'outcome data not included in analysis', there are two approaches to addressing risk of bias. One is deciding the risk of bias associated with this item for each individual trial. In this case, systematic review teams will need to set the missing percentage threshold for each response option for the step 2 of item 6 (see appendix 7 for instructions). Risk of bias assessors will determine the percentage of people not included in analysis and where that percentage falls in the risk of bias categories.

An alternative approach for core item 6 involves systematic review teams assessing risk of bias associated with missing data across the entire body of evidence at the meta-analysis level.[19 20] To test whether the inference is robust, the process of doing so begins with a complete-case analysis followed by an analysis imputing data for participants in each trial who were not included in the analysis.

For example, if for a binary outcome the complete-case analysis suggests the intervention decreases risk of undesirable event, reviewers can conduct a sensitivity analysis assuming the control group event rate in participants not included in analysis is the same as that in the participants who were included. They can further assume the event rate in intervention group participants who were not included in the analysis is higher than those included in analysis using plausible worse case assumptions.[19 20] Using this approach, risk of bias assessors can complete only step 1 in which they extract the number of participants who were not included in analysis.

Optional items

The instrument includes 8 optional items that systematic review teams could consider bringing to the attention of the risk of bias assessors (table 3; see appendix 7 for

details).

- **Discussion**

We have developed ROBUST-RCT, a simply structured and user-friendly instrument for assessing risk of bias of RCTs in systematic reviews. ROBUST-RCT provides 6 core items each of which includes two steps: the first step is evaluating what happened in individual trials and the second is judging risk of bias based on what happened. ROBUST-RCT also provides 8 optional items that systematic review authors may want to consider relevant in specific circumstances.

Strengths and limitations

We conducted preparatory work to support the development of ROBUST-RCT: a thorough collection of potential candidate items through a survey of existing RCT risk of bias instruments with an assessment of whether the potential items address risk of bias or other issues such as indirectness.[5] That process resulted in organizing the items into categories of assessing risk of bias, assessing issues other than risk of bias, and possibly assessing risk of bias.[5] A second major aspect of preparatory work was a systematic survey of meta-epidemiological studies that had addressed the impact of potential risk of bias items on effect estimates in RCTs.[18] An international panel that reviewed the preparatory material and created the instrument was balanced in both geography and gender and included experts chosen on the basis of prior publication of risk of bias methodological papers as well as highly experienced EBM educators.

We developed rigorous item selection criteria (box 2) that proved of great use in deciding on item inclusion or exclusion (table 1; appendix 3). They measured the items from different dimensions in a comprehensive and clear way. Criteria included theoretical issues, empirical support and two criteria - information required to make judgement is commonly reported, and non-expert reviewers can assess the item easily – geared to optimize the practical application of the ROBUST-RCT.

We conducted user testing with both junior reviewers and extremely experienced senior individuals. The user testing resulted in considerable refinement of the items and the presentation of the instrument (appendix 2). User testing ultimately confirmed the simplicity and the ease of practical application of ROBUST-RCT: junior systematic reviewers were able to assess the core items correctly (appendix 2). However, the user testing is limited by the relatively small number of systematic reviewers who participated.

Panelists reached consensus mainly through open discussion rather than more structured approaches such as the Delphi method. Open discussion was suitable in this case because issues of risk of bias are complex and interconnected. For instance, the issue of whether the co-interventions were balanced between groups involved the following issues: what is a sufficiently important co-intervention; if it is sufficiently important, when is there enough imbalance to consider it as high risk of bias; whether the imbalance in co-interventions is an issue of risk of bias or a function of the effect of intervention; and how easy it would be for non-expert systematic reviewers to make judgements on the above. Ultimately, the panel decided these judgments were too complex for many junior risk of bias assessors and should be included in one of the optional rather than core items. The result of the deliberation process was an extremely rich discussion of the relevant issues.

Limitations of the current ROBUST-RCT includes its addressing only risk of bias assessment of individually-randomized parallel-group trials. This will create a challenge for systematic review teams whose review includes cluster or crossover RCTs. Dealing with this situation will require referring to relevant items in instruments addressing these study designs. Our group plans to, in the future, develop extensions of ROBUST-RCT to other trial designs such as cluster or crossover trials.

Relation to prior work

In recent years, the risk of bias assessment process has become overly complex.[9 11] ROBUST-RCT is designed to address this problem by focusing on the pragmatic use of the new instrument by its target users. Strategies to achieve the goal include item selection criteria that include availability of the information required to make judgement and ease of judgement by non-expert systematic reviewers (box 2) and the user testing exercises. No prior instrument had such item selection criteria, and although Cochrane RoB 2 conducted user testing, its challenges in application suggest that its user testing did not focus on ease of application by non-expert review team members.

We considered the two steps in risk of bias assessment that are often combined in previous instruments resulting in problematic ambiguity: evaluation of whether a methodological safeguard has been implemented, and whether the failure to implement methodological safeguard results in risk of bias. Including two separate steps for assessing these different constructs increases ROBUST-RCTs transparency and conceptual clarity. Considering the different level of experience and expertise across systematic review teams, we offer flexibility regarding who completes the second step for item 1-5. A review team may require initial risk of bias assessors to complete both evaluations, or require less experienced reviewers to complete only step 1 to maximize reliability while leaving the ultimate risk of bias judgement to the more experienced review leaders. For item 6, two steps represent two approaches to assessing risk of bias for this item.

Compared to the first Cochrane risk of bias instrument,[6] instructions for ROBUST-RCT offers suggestions regarding how to classify trials into categories when the trials failed to report methodological safeguards clearly. This allows reviewers to make reasonable inferences and classify the trials as probably yes/low risk of bias or probably no/high risk of bias.

Implications

ROBUST-RCT (Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials) is a new rigorously developed, simply structured and user-friendly instrument for assessing risk of bias of RCTs in systematic reviews. We believe the development of ROBUST-RCT achieved its aim of an optimal balance between simplicity and methodological rigor and can be used by review teams with different levels of expertise when doing risk of bias assessments. While our extensive pretesting provides evidence of the feasibility and acceptability of ROBUST-RCT, wider use may reveal limitations that we could correct. We therefore encourage future users who experience such limitations to bring these to our attention.

Table 1. Initially selected core items and optional items and judgement regarding whether they met the six item selection criteria*

Items	Item selection criteria					
	Clearly risk of bias rather than others	Theoretical or logical argument for why the item is important	Information required to make judgement is commonly reported in trials	Non-expert reviewers can make the judgment easily	Problem occurs more often than rarely	Empirical evidence supports item influence on effect estimates†
Initially selected core items						
Random sequence generation	√ (category 1)	√	√	√	√	Overestimation (Moderate certainty)
Allocation concealment	√ (category 1)	√	√	√	√	Overestimation (Moderate certainty)
Blinding of participants	√ (category 1)	√	√	√	√	Any outcomes: Very uncertain Patient-reported outcomes: Overestimation (Moderate certainty) Observer-reported or objective outcomes: Very uncertain
Blinding of healthcare providers	√ (category 1)	√	√	√	√	Very uncertain
Blinding of outcome assessors	√ (category 1)	√	√	√	√	Any outcomes: Very uncertain Objective outcomes: Very uncertain Subjective outcomes: Overestimation (High certainty)
Missing outcome data	√ (category 1)	√	√	√	√	Underestimation (Low certainty)
Intention-to-treat analysis‡	× (category 3)	√	×	√	√	Very uncertain
Initially selected optional items						
Whether the baseline prognostic factors were balanced between groups	× (category 3)	√	√	×	?	Very uncertain
Whether the co-interventions	× (category 3)	√	×	×	?	Overestimation (Low certainty)

were balanced between groups in blinded trials						
Whether the outcome assessment or data collection differed between groups	√ (category 1)	√	×	×	×	No evidence
Whether the follow-up time, frequency, or intensity of outcome assessment differed between groups	√ (category 1)	√	×	×	×	No evidence
Whether the outcome measurement method was valid (i.e., validity of outcome measurement)	× (category 3)	√	×	×	×	No evidence
Whether there was selective reporting	× (category 3)	√	×	×	×	Very uncertain
Whether the trial was terminated early for benefit	× (category 3)	√	√	√	√	Overestimation (Moderate certainty)

*Appendix 3 summarizes judgement regarding whether all items in category 1 and 3 met the item selection criteria.

†Empirical evidence came from a systematic survey of meta-epidemiological studies.[18]

‡After user testing, the panel split the initial intention-to-treat analysis item into two issues: per-protocol analysis and as-treated analysis. The panel combined the per-protocol analysis issue with the missing outcome data issue together as the ultimate core item 6, and added the as-treated analysis issue as the optional item 6.

Table 2. ROBUST-RCT core items and two steps

	Step 1 Evaluate what happened	**Step 2 Judge risk of bias**
Item 1 Random sequence generation	Was the allocation sequence adequately generated	Judge risk of bias related to sequence generation
Item 2 Allocation concealment	Was the allocation adequately concealed	Judge risk of bias related to allocation concealment
Item 3 Blinding of participants	Were participants blinded	Judge risk of bias related to blinding of participants
Item 4 Blinding of healthcare providers	Were healthcare providers blinded	Judge risk of bias related to blinding of healthcare providers
Item 5 Blinding of outcome assessors	Were outcome assessors blinded	Judge risk of bias related to blinding of outcome assessors
Item 6 Outcome data not included in analysis	Extract the number of participants who were not included in analysis in each group	Judge risk of bias related to the overall percentage of participants not included in analysis
Response options	Definitely Yes; Probably Yes; Probably No; Definitely No (except for item 6)	Definitely Low; Probably Low; Probably High; Definitely High

Table 3. ROBUST-RCT optional items*

Optional item 1	Whether the baseline prognostic factors were balanced between groups
Optional item 2	Whether the co-interventions were balanced between groups in blinded trials
Optional item 3	Whether the outcome assessment or data collection differed between groups
Optional item 4	Whether the follow-up time, frequency, or intensity of outcome assessment differed between groups
Optional item 5	Whether the outcome measurement method was valid (i.e., validity of outcome measurement)
Optional item 6	When investigators conducted an as-treated analysis, was the percentage of participants not analyzed in the groups to which they were randomized sufficiently low
Optional item 7	Whether there was selective reporting
Optional item 8	Whether the trial was terminated early for benefit

*Please refer to the manual (appendix 7) for considerations regarding when systematic reviewers might or might not include the optional items.

Copyright © 2024 McMaster University, Hamilton, Ontario, Canada. Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT), authored by Wang et al, is the copyright of McMaster University (Copyright ©2024, McMaster University). The Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT) must not be copied, distributed, or used in any way without the prior written consent of McMaster University. Contact the McMaster Industry Liaison Office at McMaster University, email: milo@mcmaster.ca for licensing details.

Acknowledgements: We thank Rod Jackson for his contribution to the instrument development. We thank Chunyang Bai, Vamsikalyan Borra, Fatemeh Mirzayeh Fashami, Yasaman Hamidianshirazi, Tanvir Jassal, Sumanth Khadke, Bhargav Makwana, Sahith Rajkumar, Christof Schönenberger, Johannes Schwenke, Jessyca Silva, Michael Wu, Daniel Xie, Shuihua Yang, Puwen Zhang, Qiukui Hao, Susan Hillier, Meixuan Li, Gabriele Meyer, Vassiliki Sinopoulou, Nicole Skoetz, Kehu Yang, and Liang Yao for participating the user testing of the instrument. We thank Prashanti Eachempati for her idea about the Excel sheet including the drop-down listing and colour coding, which inspired the creation of the Excel sheet used in this publication. We thank Stefan Schandelmaier, Prashanti Eachempati, and Derek Chu for their suggestions and comments after using the instrument.

Contributors: GG and YW developed the initial idea. Operations committee members (GG, YW, RBP, RAS, DZ) identified the need for a new instrument, developed a protocol, recruited the panel, organized materials, presented proposals to the panel, and created drafts of the instrument and associated materials. Panel members (GG, YW, RBP, RAS, DZ, MB, PG, EAA, SAO, DB, CG, LLG, JLH, PR, KFS, DJT, SK, RJ, LML) participated in the open discussion in panel meetings. GG and YW are co-chairs of the panel meetings. YW and GG drafted the manuscript, and others reviewed and revised the manuscript. All authors approved the final version of the manuscript, instrument, and manual.

Funding: This work was supported by the Einstein Foundation Berlin as part of the Einstein Foundation Award for Promoting Quality in Research. The contents are those

of the authors and do not necessarily represent the official views of, nor an endorsement by, the Einstein Foundation or the award jury.

Competing interests: Authors do not have any personal financial interests that are relevant to the work.

Patient and public involvement: Patients were not involved in this work.

References

1. Guyatt GH, Rennie D, Meade MO, et al. Users' guide to the medical literature : a manual for evidence-based clinical practice. New York: McGraw-Hill Education 2015.
2. Verhagen AP, de Vet HC, de Bie RA, et al. The art of quality assessment of RCTs included in systematic reviews. *J Clin Epidemiol* 2001;54(7):651-4. doi: 10.1016/s0895-4356(00)00360-7
3. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026 [published Online First: 20101231]
4. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol* 2011;64(4):407-15. doi: 10.1016/j.jclinepi.2010.07.017 [published Online First: 20110119]
5. Wang Y, Ghadimi M, Wang Q, et al. Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues. *J Clin Epidemiol* 2022;152:218-25. doi: 10.1016/j.jclinepi.2022.10.018 [published Online First: 20221028]
6. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928. doi: 10.1136/bmj.d5928 [published Online First: 20111018]
7. Akl EA, Sun X, Busse JW, et al. Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. *J Clin Epidemiol* 2012;65(3):262-7. doi: 10.1016/j.jclinepi.2011.04.015 [published Online First: 20111224]
8. Savovic J, Weeks L, Sterne JA, et al. Evaluation of the Cochrane Collaboration's tool for assessing the risk of bias in randomized trials: focus groups, online survey, proposed recommendations and their implementation. *Syst Rev* 2014;3:37. doi: 10.1186/2046-4053-3-37 [published Online First: 20140415]
9. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898. doi: 10.1136/bmj.l4898 [published Online First: 2019/08/30]
10. Moore THM, Higgins JPT, Dwan K. Ten tips for successful assessment of risk of bias in randomized trials using the RoB 2 tool: Early lessons from Cochrane. *Cochrane Evidence Synthesis and Methods* 2023;1(10):e12031. doi: https://doi.org/10.1002/cesm.12031
11. Kuehn R, Wang Y, Guyatt G. Overly complex methods may impair pragmatic use of core evidence-based medicine principles. *BMJ Evid Based Med* 2024 doi: 10.1136/bmjebm-2024-112868 [published Online First: 20240307]
12. Babic A, Barcot O, Viskovic T, et al. Frequency of use and adequacy of Cochrane risk of bias tool 2 in non-Cochrane systematic reviews published in 2020: Meta-research study. *Research synthesis methods* 2024 doi: 10.1002/jrsm.1695 [published Online First: 20240123]
13. Martimbianco ALC, Sa KMM, Santos GM, et al. Most Cochrane systematic reviews and protocols did not adhere to the Cochrane's risk of bias 2.0 tool. *Rev Assoc Med Bras (1992)* 2023;69(3):469-72. doi: 10.1590/1806-9282.20221593 [published Online First: 20230220]
14. Crocker TF, Lam N, Jordao M, et al. 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. *J Clin Epidemiol* 2023;161:39-45. doi: 10.1016/j.jclinepi.2023.06.015 [published Online First: 20230624]

15. Minozzi S, Cinquini M, Gianola S, et al. 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. doi: 10.1016/j.jclinepi.2020.06.015 [published Online First: 2020/06/21]
16. Babic A, Pijuk A, Brazdilova L, et al. The judgement of biases included in the category "other bias" in Cochrane systematic reviews of interventions: a systematic survey. *BMC Med Res Methodol* 2019;19(1):77. doi: 10.1186/s12874-019-0718-8 [published Online First: 20190411]
17. Jorgensen L, Paludan-Muller AS, Laursen DR, 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. doi: 10.1186/s13643-016-0259-8 [published Online First: 20160510]
18. Wang Y, Parpia S, Couban R, et al. Compelling evidence from meta-epidemiological studies demonstrates overestimation of effects in randomized trials that fail to optimize randomization and blind patients and outcome assessors. *J Clin Epidemiol* 2024;165:111211. doi: 10.1016/j.jclinepi.2023.11.001 [published Online First: 20231107]
19. Goldkuhle M, Bender R, Akl EA, et al. GRADE Guidelines: 29. Rating the certainty in time-to-event outcomes-Study limitations due to censoring of participants with missing data in intervention studies. *J Clin Epidemiol* 2021;129:126-37. doi: 10.1016/j.jclinepi.2020.09.017 [published Online First: 20200930]
20. Guyatt GH, Ebrahim S, Alonso-Coello P, et al. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. *J Clin Epidemiol* 2017;87:14-22. doi: 10.1016/j.jclinepi.2017.05.005 [published Online First: 2017/05/23]

Appendix 1. Protocol

Protocol for the Development of an Instrument for Assessing Risk of Bias of Randomized Trials in the Context of Systematic Reviews

1. Background

Clinical researchers and methodologists generally agree that randomized controlled trials (RCTs) and systematic reviews of such trials provide the most trustworthy evidence regarding causal inferences and effects of interventions.[1] The effect estimates presented in RCTs can, however, be biased due to flaws in design and execution of the studies. Given this concern, risk of bias assessment of individual studies has become an essential part in systematic reviews which influences the decision of whether to rate down GRADE certainty of evidence due to risk of bias.[2]

Investigators have developed many instruments for assessing risk of bias of RCTs.[1 3-7] However, even the most widely used tools - those developed by the Cochrane Collaboration[1 4] - suffer from limitations. Limitations of the first Cochrane risk of bias tool[1] include the "unclear" response option that fails to take advantage of reasonable inferences regarding presence or absence of risk of bias,[8 9] and the confusion about rating domains of incomplete outcome data and selective reporting.[10] Aiming to overcome these and other limitations, the Cochrane group developed the revised Cochrane instrument (RoB 2).[4] Although sophisticated, the RoB 2 has proved to have low inter-rater reliability, complexity of implementation, and challenges in its application.[11]

Therefore, we aim to develop a new user-friendly and credible instrument for assessing risk of bias of RCTs in the context of systematic reviews.

2. Ground rules of the instrument

The operations committee developed the following ground rules:

- The instrument aims to assess risk of bias of RCTs in the context of systematic reviews.

- The objective is to develop a user-friendly instrument: item presentation will be simple and straightforward; making judgements not overly complex or difficult.
- We define bias as a systematic error or systematic deviation from the truth.
- We assume that systematic reviewers will use the GRADE approach to assess certainty of evidence.
- Decisions should be consistent with the GRADE system in distinguishing risk of bias from imprecision (random error), indirectness (applicability), and publication bias. Reporting quality represents another issue to distinguish from risk of bias.
- The instrument currently addresses only risk of bias assessment of individually-randomized parallel-group trials. We leave the risk of bias assessment of cluster-trials and crossover trials for future consideration.
- This instrument will not include items that address the detection of fraud.

3. Methods

3.1. ASSEMBLE GROUP

3.1.1. Operations committee

The operations committee includes 5 members (Gordon Guyatt, Romina Brignardello-Petersen, Reed A Siemieniuk, Dena Zeraatkar, and Ying Wang). The operations committee came up with the idea of developing the new instrument and is responsible for execution of review work, organizing materials, and presenting proposals to the panel.

3.1.2. Panel

To complement the operations committee, we will invite additional experts in risk of bias methodology and evidence-based medicine education to join the panel.

Operation committee members invited Matthias Briel and Paul Glasziou to join as methodological experts because they could have substantial contribution. Operation

committee members suggested Sheri Keitz to join because she is an exceptionally astute evidence-based medicine educator.

We will invite 10 other risk of bias methodological experts and 2 other experts in evidence-based medicine education, which we will randomly select from a list of risk of bias methodological experts and a list of internationally recognized experts in evidence-based medicine education, stratified by gender and region.

Risk of bias methodological experts

We will collect potential risk of bias methodological experts from a list of authors of risk of bias methodological papers.

We will identify risk of bias methodological papers from four resources:

a) References of existing RCT risk of bias instruments;
b) References of RCT risk of bias instrument guidance documents describing the use of RCT risk of bias instruments;
c) Eligible papers that are suggested by operations committee members;
d) References of eligible papers that are identified from the above three resources.

Papers in which authors stated explicitly or implicitly indicated that what they addressed in the paper is a risk of bias issue, are eligible. Table 1 presents possible classification of eligible papers.

Eligible risk of bias methodological experts should be first or last or corresponding author of at least one eligible paper and co-author of at least two other eligible papers.

Evidence-based medicine educators

The operations committee members will suggest a list of internationally recognized experts in evidence-based medicine education from which we will randomly select two.

3.2. DEVELOPMENT OF THE INSTRUMENT

Development of the new RCT risk of bias instrument will consist of 4 steps: collection of potential candidate items for the instrument and develop an organized item list; collection of empirical evidence; a panel process to identify items that will be included in the new instrument and design the instrument; and user testing of the new instrument.

Step 1 - Collect potential candidate items and develop an organized item list

We will collect potential candidate items from three resources: included items of the existing RCT risk of bias instruments; items regarded as "other bias" in systematic reviews; comments on Cochrane risk of bias tool 1.0.

a) *We will conduct a systematic survey of existing RCT risk of bias instruments published in the last decade*

This systematic survey will include instruments that purport to assess risk of bias, internal validity, quality, or methodological quality of RCTs. Typical types of the instruments include scales, checklists, and domain-based tools. We will only include instruments published (or presented in websites) in 2010 or later.

We will search Medline, Embase, Web of Science, and Scopus from 2010. We will develop the search strategy in collaboration with an experienced research librarian. We will scan the reference lists of existing systematic surveys of similar topics to identify other potential eligible studies. We will extract included items of the existing instruments.

b) *Collect the items regarded as "other bias" in systematic reviews from an existing systematic survey*

Cochrane risk of bias tool 1.0 has a domain of “other bias”. A systematic survey published in 2019 summarized the items that Cochrane reviews regarded as “other bias” when they assessing risk of bias using Cochrane risk of bias tool 1.0.[12] We will

extract items from this systematic survey.

c) *Collect items from an existing systematic survey of published comments of Cochrane risk of bias tool 1.0*

A review published in 2016 summarized the comments of Cochrane risk of bias tool 1.0. Some of the comments were related to the additional items that may need to be included in the RCT risk of bias instrument.[13] We will extract items from this systematic survey.

After collecting items, we will conduct a survey of panel members regarding what issue each item addresses: risk of bias, applicability (imprecision), random error (imprecision), reporting quality, or none of the above. Based on the number of respondents that choose each option, we will classify items into three categories:

Category 1: items that the majority thought address risk of bias

Category 2: items that the majority thought do not address risk of bias

Category 3: items with substantial disagreement regarding whether they address risk of bias

Step 2 – Collect empirical evidence

Meta-epidemiological studies provide empirical evidence regarding how risk of bias issues influence effect estimates. We will conduct a systematic survey of meta-epidemiological studies examining the influence of potential risk of bias issues (items in category 1 and 3) on effect estimates in RCTs.

Three existing systematic surveys on this topic searched till to 2015.[14-16] We will review meta-epidemiological studies included in these three systematic surveys and assess their eligibility. We will conduct an updated search of Medline, Embase, Web of Science, and the Cochrane Database of Systematic Reviews from 2015.

Step 3 – Panel process

We will hold panel meetings to discuss and decide which items to be included in the new instrument; response options for single items, and to develop a manual that helps systematic reviewers completing the instrument.

We will use the below item selection criteria to help deciding which items to be included in the instrument. All criteria are desirable but not essential, i.e., final decision is up to the panel.

- Clearly risk of bias issue rather than imprecision, indirectness, publication bias, or reporting quality
- Theoretical or logical argument for why the item is important
- Information required to make judgement on the item is commonly reported in trials
- Non-expert systematic reviewers can make the judgment easily
- Problem occurs more often than rarely
- Empirical evidence supports item influence on effect estimates

The panel will first discuss items in category 2 (items that the majority thought not address risk of bias), then category 1 (items that the majority thought address risk of bias), and finally category 3 (items with substantial disagreement regarding whether they address risk of bias).

Since some items are highly related and only differ in details, during the item selection process, items may be first considered by groups if necessary. The panel will discuss the detailed issues during item selection or when developing instructions for items.

Step 4 – User testing

a) User testing with junior systematic reviewers

We will conduct a user testing of the new instrument to identify challenges experienced by potential users in applying the instrument, which will be used to improve the instrument. We will include people who have assessed risk of bias of RCTs

for at least one systematic review and have never led any systematic review of RCTs themselves. We plan to first recruit 15 people. To achieve saturation, we will continue to enroll people until they do not identify any new issues or problems in applying the new instrument.

Our panel members will send an invitation email to the potential eligible participants that they know. In the email, panel members will introduce this study, confirm the potential participants' eligibility, and ask their interests in participating in this study. The consent form will be sent with the email. If potential participants' eligibility is determined and they agree to participate, they will sign the consent form and send it back to the panel.

After obtaining consent, we will email participants a link to a LimeSurvey survey to collect their demographic information (Box 1).

Box 1. Demographic survey

- Participant ID
- Gender
- Country
- Clinical background: physician vs. pharmacist vs. nurse vs. others (specify) vs. no clinical background
- Student status: undergraduate student vs. master student vs. PhD student vs. not a student
- How many systematic reviews have you assessed risk of bias of included RCTs for: 1-2 vs. 3-5 vs. >5

Based on the demographic information, we will select 15 people that vary with respect to gender, country, clinical background, student status, and number of systematic reviews in which they have assessed risk of bias of RCTs.

We will select trials based on the following criteria:

i) Trials must be individually randomized parallel-group trials (described as

"randomized" by trial authors).↵

ii) We will prioritize trials with which systematic reviewers will face difficulty in assessing as many items in the instrument as possible (for step 1 we considered the trials in which judgements were probably yes or probably no as challenging; for step 2 we considered it challenging when the assessment involved judgement regarding whether the failure of implementation of methodological safeguard resulted in risk of bias). Challenge in each item must be faced in at least one final selected trial.↵

iii) Risk of bias assessments for the final selected trials vary across the response options for each item.↵

↵

Panel members will suggest potential trials. To identify trials that meet our criteria, Ying Wang and Gordon Guyatt will apply the risk of bias instrument to the trials that the team members suggest. We plan to first select 5 trials. If 5 trials prove insufficient to meet our criteria, we will select more trials until we find sufficient trials that raise issues with each item. Ying will discuss trial selection with Gordon Guyatt and possibly with other panel members.↵

↵

We will conduct a think-aloud interview with each of the 15 people online via Zoom. Prior to the interviews, we will send one trial and draft of the instrument and the manual to each participant. During the interview, participants will apply the instrument to the assigned trial. As they apply the instrument, we will instruct them to read the question, instruction, and manual for each item aloud and verbalize what they are thinking as they do so. When they apply each item to the trial, they will verbalize all the thought processes that lead to their assessment. The aim is to understand any challenges that participants have in comprehending and applying items to the trials. After they finish applying the instrument, we will ask their overall experience. We will record and transcribe the interviews, and extract participants' challenges in comprehending and applying the instrument.↵

↵

As soon as we find a clear problem in the instrument or manual, we will make a correction immediately and continue the use testing with the corrected version.↵

↵

b) User testing with systematic review experts↵

We will conduct another user testing of the instrument with systematic review experts to identify concerns or questions systematic review leaders may have about the instrument and the manual, which will be used to improve the instrument.↵

↵

We will search Cochrane Library, randomly select Cochrane systematic reviews and identify the first or last or corresponding authors. If the authors have led at least five systematic reviews of RCTs, by checking in PubMed and Google, we will invite them to participate in the user testing. We will first recruit 8 participants that vary with respect to gender, country, clinical background, and number of systematic reviews they have ever led. To achieve saturation, we will continue to enroll participants until they do not identify any new issues or problems about instrument.↵

↵

We will send them the instrument and the manual one week before the interview and ask them to read these documents before the interview.↵

↵

During the interviews, we will collect participants' concerns or suggestions on the instrument and the manual. We will record and transcribe the interviews. We will extract participants' concerns or suggestions and any problems they may have if using this instrument.↵

↵

As soon as we find a clear problem in these documents, we will make a correction immediately and continue the use testing with the corrected version.↵

↵

Table 1. Possible categories of risk of bias methodological papers*

Classification	Subclassification
Overall risk of bias	Studies addressing categorization/dimension of bias in RCTs Studies illustrating approach to incorporating risk of bias assessment of RCTs into meta-analysis Studies about RCT risk of bias instrument
Random sequence generation	Empirical study illustrating how random sequence generation influences results Theoretical discussion about why random sequence generation is important Studies introducing/criticizing specific randomization approach Studies about baseline imbalance
Allocation concealment	Empirical study illustrating how allocation concealment influences results Theoretical discussion about why allocation concealment is important Studies introducing/criticizing specific allocation concealment approach Studies about deciphering allocation sequences Studies about reporting of allocation concealment
Blinding	Empirical study illustrating how blinding influences results Theoretical discussion about why blinding is important Introducing/criticizing blinding method Interpretation of blinding terminology/ reporting of blinding status Testing the successful of blinding

Missing outcome data	Empirical study illustrating how missing data influences results Methodological paper introducing how to deal with missing outcome data in RCTs Methodological paper introducing how to deal with missing outcome data in meta-analyses Systematic survey investigating how RCTs dealing with missing outcome data Systematic survey investigating how meta-analyses dealing with missing outcome data Strategies for identifying missing data Relationship between missing data and intention to treat
Intention to treat	Methodological papers advocating intention to treat analysis Meaning/definition of intention to treat Empirical study illustrating how deviation from intention to treat analysis influences results Effect of adhering to intervention Post-randomization exclusion
Selective reporting	Empirical study illustrating how selective reporting influences results Assessing selective reporting in RCTs General discussion about selective reporting Categorization of selective reporting in RCTs
Stop early for benefit	Empirical study illustrating how stop early for benefit influences results Theoretical discussion about why stop early for benefit is or is not a risk of bias issue Simulation study illustrating how stop early for benefit influences results Empirical study examining the prevalence of stop early for benefit or reviews how to deal with stop early for benefit
Center status	Empirical study illustrating how center status influences results
Funding	Empirical study illustrating how funding influences results

	Theoretical discussion about why funding is or is not a risk of bias issue
Empirical study/ meta-epidemiological study involving more than one risk of bias issue	Empirical study illustrating how items influence results Empirical study examining the prevalence of several items
Making judgement when poor reporting	Identify which risk of bias item(s) is/are examined
Cluster/crossover RCTs	NA

*If authors stated explicitly or implicitly indicated that what they addressed is a risk of bias issue, the papers are eligible.

References

1. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928. doi: 10.1136/bmj.d5928 [published Online First: 2011/10/20]
2. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *J Clin Epidemiol* 2011;64(4):407-15. doi: 10.1016/j.jclinepi.2010.07.017 [published Online First: 2011/01/21]
3. Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol* 1998;51(12):1235-41. doi: 10.1016/s0895-4356(98)00131-0 [published Online First: 1999/03/23]
4. Sterne JAC, Savovic J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898. doi: 10.1136/bmj.l4898 [published Online First: 2019/08/30]
5. Sherrington C, Herbert RD, Maher CG, et al. PEDro. A database of randomized trials and systematic reviews in physiotherapy. *Manual therapy* 2000;5(4):223-6. doi: 10.1054/math.2000.0372 [published Online First: 2000/10/29]
6. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled clinical trials* 1996;17(1):1-12. doi: 10.1016/0197-2456(95)00134-4 [published Online First: 1996/02/01]
7. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of epidemiology and community health* 1998;52(6):377-84. doi: 10.1136/jech.52.6.377 [published Online First: 1998/10/09]
8. Akl EA, Sun X, Busse JW, et al. Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. *J Clin Epidemiol* 2012;65(3):262-7. doi: 10.1016/j.jclinepi.2011.04.015 [published Online First: 2011/12/28]
9. Guyatt GH, Busse J. Methods Commentary: Risk of Bias in Randomized Trials 1.
10. Savovic J, Weeks L, Sterne JA, et al. Evaluation of the Cochrane Collaboration's tool for assessing the risk of bias in randomized trials: focus groups, online survey, proposed recommendations and their implementation. *Syst Rev* 2014;3:37. doi: 10.1186/2046-4053-3-37 [published Online First: 20140415]
11. Minozzi S, Cinquini M, Gianola S, et al. 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. doi: 10.1016/j.jclinepi.2020.06.015 [published Online First: 2020/06/21]
12. Babic A, Pijuk A, Brazdilova L, et al. The judgement of biases included in the category "other bias" in Cochrane systematic reviews of interventions: a systematic survey. *BMC Med Res Methodol* 2019;19(1):77. doi: 10.1186/s12874-019-0718-8 [published Online First: 20190411]
13. Jorgensen L, Paludan-Muller AS, Laursen DR, 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. doi: 10.1186/s13643-016-0259-8 [published Online First: 20160510]

14. Berkman AD, Santaguida PL, M. V, et al. The Empirical Evidence of Bias in Trials Measuring Treatment Differences: Agency for Healthcare Research and Quality (US) 2014.
15. Dechartres A, Trinquart L, Faber T, et al. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. *J Clin Epidemiol* 2016;77:24-37. doi: 10.1016/j.jclinepi.2016.04.005 [published Online First: 20160429]
16. Page MJ, Higgins JP, Clayton G, et al. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. *PloS one* 2016;11(7):e0159267. doi: 10.1371/journal.pone.0159267 [published Online First: 20160711]

Appendix 2. Details of the user testing of the ROBUST-RCT

Table 1. Characteristics of the junior systematic reviewers in the user-testing

Characteristic	Number of participants (total 15)
Female	7
Country	
Canada	7
China	3
India	2
Switzerland	2
US	1
Clinical background	
Physician	7
Pharmacist	2
Dietitian	1
No clinical background	5
Student status	
PhD student	2
Master student	6
Undergraduate student	3
Not student	4
Number of systematic reviews in which they have assessed risk of bias of randomized trials	
1-2	10
3-5	3
>5	2

Table 2. Trials that used in user testing with junior systematic reviewers

1. Chochinov HM, Kristjanson LJ, Breitbart W, et al. Effect of dignity therapy on distress and end-of-life experience in terminally ill patients: a randomised controlled trial. Lancet Oncol. 2011 Aug;12(8):753-62. doi: 10.1016/S1470-2045(11)70153-X.
2. Hansen CD, Gram-Kampmann EM, Hansen JK, et al. Effect of Calorie-Unrestricted Low-Carbohydrate, High-Fat Diet Versus High-Carbohydrate, Low-Fat Diet on Type 2 Diabetes and Nonalcoholic Fatty Liver Disease : A Randomized Controlled Trial. Ann Intern Med. 2023 Jan;176(1):10-21. doi: 10.7326/M22-1787.
3. Lou W, Xia Y, Xiang P, et al. Prevention of upper gastrointestinal bleeding in critically ill Chinese patients: a randomized, double-blind study evaluating esomeprazole and cimetidine. Curr Med Res Opin. 2018 Aug;34(8):1449-1455. doi: 10.1080/03007995.2018.1464132.
4. Nayyab I, Ghous M, Shakil Ur Rehman S, et al. The effects of an exercise programme for core muscle strengthening in patients with low back pain after Caesarian-section: A single blind randomized controlled trial. J Pak Med Assoc. 2021 May;71(5):1319-1325. doi: 10.47391/JPMA.596.
5. Parekh DJ, Reis IM, Castle EP, et al. Robot-assisted radical cystectomy versus open radical cystectomy in patients with bladder cancer (RAZOR): an open-label, randomised, phase 3, non-inferiority trial. Lancet. 2018 Jun 23;391(10139):2525-2536. doi: 10.1016/S0140-6736(18)30996-6.

Table 3. Summary of feedback from user testing with junior systematic reviewers and resulting changes

Items	Feedback or problem found in the user testing	Resulting changes
Instrument version A (Ultimate first step)		
Item 1: Was the allocation sequence adequately generated	1. One out of the first five reviewers got confused by the example "shuffling cards or envelopes" in definitely yes.	1. Since we provide a number of examples, and shuffling cards or envelopes is seldom used, the panel removed this specific example.
	2. The initial instruction for probably yes was "Trial was described as 'randomized' without further details regarding the method of generating the allocation sequence, and the text described the allocation concealment method". As long as the trial reported using central allocation, drug containers or envelopes, the criterion is satisfied (e.g., no need to explicitly state sequentially numbered opaque sealed envelopes). One out of the second five reviewers had problem assessing this correctly.	2. We added a parenthesis as follows: "Trial was described as 'randomized' without further details regarding the method of generating the allocation sequence, and the text described the allocation concealment method (stated using central allocation, drug containers, or envelopes)".
Item 2: Was the allocation adequately concealed	1. Initially we had independent central allocation and independent pharmacy-controlled randomization in definitely yes; and central allocation or pharmacy-controlled randomization without specification of independent in probably yes ("independent" means that the investigator generating the sequence is different from the investigator enrolling and assigning participants). Four out of the first five reviewers had problem understanding "independent". This led to mistakes in assessment (in the two trials where this was an issue, they rated as definitely yes when it should have been probably yes and rated as probably yes when it should have been definitely yes).	1. The panel decided to omit the independent issue. Thus, in the latest instrument, central allocation and pharmacy-controlled randomization should be assessed as definitely yes.
	2. For the envelopes approach, two out of the second five reviewers	2. We made the difference in italics.

	assessed wrongly because they ignored the difference between "sequentially numbered opaque sealed envelopes with evidence that they were opened sequentially" in definitely yes and "sequentially numbered, opaque, sealed envelopes without further details" in probably yes.	
	3. Two out of the second five reviewers thought the difference between probably yes and probably no was not very clear (probably yes: "Drug trial in which participants and healthcare providers were blinded..."; probably no: "Unblinded drug trial or non-drug trial...")	3. We made the difference bold.
Item 3: Were participants blinded	No problem found.	No change.
Item 4: Were healthcare providers blinded	No problem found.	No change.
Item 5: Were outcome assessors blinded	Two out the first five reviewers assessed wrongly because they didn't understand that, when the participant completed a questionnaire, participants were the outcome assessors.	We added "When the outcome is participants self-report, e.g., completing a questionnaire (such as quality of life, disability index), participants are the outcome assessors".
Item 6: Section 1. Extract the number of participants who were not included in analysis in each group Section 2. Was the overall percentage of participants not included in analysis acceptably low	1. In the instrument used in the first five interviews, the initial item 6 was: Section 1. Extract number of participants with missing outcome data in each group Section 2. Did the trial achieve an acceptably low percentage of participants with missing outcome data The initial item 7 was: For participants whose outcome data were available, were all or almost all of them analyzed in the group to which they were randomized	1. The panel split the initial item 7 into two issues: per-protocol analysis and as-treated analysis. The panel combined the per-protocol analysis issue with the initial item 6 together, as the current item 6. Thus, the current item 6 addresses the participants who were not included in analysis irrespective of the reasons. The following user testing suggested the new item works well. The panel added the as-treated analysis issue as an optional item (optional item 6) since it happens rarely.

	Four out of the first five reviewers, when they assessing different trials, assessed wrongly for the initial item 7 - they had difficulty differentiating the participants who were left out of analysis because of decisions by the trialists to not include non-compliant individuals (i.e., should be addressed in initial item 7), and the participants cannot be included in analysis because their outcome data were not available to the trialists (should be addressed in initial item 6).	
	2. Reviewers had difficulty getting the number of participants not included in analysis correctly for time-to-event outcomes.	2. We added "For time-to-event outcomes, also count the participants who were censored because of missing follow-up data in 'N not analyzed'."
Overall experience in applying version A (ultimate first step)	All reviewers thought items and instructions in version A are clear and easy to apply.	No change.
Instrument version B (Ultimate second step)		
Item 1: Judge risk of bias related to sequence generation	Same as version A.	Same as version A.
Item 2: Judge risk of bias related to allocation concealment	Same as version A.	Same as version A.
Item 3: Judge risk of bias related to blinding of participants	Reviewer suggested bolding the adverbs of degree: very unlikely, unlikely, likely and very likely.	We made the words very unlikely, unlikely, likely and very likely bold.
Item 4: Judge risk of bias related to blinding of healthcare	No problem found.	No change.

providers		
Item 5: Judge risk of bias related to blinding of outcome assessors	No problem found.	No change.
Item 6: Section 1. Extract the number of participants who were not included in analysis in each group Section 2. Judge risk of bias related to the overall percentage of participants not included in analysis	Same as version A.	Same as version A.
Overall experience in applying version B (ultimate second step)	All reviewers thought items and instructions in version B are clear. Four reviewers thought version B makes more sense because they have get used to assess risk of bias directly.	No change.

Table 4. Characteristics of the systematic review experts in the user-testing

Characteristic	Number of participants (total 8)
Female	5
Country	
Canada	2
China	2
Germany	2
Australia	1
UK	1
Clinical background	
Physician	4
Nurse	1
No clinical background	3

Table 5. Summary of feedback from user testing with systematic review experts and resulting changes

Document	Feedback or problem found in the user testing	Resulting changes
Instrument	Participants questioned the rationale for using only version A. One person suggested to make the two instrument versions as two steps for assessing risk of bias: first step is evaluating what happened, and second step is judging risk of bias based on what happened.	We changed to adopt the two-step approach: for each core item, there are two steps: first step is evaluating what happened, and second step is judging risk of bias based on what happened.
	One participant suggested to add in each item a space for "support for judgement".	We added a space for "support for judgement" for each item.
	One participant suggested to add a space for systematic reviewers to add optional items in the instrument.	We added a space for optional items in the last page of the instrument.
	Tow participants suggested to think of a way to let systematic reviewers enter their risk of bias assessment results and generate a risk of bias assessment table for all trials in the systematic review which can be put in the systematic review manuscripts.	We developed an excel where systematic reviewers can enter their risk of bias assessment results and thus generate a risk of bias assessment table for all trials in the systematic review (appendix 5).
Manual	One participant suggested to rewrite the titles for optional items to make them easier to understand. For instance, optional item 1 "Baseline prognostic factors balance vs. imbalance", change to "Whether the baseline prognostic factors balanced between groups ".	We rewrote the titles for optional items.
	One participant suggested to add some examples for behavioral or psychological trials in the manual.	We added examples for behavioral and psychological trials.
	One participant suggested to change the sentence "Trials usually do not report number of participants who were not included in analysis for individual outcomes" to "Trials may not report number of participants who were not included in analysis for individual outcomes".	We changed to sentence to "Trials may not report number of participants who were not included in analysis for individual outcomes".

Appendix 3. Panel's initial decisions on item selection and format of the instrument

Table 1. Judgment regarding the extent to which items in category 1 (items that the majority judged as addressing risk of bias in a survey) meet the six item selection criteria and item selection decisions

Items	Item selection criteria						Item selection decision
	Clearly risk of bias rather than others	Theoretical or logical argument for why the item is important	Information required to make judgement is commonly reported in trials	Non-expert reviewers can make the judgment easily	Problem occurs more often than rarely	Empirical evidence supports item influence on effect estimates	
Random sequence generation	√ (category 1)	√	√	√	√	Overestimation (Moderate certainty)	**Include as core item 1**
Allocation concealment	√ (category 1)	√	√	√	√	Overestimation (Moderate certainty)	**Include as core item 2**
Blinding of participants	√ (category 1)	√	√	√	√	Any outcomes: Very uncertain Patient-reported outcomes: Overestimation (Moderate certainty) Observer-reported or objective outcomes: Very uncertain	**Include as core item 3**
Blinding of healthcare providers	√ (category 1)	√	√	√	√	Very uncertain	**Include as core item 4**
Blinding of data collectors	√ (category 1)	√	×	×	√	Very uncertain	**Exclude**
Blinding of outcome assessors	√ (category 1)	√	√	√	√	Any outcomes: Very uncertain Objective outcomes: Very uncertain Subjective outcomes: Overestimation (High certainty)	**Include as core item 5**
Blinding of data analysts	√ (category 1)	√	×	×	√	Very uncertain	**Exclude**
Whether the outcome assessment or data collection differed between groups	√ (category 1)	√	×	×	×	No evidence	**Include as optional item 3**
Whether the follow-up time, frequency, or intensity of outcome assessment differed between groups	√ (category 1)	√	×	×	×	No evidence	**Include as optional item 4**
Missing outcome data	√ (category 1)	√	√	√	√	Underestimation (Low certainty)	**Initially include as a core item. Afterward, based on use**

							testing feedback, the panel combined the missing outcome data issue and per-protocol analysis issue together as a new core item (core item 6).

Table 2. Decisions for items in category 2 (items that the majority judged as not addressing risk of bias in a survey)*

Items	Item selection decision	Reasons for excluding
Whether the outcome measurement was reliable (i.e., reliability of outcome measurement)	**Exclude**	Imprecision (random error) rather than risk of bias.
Whether the follow-up time was adequate to identify the outcome of interest	**Exclude**	Indirectness (applicability) rather than risk of bias.
Whether the sample size was big enough	**Exclude**	Imprecision (random error) rather than risk of bias.
Whether the sampling approach (approach to selecting participants from whole population) was appropriate	**Exclude**	Indirectness (applicability) rather than risk of bias.
Whether there was conflict of interest	**Exclude**	Not directly related to risk of bias – may influence effect estimate but must through other mechanisms.
Whether there was funding	**Exclude**	Not directly related to risk of bias – may influence effect estimate but must through other mechanisms.
Whether the results were comparable for all trial sites	**Exclude**	Not related to risk of bias.
Whether there was run-in period before randomization	**Exclude**	Not related to risk of bias.
Whether the inclusion and exclusion criteria were appropriate	**Exclude**	Indirectness (applicability) rather than risk of bias.

*All items in category 2 were excluded.

Table 3. Judgment regarding the extent to which items in category 3 (items with substantial disagreement regarding addressing risk of bias or not) meet the six item selection criteria and item selection decisions

Items	Item selection criteria						Item selection decision
	Clearly risk of bias rather than others	Theoretical or logical argument for why the item is important	Information required to make judgement is commonly reported in trials	Non-expert reviewers can make the judgment easily	Problem occurs more often than rarely	Empirical evidence supports item influence on effect estimates	
Whether the baseline prognostic factors were balanced between groups	×	√	√	×	?	Very uncertain	**Include as optional item 1**
Whether the co-interventions were balanced between groups in blinded trials	×	√	×	×	?	Overestimation (Low certainty)	**Include as optional item 2**
Whether the interventions were implemented as intended	×	√	×	×	√	Very uncertain	**Exclude**
Whether the participants were adhered to the assigned interventions (nonadherence could be imperfect compliance, cessation, crossover, or switch to another active intervention)	×	×	×	×	√	Very uncertain	**Exclude**
Whether the outcome measurement method was valid (i.e., validity of outcome measurement)	×	√	×	×	×	No evidence	**Include as optional item 5**
Whether there was non-differential outcome measurement error (errors that were unrelated to intervention assignment)	×	√	×	×	×	No evidence	**Exclude**
Whether the method of outcome measurement or data collection was	×	×	×	×	×	No evidence	**Exclude**

sensitive to plausible intervention effects							
Whether there was selective reporting	×	√	×	×	×	Very uncertain	**Include as optional item 7**
Intention-to-treat analysis	×	√	×	√	√	Very uncertain	**Initially include as a core item. Afterward, based on use testing feedback, the panel split this item into two types of deviation: per-protocol analysis and as-treated analysis. The panel combined the per-protocol analysis issue with the missing outcome data issue together as a new core item (core item 6). The panel added the as-treated analysis issue as optional item 6: "when investigators conducted an as-treated analysis, was the percentage of participants not analyzed in the groups to which they were randomized sufficiently low".**
Whether the trial was terminated early	×	√	√	√	√	Stop early for benefit: Overestimation (Moderate certainty)	**Include as optional item 8. The panel rewrote the optional item as "whether the trial was terminated early for benefit". Although trials that terminated early for futility may underestimate effect, the underestimation was much less dramatic. As for trials terminated early for harm, although they would overestimate the harm, if a trial showed an early signal of harm with the intervention it may be unethical to enroll more patients into the intervention group.**

Table 4. Initial instrument version A (finally changed to the first step for assessing risk of bias)

Task of systematic reviewers who assess individual trials (i.e., front-line systematic reviewers)	Evaluate what happened for each item Item 1: Was the allocation sequence adequately generated Item 2: Was the allocation adequately concealed Item 3: Were participants blinded Item 4: Were healthcare providers blinded Item 5: Were outcome assessors blinded Item 6 - Section 1 (necessary): Extract the number of participants who were not included in analysis in each group - Section 2 (optional): Was the overall percentage of participants not included in analysis acceptably low
Response options	Definitely Yes; Probably Yes; Probably No; Definitely No (except for item 6 section 1)
Who make risk of bias judgement	Not front-line systematic reviewers; rather, systematic review team members who assess GRADE certainty of evidence across trials (usually review leaders)

Table 5. Initial instrument version B (finally changed to the second step for assessing risk of bias)

Task of systematic reviewers who assess individual trials (i.e., front-line systematic reviewers)	Judge risk of bias related to each item Item 1: Judge risk of bias related to sequence generation Item 2: Judge risk of bias related to allocation concealment Item 3: Judge risk of bias related to blinding of participants Item 4: Judge risk of bias related to blinding of healthcare providers Item 5: Judge risk of bias related to blinding of outcome assessors Item 6 - Section 1 (necessary): Extract the number of participants who were not included in analysis in each group - Section 2 (optional): Judge risk of bias related to the overall percentage of participants not included in analysis
Response options	Definitely Low; Probably Low; Probably High; Definitely High (except for item 6 section 1)
Who make risk of bias judgement	Front-line systematic reviewers

Appendix 4 (PDF) and 5 (Word). ROBUST-RCT

Risk of Bias Instrument for Use in Systematic Reviews – for Randomized Controlled Trials (ROBUST-RCT)

Study reference:

State the outcome(s) that are being assessed for risk of bias:

Copyright © 2024 McMaster University, Hamilton, Ontario, Canada. Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT), authored by Wang et al, is the copyright of McMaster University (Copyright ©2024, McMaster University). The Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT) must not be copied, distributed, or used in any way without the prior written consent of McMaster University. Contact the McMaster Industry Liaison Office at McMaster University, email: milo@mcmaster.ca for licensing details.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Item 1: Random sequence generation

Step 1: Was the allocation sequence adequately generated*	Step 2: Judge risk of bias related to sequence generation*	Instructions
☐ Definitely Yes	☐ Definitely Low	Trial explicitly stated use of an adequate method of generating the random allocation sequence. Examples include: random number table; random number generator; throwing dice; drawing of lots; minimization.
☐ Probably Yes	☐ Probably Low	Trial was described as "randomized" without further details regarding the method of generating the allocation sequence, and: - the text mentioned simple randomization, block randomization, or stratified randomization; or - the text described the allocation concealment method (stated using central allocation, drug containers, or envelopes).
☐ Probably No	☐ Probably High	Trial was described as "randomized" without further details regarding the method of generating the allocation sequence, and it does not meet any of the criteria for "Probably Yes/Low".
☐ Definitely No	☐ Definitely High	Trial used a non-randomized allocation sequence that would be recognized as "quasi-randomization". Examples include: allocation based on dates of birth or admission; patient's hospital record number; alteration or rotation; allocation decided by clinicians or participants; allocation based on the results of a laboratory test; allocation by availability of the intervention.

*Instructions for step 1 and step 2 are the same because no additional judgement is involved.

Support for judgement:

Item 2: Allocation concealment

Step 1: Was the allocation adequately concealed*	Step 2: Judge risk of bias related to allocation concealment*	Instructions
☐ **Definitely Yes**	☐ **Definitely Low**	Trial used a clearly satisfactory allocation concealment method: central allocation (e.g., telephone, web-based); pharmacy-controlled randomization (stated using sequentially numbered sealed drug containers); trial explicitly stated using sequentially numbered opaque sealed envelopes *with evidence that they were opened sequentially*.
☐ **Probably Yes**	☐ **Probably Low**	• Trial explicitly stated using sequentially numbered, opaque, sealed envelopes *without further details*. • *Drug trial in which participants and healthcare providers were blinded*: - with no further information on the allocation concealment method; or - stated using envelopes but it remains unclear whether the envelopes were sequentially numbered opaque and sealed; or - stated using drug containers but it remains unclear whether the drug containers were sequentially numbered and sealed.
☐ **Probably No**	☐ **Probably High**	• *Unblinded drug trial or non-drug trial*: - with no further information on the allocation concealment method; or - stated using envelopes but it remains unclear whether the envelopes were sequentially numbered opaque and sealed; or - stated using drug containers but it remains unclear whether the drug containers were sequentially numbered and sealed.
☐ **Definitely No**	☐ **Definitely High**	• Trial used an open random allocation schedule. • Trial used a non-randomized allocation sequence that would be recognized as "quasi-randomization".

*Instructions for step 1 and step 2 are the same because no additional judgement is involved.

Support for judgement:

Item 3: Blinding of participants

Step 1: Were participants blinded

☐ **Definitely Yes**	Trial explicitly stated that participants were blinded.
☐ **Probably Yes**	No explicit statement about blinding of participants, and: - it is a placebo-controlled drug trial; or - it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or - trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgment is that one of the blinded groups is the participants; or - participants not capable of distinguishing if they are receiving active or control intervention (e.g., neonates, severely demented).
☐ **Probably No**	No explicit statement about blinding of participants, and: - it is an active control drug trial (A vs. B) and no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the participants.
☐ **Definitely No**	• Trial explicitly stated that participants were not blinded. • Trial was described as "open-label" or "unblinded".
Support for judgement:	

Step 2: Judge risk of bias related to blinding of participants

Issues to consider:

i) Were participants blinded (step 1)

ii) If unblinded, how likely unblinding of participants has influenced the outcome

- *How likely are participants expectations regarding effect of intervention to have influenced the outcome*
- *How likely are participant-initiated co-interventions to have influenced the outcome*

☐ **Definitely Low**	• Participants were definitely blinded; OR • Unblinding of participants *very unlikely* to have influenced the outcome because: very unlikely participants expectations regarding effect of intervention have influenced the outcome and very unlikely participant-initiated co-interventions have influenced the outcome.
☐ **Probably Low**	• Participants were probably blinded; OR • Unblinding of participants *unlikely* to have influenced the outcome because: unlikely participants expectations regarding effect of intervention have influenced the outcome and unlikely participant-initiated co-interventions have influenced the outcome.
☐ **Probably High**	Participants were definitely or probably not blinded, AND unblinding of participants *likely* to have influenced the outcome because participants expectations regarding effect of intervention likely to have influenced the outcome or participant-initiated co-interventions likely to have influenced the outcome.
☐ **Definitely High**	Participants were definitely or probably not blinded, AND unblinding of participants *very likely* to have influenced the outcome through participants expectations regarding effect of intervention or through participant-initiated co-interventions.
Support for judgement:	

Item 4: Blinding of healthcare providers

Step 1: Were healthcare providers blinded

☐ **Definitely Yes**	Trial explicitly stated that healthcare providers were blinded.
☐ **Probably Yes**	No explicit statement about blinding of healthcare providers, and: - it is a placebo-controlled drug trial; or - it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or - trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgment is that one of the blinded groups is the healthcare providers.
☐ **Probably No**	No explicit statement about blinding of healthcare providers, and: - it is an active control drug trial (A vs. B) but no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the healthcare providers.
☐ **Definitely No**	• Trial explicitly stated that healthcare providers were not blinded. • Trial was described as "open-label" or "unblinded".
Support for judgement:	

Step 2: Judge risk of bias related to blinding of healthcare providers

Issues to consider:

i) Were healthcare providers blinded (step 1)

ii) *If unblinded, how likely unblinding of healthcare providers has influenced the outcome*

- *How likely healthcare provider-initiated co-interventions have influenced the outcome*

☐ **Definitely Low**	Healthcare providers were definitely blinded.
☐ **Probably Low**	• Healthcare providers were probably blinded; OR • Unblinding of healthcare providers unlikely to have influenced the outcome because: - Unlikely there is any healthcare provider-initiated co-intervention that could potentially influence the outcome; or - Investigators have documented all healthcare provider-initiated co-interventions that could potentially influence the outcome and demonstrated similarity in use of all these co-interventions between groups.
☐ **Probably High**	• Healthcare providers were definitely or probably not blinded, AND unblinding of healthcare providers likely to have influenced the outcome because there are healthcare provider-initiated co-interventions that could potentially influence the outcome.
☐ **Definitely High**	• Healthcare providers were definitely or probably not blinded, AND unblinding of healthcare providers very likely to have influenced the outcome because there are healthcare provider-initiated co-interventions that could potentially influence the outcome and investigators have documented dissimilarity in any of these co-interventions between groups.
Support for judgement:	

Item 5: Blinding of outcome assessors

Step 1: Were outcome assessors blinded

When the outcome is participants self-report, e.g., completing a questionnaire (such as quality of life, disability index), participants are the outcome assessors.

☐ **Definitely Yes**	Trial explicitly stated that the outcome assessors or adjudicators (people making the measurement or assessment) were blinded.
☐ **Probably Yes**	No explicit statement about blinding of outcome assessors or adjudicators, and: - it is a placebo-controlled drug trial; or - it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or - trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgement is that one of the blinded groups is the outcome assessors.
☐ **Probably No**	No explicit statement about blinding of outcome assessors or adjudicators, and: - it is an active control drug trial (A vs. B) but no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the outcome assessors.
☐ **Definitely No**	• Trial explicitly stated that the outcome assessors or adjudicators were not blinded. • Trial was described as "open-label" or "unblinded".
Support for judgement:	

Step 2: Judge risk of bias related to blinding of outcome assessors

Issues to consider:

i) Were outcome assessors blinded (step 1)

ii) If unblinded, how likely unblinding of outcome assessors has influenced the outcome assessment

- *Degree of judgement/subjectivity involved in the outcome assessment (more judgment/subjectivity more likelihood of bias)*

☐ **Definitely Low**	• Outcome assessors were definitely blinded; OR • Outcome is all-cause mortality.
☐ **Probably Low**	• Outcome assessors were probably blinded; OR • Unblinding of outcome assessors likely could not have influenced the outcome assessment because the outcome assessment involves minimal judgement (other objective outcomes, e.g., laboratory measurement, hospital admission, mechanical ventilation).
☐ **Probably High**	Outcome assessors were definitely or probably not blinded, AND unblinding of outcome assessors likely could have influenced the outcome assessment because the outcome assessment involves some judgment (e.g., cause-specific mortality).
☐ **Definitely High**	Outcome assessors were definitely or probably not blinded, AND unblinding of outcome assessors could have influenced the outcome assessment because the outcome assessment involves considerable judgment by participant or adjudicator (e.g., symptoms and symptom scores, quality of life, seizure occurrence).
Support for judgement:	

Item 6: Outcome data not included in analysis

Step 1: Extract the number of participants who were not included in analysis in each group

Note: For time-to-event outcomes, also count the participants who were censored because of missing follow-up data in 'N not analyzed'.

	N not analyzed	N total (usually N randomized)	Percentage not analyzed (N not analyzed / N total)
Intervention group			
Control group			
Overall			

Step 2: Judge risk of bias related to the overall percentage of participants not included in analysis

Issue to consider: Was the overall percentage of participants not included in analysis acceptably low

Note: Systematic review teams need to set and fill in the threshold for each response option. See manual for instructions and example thresholds.

☐ **Definitely Low**	Percentage of participants not included in analysis is < %
☐ **Probably Low**	• Percentage of participants not included in analysis is % to < % • If the trial did not mention whether there were participants not included in analysis, a substantial loss to follow-up is unlikely (e.g., ICU mortality)
☐ **Probably High**	• Percentage of participants not included in analysis is % to < % • If the trial did not mention whether there were participants not included in analysis, a substantial loss to follow-up is likely (e.g., 1-year quality of life)
☐ **Definitely High**	Percentage of participants not included in analysis is ≥ %
Support for judgement:	

Which if any optional item(s) to assess

We provide eight optional items (see manual).

We strongly recommend that the systematic review teams consider each of the optional items carefully and judge whether they are relevant to their particular reviews.

If systematic review teams decide to include optional item(s), specify the optional items that need to be considered and provide instructions for assessing the items (see examples in manual).

Optional item	Instruction	Judgement	Support for judgement

Appendix 6. ROBUST-RCT Excel (see the manuscript for details)

Appendix 7. Manual

Risk of Bias Instrument for Use in Systematic Reviews – for Randomized Controlled Trials (ROBUST-RCT)

Manual

Table of contents

Two steps for assessing risk of bias 2
State the outcome(s) that are being assessed for risk of bias 4
Item 1 Random sequence generation 6
Item 2 Allocation concealment 8
Item 3 Blinding of participants 11
Step 1: Were participants blinded 11
Step 2: Judge risk of bias related to blinding of participants 12
Item 4 Blinding of healthcare providers 15
Step 1: Were healthcare providers blinded 15
Step 2: Judge risk of bias related to blinding of healthcare providers 16
Item 5 Blinding of outcome assessors 19
Step 1: Were outcome assessors blinded 19
Step 2: Judge risk of bias related to blinding of outcome assessors 20
Item 6 Outcome data not included in analysis 23
Two approaches for assessing risk of bias related to item 6 23
Set thresholds for step 2 23
Specify post-randomization exclusion without bias (optional) 24
Manual for completing step 1 25
Manual for completing step 2 26
Decide which if any optional items to include 28
References 35

Copyright © 2024 McMaster University, Hamilton, Ontario, Canada. Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT), authored by Wang et al, is the copyright of McMaster University (Copyright ©2024, McMaster University). The Risk of Bias Instrument for Use in Systematic Reviews - for Randomized Controlled Trials (ROBUST-RCT) must not be copied, distributed, or used in any way without the prior written consent of McMaster University. Contact the McMaster Industry Liaison Office at McMaster University, email: milo@mcmaster.ca for licensing details.
This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

Two steps for assessing risk of bias

ROBUST-RCT suggests two steps for assessing risk of bias for each core item.

The first step is evaluating what happened, that is, whether the methodological safeguard has been implemented (e.g., step 1 of item 3 is judging if participants were blinded). For all but the last item, response options include definitely yes, probably yes, probably no, and definitely no.

The second step is judging risk of bias based on what happened (e.g., step 2 of item 3 is judging risk of bias related to blinding of participants). It requires systematic review team members to decide the extent to which any deficits in instituting methodological safeguards actually resulted in risk of bias. Response options include definitely low, probably low, probably high, and definitely high risk of bias. Table 1 presents the two steps for each core item.

Systematic reviewers assessing individual trials (i.e., risk of bias assessors) can complete both steps. However, for core items 1-5, if the risk of bias assessors that the systematic review team recruits are less experienced and may face difficulty in judging risk of bias, review leaders may ask the risk of bias assessors to complete only step 1 and leave step 2 to the reviewers with more experience.

For core item 6 'outcome data not included in analysis', there are two approaches to addressing risk of bias. One is deciding the risk of bias associated with this item for each individual trial. In this case, systematic review teams will need to set the missing percentage threshold for each response option for the step 2 of item 6 (see instructions for setting thresholds in page 23). Risk of bias assessors will determine the percentage of people not included in analysis and where that percentage falls in the risk of bias categories.

Another approach for core item 6 involves systematic review teams assessing risk of bias associated with missing data across the entire body of evidence at the meta-analysis level.[1 2] The process of doing so begins with a complete-case analysis followed by an analysis imputing data for participants in each trial who were not included in the analysis (see details in page 23).[1 2] In this case, risk of bias assessors can complete only step 1 in which they extract the number of participants who were not included in analysis.

Table 1. ROBUST-RCT core items and two steps

	Step 1 Evaluate what happened	Step 2 Judge risk of bias
Item 1 Random sequence generation	Was the allocation sequence adequately generated	Judge risk of bias related to sequence generation
Item 2 Allocation concealment	Was the allocation adequately concealed	Judge risk of bias related to allocation concealment
Item 3 Blinding of participants	Were participants blinded	Judge risk of bias related to blinding of participants
Item 4 Blinding of healthcare providers	Were healthcare providers blinded	Judge risk of bias related to blinding of healthcare providers
Item 5 Blinding of outcome assessors	Were outcome assessors blinded	Judge risk of bias related to blinding of outcome assessors
Item 6 Outcome data not included in analysis	Extract the number of participants who were not included in analysis in each group	Judge risk of bias related to the overall percentage of participants not included in analysis
Response options	Definitely Yes; Probably Yes; Probably No; Definitely No (except for item 6)	Definitely Low; Probably Low; Probably High; Definitely High

State the outcome(s) that are being assessed for risk of bias

Systematic review teams need to specify the outcome(s) that are being assessed for risk of bias (fill in the first page of the instrument).

Study reference:

State the outcome(s) that are being assessed for risk of bias:

Systematic review teams need to consider whether to direct systematic reviewers to make one assessment for all outcomes, a group of outcomes, or for each outcome separately (Table 2).

Table 2. Considerations regarding assessing all outcomes, a group of outcomes, or each outcome separately

Items	Considerations
Item 1 Random sequence generation Item 2 Allocation concealment	Same for all outcomes; can assess once for all outcomes.
Item 3 Blinding of participants Item 4 Blinding of healthcare providers	For step 1 judging if participants/healthcare providers were blinded, judgement is likely to be the same for all outcomes; if the systematic reviewers assessing individual trials need to complete only step 1, they could assess once for all outcomes. If the systematic reviewers assessing individual trials need to complete both steps, the review team should consider for which outcomes extent of risk of bias is likely to differ. When this is likely, the safest approach is to ask systematic reviewers to rate each outcome separately.
Item 5 Blinding of outcome assessors	If the systematic reviewers assessing individual trials need to complete only step 1, the review team should consider whether the outcome assessors were the same for different outcomes. If not, they should consider whether it is plausible that blinding may differ across the different outcome assessors. If that is likely, the review team may choose to assess each outcome separately. If the systematic reviewers assessing individual trials need to complete both steps, the review team should consider for which outcomes extent of risk of bias is likely to differ. When this is likely, the safest approach is to ask systematic reviewers to rate each outcome separately.
Item 6 Outcome data not included in analysis	Trials may not report number of participants who were not included in analysis for individual outcomes. In such instances, rating for item 6 will likely be identical for different outcomes. The review team should, however, consider the possibility that trialists may have reported participants not included in analysis separately for each outcome, in which case item 6 should be rated separately. The

	review leaders may want to check this themselves and then specify for risk of bias assessors which outcomes should be assessed separately, or direct risk of bias assessors to be themselves alert to these issues. They may want to consider the possibility that although loss to follow-up is specified for only a single outcome (e.g., mortality), loss to follow-up although unspecified may be greater for another outcome (e.g., quality of life).

Item 1 Random sequence generation

Step 1: Was the allocation sequence adequately generated
Step 2: Judge risk of bias related to sequence generation
Instructions for step 1 and step 2 are the same.

Definitely Yes/Low
Trial explicitly stated use of an adequate method of generating the random allocation sequence. Examples include: random number table; random number generator; throwing dice; drawing of lots; minimization.
Explanation:
Adequate method of generating the random allocation sequence refers to the method that incorporates a random element and thus can generate a random and unpredictable sequence. *Minimization* is a method of ensuring intervention groups are closely similar for multiple prognostic factors, even in small trials.[3] Using minimization, the first participant is allocated randomly. For each subsequent participant, investigators determine assignment to the intervention that would lead to better balance between the groups over all the identified prognostic factors.
Example: A trial stated "Randomization was performed with a computer-generated allocation sequence…".[4]

Probably Yes/Low
Trial was described as "randomized" without further details regarding the method of generating the allocation sequence, and:
- **the text mentioned simple randomization, block randomization, or stratified randomization; or**
- **the text described the allocation concealment method (stated using central allocation, drug containers, or envelopes).**

Explanation:
Sometimes, in trial reports, authors described the trials as "randomized" or stated "randomly allocated". They did not, however, explicitly report the method they used for generating the allocation sequence (i.e., did not specify any of the methods in 'Definitely Yes/Low' or 'Definitely No/High').

Simple randomization (unrestricted randomization) means allocating each participant at random independently with no constraints.[5] If a trial did not specify the method trialists used for generating the allocation sequence, but it reported that it adopted simple or unrestricted randomization approach, then trialists probably have used an adequate method of generating the random sequence.

Block randomization (one type of restricted randomization) is used when trialists want to ensure that the numbers of participants in intervention and control groups reach a particular ratio (e.g., 1:1).[5] If a trial reported that it adopted block randomization or mentioned blocking, although it did not specify the method used for generating the allocation sequence for each block, it is likely that trialists used an adequate method of generating the random sequence.

Stratified randomization is used when trialists want to ensure that the participants in intervention and control groups are balanced regarding a particular prognostic factor.[5] If a trial did not specify the method trialists used for generating the allocation sequence, but it reported that it adopted stratified randomization, then trialists probably have used an adequate method of generating the random sequence.

Allocation concealment method: If authors described the trial as "randomized" or stated "randomly allocated", and reported using central allocation, pharmacy-controlled randomization (or using drug containers), or envelopes, trialists probably have generated the sequence adequately.
Example:

1. A trial stated "Patients were randomly allocated into two groups using blocked randomization" but it did not report how they generated the random sequence.[6]
2. A trial stated "Randomization was by sealed envelope using the permuted block design" but it did not report how they generated the random sequence.[7]

Probably No/High

Trial was described as "randomized" without further details regarding the method of generating the allocation sequence, and it does not meet any of the criteria for "Probably Yes/Low".
Example: A trial stated "patients were randomly allocated" without further details regarding the sequence generation method, and it did not mention simple, block, or stratified randomization, and it did not describe the allocation concealment method.[8]

Definitely No/High

Trial used a non-randomized allocation sequence that would be recognized as "quasi-randomization". Examples include: allocation based on dates of birth or admission; patient's hospital record number; alteration or rotation; allocation decided by clinicians or participants; allocation based on the results of a laboratory test; allocation by availability of the intervention.
Explanation: *"Quasi-randomization"* methods are actually non-randomized allocation methods because they are not based on a random/chance process and they cannot generate an unpredictable allocation sequence.[9]
Example: A trial stated "Patients were randomized to each respective study arm by date of admission. Treatment regimens alternated every other month. Patients admitted into the unit during odd-numbered months received famotidine 20 mg IV every 12 h; patients admitted during even numbered months received lansoprazole suspension 30 mg suspended in 10 ml of an 8.4% sodium bicarbonate solution or apple juice, and administered via NG tube daily."[10]

Item 2 Allocation concealment

Step 1: Was the allocation adequately concealed
Step 2: Judge risk of bias related to allocation concealment
Instructions for step 1 and step 2 are the same.

Definitely Yes/Low
Trial used a clearly satisfactory allocation concealment method: central allocation (e.g., telephone, web-based); pharmacy-controlled randomization (stated using sequentially numbered sealed drug containers); trial explicitly stated using sequentially numbered opaque sealed envelopes with evidence that they were opened sequentially.
Explanation:

Allocation concealment means trialists take procedures to prevent participants and investigators involved in the enrollment and assignment from foreknowing or predicting the forthcoming allocations.

Central allocation (or remote allocation/randomization) means the allocation is controlled by a third party who is independent from the investigators enrolling participants. After participants' eligibility is determined, the investigator will inform the third party (e.g., by electronic communication, telephone, or enter participant's information into the system if web-based e.g. REDCap, or through Interactive Voice Response System) to perform the allocation.

Explicitly stated using sequentially numbered opaque sealed envelopes with evidence that they were opened sequentially: Concealment using envelopes is susceptible to manipulation.[11] Systematic reviewers can evaluate trials using envelopes as definitely concealed only when sufficient details support rigor of the approach. First, trials need to explicitly state that the envelopes were sequentially numbered, opaque and sealed. In addition, there is evidence that the envelopes were opened sequentially only after the envelope has been irreversibly assigned to the participant (e.g., participant's details were written on the outside of the envelope and transferred to the assignment card by carbon or pressure-sensitive paper inside the envelope).[12]
Example:

1. A trial stated "Copenhagen Trial Unit is responsible for centralized and web-based 1:1 randomization according to a computer-generated allocation sequence list".[4 13] This trial used central allocation ("centralized and web-based").
2. A trial stated "After providing informed consent, the participants were randomly allocated in a 2:1 ratio to either an LCHF or a HCLF diet using computer-generated (REDCap) random allocation sequence with permuted blocks of 6 and 9 stratified by sex and the number of antidiabetic drugs (<2 versus ≥2) to balance the groups according to disease severity and to avoid potential gender differences". REDCap includes a randomization module that helps trialists implementing the allocation process and achieving allocation concealment.[14]

Probably Yes/Low

- **Trial explicitly stated using sequentially numbered, opaque, sealed envelopes without further details.**
  Explanation: *Without further details* means no evidence supporting the envelopes were opened sequentially only after the envelope has been irreversibly assigned to the participant (e.g., participant's details were written on the outside of the envelope and transferred to the assignment card by carbon or pressure-sensitive paper inside the envelope).
  Example: "Treatment regimens were included in opaque sealed numbered envelopes and the

envelope with the lowest number was always used for the consecutive patient".[15]

- **Drug in which participants and healthcare providers were blinded, with no further information on the allocation concealment method OR stated using envelopes but it remains unclear whether the envelopes were sequentially numbered opaque and sealed OR stated using drug containers but it remains unclear whether the drug containers were sequentially numbered and sealed.**
  Explanation: Typically, in a drug trial in which participants and healthcare providers are blinded, the medication is prepared by the pharmacy and leaves the pharmacy in a container that specifies this is for a particular participant without any further designation; thus, it is virtually impossible that the allocation is not concealed. To meet this criterion, judgement has to be that the trials are definitely or probably blinded (blinding of participants and blinding of healthcare providers items are evaluated as 'Definitely Yes' or 'Probably Yes').
  Example: A trial reported as double-blind (participants and healthcare providers were probably blinded) compared pantoprazole or placebo for stress ulcer prophylaxis, with no further information on the allocation concealment method.[16]

In addition:

1. If a trial explicitly stated it is concealed, but did not report any detail about allocation concealment method, systematic reviewers should assess this item as 'Probably Yes' (Version A) or 'Probably Low' (Version B).
2. If a trial used a minimization approach to allocating participants, participants and the individual investigator who enrolls the participants are unlikely to know how the multiple prognostic factors are accruing, thus systematic reviewers should assess this item as 'Probably Yes' (Version A) or 'Probably Low' (Version B).

Probably No/High

Unblinded drug trial or non-drug trial, with no further information on the allocation concealment method OR stated using envelopes but it remains unclear whether the envelopes were sequentially numbered opaque and sealed OR stated using drug containers but it remains unclear whether the drug containers were sequentially numbered and sealed.

Example:

1. An unblinded trial compared stress ulcer prophylaxis with lansoprazole OD 30 mg once daily versus no prophylaxis, with no further information on the allocation concealment method.[6]
2. An open-label trial stated "randomization was performed by the research nurse by using previously prepared closed and opaque envelopes". It is unclear whether the envelopes were sequentially numbered.[17]

Definitely No/High

- **Trial used an open random allocation schedule.**
  Explanation: *Open random allocation schedule* means trialists did not attempt to conceal allocation or there is clear evidence that the concealment method was compromised. Examples include: the allocation schedule was posted on a bulletin board; it is clearly that the same investigator generated the sequence and enrolled and assigned participants; drug containers were clearly not sequentially numbered, not sealed, or not of identical appearance or weight; envelopes were clearly not sequentially numbered, not opaque or sealed, or not opened sequentially.
- **Trial used a non-randomized allocation sequence that would be recognized as "quasi-**

randomization".

Explanation: *"Quasi-randomization"* cannot generate an unpredictable allocation sequence. Examples include: allocation based on dates of birth or admission; patient's hospital record number; alteration or rotation; allocation decided by clinicians or participants; allocation based on the results of a laboratory test; allocation by availability of the intervention.

Example: "Patients were randomized to each respective study arm by date of admission. Treatment regimens alternated every other month. Patients admitted into the unit during odd-numbered months received famotidine 20 mg IV every 12 h; patients admitted during even numbered months received lansoprazole suspension 30 mg suspended in 10 ml of an 8.4% sodium bicarbonate solution or apple juice, and administered via NG tube daily."[10]

Item 3 Blinding of participants

Step 1: Were participants blinded

Definitely Yes

Trial explicitly stated that participants were blinded.

Explanation: Systematic reviewers need to identify who are the *participants*. Most trials include patients as participants (they may explicitly state patients were blinded), but in some trials, participants are healthy people.

Example: A trial stated "patients, families, clinicians, and research staff were blinded."[18]

Probably Yes

No explicit statement about blinding of participants, and:

- **it is a placebo-controlled drug trial; or**
- **it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or**
- **trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgment is that one of the blinded groups is the participants; or**
- **participants not capable of distinguishing if they are receiving active or control intervention (e.g., neonates, severely demented).**

Example: A trial, although with no explicit statement about blinding of participants, stated "we conducted a prospective randomized double-blind parallel-group study. Study participants were randomly assigned to receive pantoprazole (40 mg in 10 mL of 0.9% saline IV) or placebo (10 mL of 0.9% saline IV)".[16] Here, double-blind is ambiguous as to who was blinded, but this is a placebo-controlled drug trial and it seems extraordinarily unlikely that patients could distinguish between a bag of IV fluid that does or does not contain pantoprazole.

Probably No

No explicit statement about blinding of participants, and:

- **it is an active control drug trial (A vs. B) and no mention of "double dummy" or that medications were identical or matched; or**
- **it is a non-drug trial; or**
- **trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the participants.**

Example: A randomized trial compared the effect of tidal peritoneal dialysis versus continuous renal replacement therapy on critically ill patients with acute kidney injury.[19] Since there is no explicit statement about blinding of participants and this is a non-drug trial in which it would appear to be impossible to blind patients to peritoneal dialysis or continuous renal replacement therapy, participants were probably not blinded.

Definitely No

- **Trial explicitly stated that participants were not blinded.**
- **Trial was described as "open-label" or "unblinded".**

Example:

1. A trial stated "Face-to-face treatment meant it was not possible to blind participants or the physical therapists delivering the interventions".[20]
2. A trial stated "This was a prospective, randomized, open-label, multicenter study".[17] "Open

label" constitutes an explicit statement of no blinding.

Step 2: Judge risk of bias related to blinding of participants

Issues to consider:

i) Were participants blinded (step 1)

ii) If unblinded, how likely unblinding of participants has influenced the outcome

- **How likely are participants expectations regarding effect of intervention to have influenced the outcome**

  Explanation:

  Unblinding of participants may cause the participants in intervention and control groups to have different expectations regarding the effect of the intervention they received and preconceptions about their outcomes, which may affect their actual outcome.[21]

  To judge how likely are participants expectations regarding effect of intervention to have influenced participants' outcome, systematic reviewers should consider two issues. First, how likely are unblinding of participants in intervention and control groups to have different expectations regarding the effect of the intervention they received. This is more likely in trials comparing an active intervention versus an inactive control than in trials comparing two active interventions in which participants are unlikely to have expectations regarding which intervention is superior. Second, how likely is the outcome to be influenced by participants' expectations regarding effect of intervention – this is decided by the type of outcome.

- **How likely are participant-initiated co-interventions to have influenced the outcome**

  Explanation:

  Unblinding of participants may lead to differential participant-initiated co-interventions between groups, thus influencing participants' outcome.

  Participant-initiated co-interventions means any additional interventions that could potentially influence the outcome of interest that can be initiated by participants.

  To judge how likely are participant-initiated co-interventions to have influenced participants' outcome in a trial, systematic reviewers should consider two issues. First is the comparator in the trial. Trials comparing an active intervention versus an inactive control are more likely to have differential participant-initiated co-interventions than trials comparing two active interventions (participants in control group are more likely to seek other treatments when they know they receive an ineffective intervention or no treatment). Second, how easy was it for the participants to obtain co-interventions that had an appreciable impact on the outcome.

Definitely Low

- **Participants were definitely blinded.**

  Explanation: Trial explicitly stated that participants were blinded (systematic reviewers need to identify who are the participants).

  Example: A trial stated "patients, families, clinicians, and research staff were blinded."[18]

OR

- **Unblinding of participants very unlikely to have influenced the outcome because very**

unlikely participants expectations regarding effect of intervention have influenced the outcome and very unlikely participant-initiated co-interventions have influenced the outcome.

Example: An open randomized trial compared intrathoracic versus cervical anastomosis after totally or hybrid minimally invasive esophagectomy for esophageal cancer.[22] One outcome is in-hospital mortality. Although patients were unblinded (open-label), patients receiving intrathoracic and patients receiving cervical anastomosis were very unlikely to have different expectations regarding the effect of the intervention. Moreover, even if they did have different expectations, these were very unlikely to have influenced mortality. Finally, it is very likely there is no patient-initiated co-intervention that could influence mortality. Thus, for the outcome in-hospital mortality, there is 'Definitely Low' risk of bias related to blinding of participants.

Probably Low

- **Participants were probably blinded.**

  Explanation:

Probably Yes in Step 1	No explicit statement about blinding of participants, and: - it is a placebo-controlled drug trial; or - it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or - trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgment is that one of the blinded groups is the participants; or - participants not capable of distinguishing if they are receiving active or control intervention (e.g., neonates, severely demented).

Example: A trial, although with no explicit statement about blinding of participants, stated "we conducted a prospective randomized double-blind parallel-group study. Study participants were randomly assigned to receive pantoprazole (40 mg in 10 mL of 0.9% saline IV) or placebo (10 mL of 0.9% saline IV)".[16] Here, double-blind is ambiguous as to who was blinded, but this is a placebo-controlled drug trial and it seems extraordinarily unlikely that patients could distinguish between a bag of IV fluid that does or does not contain pantoprazole.

OR

- **Unblinding of participants unlikely to have influenced the outcome because unlikely participants expectations regarding effect of intervention have influenced the outcome and unlikely participant-initiated co-interventions have influenced the outcome.**

  Example: An open randomized trial compared intrathoracic versus cervical anastomosis after totally or hybrid minimally invasive esophagectomy for esophageal cancer.[22] One outcome is length of hospital stay. Although patients were unblinded, it is unlikely that patient's expectation regarding effect of intervention or patient-initiated co-interventions have influenced length of hospital stay. Thus, for the outcome length of hospital stay, there is 'Probably Low' risk of bias related to blinding of participants.

Probably High

Participants were definitely or probably not blinded, AND unblinding of participants likely to have influenced the outcome because participants expectations regarding effect of intervention likely to have influenced the outcome or participant-initiated co-interventions likely to have influenced the outcome.

Explanation:

Probably No in Step 1	No explicit statement about blinding of participants, and: - it is an active control drug trial (A vs. B) and no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the participants.
Definitely No in Step 1	• Trial explicitly stated that participants were not blinded. • Trial was described as "open-label" or "unblinded".

Example: In a randomized, open-label trial, investigators randomly allocated 80 adult patients with acute low back pain into two groups: 40 patients received ibuprofen 400 mg three times daily for three days; 40 patients received a fixed-dose combination tablet of ibuprofen 200 mg plus paracetamol 325 mg three times daily for three days.[23] This is an open-label trial so patients were definitely not blinded. Because one group received a single drug but another received combination therapy, patients in intervention and control groups were likely to have different expectations regarding the efficacy of intervention. Thus, for the outcome pain intensity, it could be assessed as 'Probably High' risk of bias related to blinding of participants.

Definitely High

Participants were definitely or probably not blinded, AND unblinding of participants very likely to have influenced the outcome through participants expectations regarding effect of intervention or through participant-initiated co-interventions.

Explanation:

Probably No in Step 1	No explicit statement about blinding of participants, and: - it is an active control drug trial (A vs. B) and no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the participants.
Definitely No in Step 1	• Trial explicitly stated that participants were not blinded; • Trial was described as "open-label" or "unblinded".

Example: A randomized, open-label trial compared melatonin and standard care versus standard care alone in hospitalized patients with COVID-19.[24] The primary outcome was sleep quality. This was an open-label trial, so patients were definitely not blinded. Sleep quality is likely to be influenced by patients' expectation regarding effect of intervention (patients receiving melatonin probably think melatonin could improve sleep quality and this belief may improve their sleep quality) or patient-initiated co-interventions (patients in control group were more likely to use hypnotics by themselves). Thus, for the outcome sleep quality, it could be assessed as 'Definitely High' risk of bias related to blinding of participants.

Item 4 Blinding of healthcare providers

Step 1: Were healthcare providers blinded

Definitely Yes

Trial explicitly stated that healthcare providers were blinded.

Explanation: *Healthcare providers* refer to the people who provide care and administer interventions. Some trials may not use the word "healthcare providers" explicitly, so systematic reviewers need to identify who were the healthcare providers in the trial. Most commonly they are physicians, nurses or members of other allied health professions.

Example: A trial stated "Physicians, bedside nurses and clinical pharmacists, other healthcare personnel, investigators, adjudicators, and the data analyst were blinded".[18]

Probably Yes

No explicit statement about blinding of healthcare providers, and:

- **it is a placebo-controlled drug trial; or**
- **it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or**
- **trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgment is that one of the blinded groups is the healthcare providers.**

Example: A trial, although with no explicit statement about blinding of healthcare providers, stated "we conducted a prospective randomized double-blind parallel-group study. Study participants were randomly assigned to receive pantoprazole (40 mg in 10 mL of 0.9% saline IV) or placebo (10 mL of 0.9% saline IV)".[16] Here, double-blind is ambiguous as to who was blinded, but this is a placebo-controlled drug trial and it seems extraordinarily unlikely that clinicians could distinguish between a bag of IV fluid that does or does not contain pantoprazole.

Probably No

No explicit statement about blinding of healthcare providers, and:

- **it is an active control drug trial (A vs. B) but no mention of "double dummy" or that medications were identical or matched; or**
- **it is a non-drug trial; or**
- **trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the healthcare providers.**

Example: A randomized trial compared the effect of tidal peritoneal dialysis versus continuous renal replacement therapy on critically ill patients with acute kidney injury.[19] Since there is no explicit statement about blinding of healthcare providers and this is a non-drug trial in which providers could easily distinguish whether patients are receiving peritoneal dialysis or continuous renal replacement therapy, healthcare providers were probably not blinded.

Definitely No

- **Trial explicitly stated that healthcare providers were not blinded.**
- **Trial was described as "open-label" or "unblinded".**

Example:

1. A trial stated "Face-to-face treatment meant it was not possible to blind participants or the physical therapists delivering the interventions".[20]
2. A trial stated "This was a prospective, randomized, open-label, multicenter study".[17] "Open

label" constitutes an explicit statement of no blinding.

Step 2: Judge risk of bias related to blinding of healthcare providers

Issues to consider:

i) Were healthcare providers blinded (step 1)

ii) If unblinded, how likely unblinding of healthcare providers has influenced the outcome

- **How likely healthcare provider-initiated co-interventions have influenced the outcome**

Definitely Low

Healthcare providers were definitely blinded.

Example: A trial stated "Physicians, bedside nurses and clinical pharmacists, other healthcare personnel, investigators, adjudicators, and the data analyst were blinded".[18]

Probably Low

- **Healthcare providers were probably blinded.**

  Explanation:

Probably Yes in Step 1	No explicit statement about blinding of healthcare providers, and: - it is a placebo-controlled drug trial; or - it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or - trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgment is that one of the blinded groups is the healthcare providers.

Example: A trial, although with no explicit statement about blinding of healthcare providers, stated "we conducted a prospective randomized double-blind parallel-group study. Study participants were randomly assigned to receive pantoprazole (40 mg in 10 mL of 0.9% saline IV) or placebo (10 mL of 0.9% saline IV)".[16] Here, double-blind is ambiguous as to who was blinded, but this is a placebo-controlled drug trial and it seems extraordinarily unlikely that clinicians could distinguish between a bag of IV fluid that does or does not contain pantoprazole.

OR

- **Unblinding of healthcare providers unlikely to have influenced the outcome because:**
  - **Unlikely there is any healthcare provider-initiated co-intervention that could potentially influence the outcome; or**
  - **Investigators have documented all healthcare provider-initiated co-interventions that could potentially influence the outcome and demonstrated similarity in use of all these co-interventions between groups.**

  Explanation: *Healthcare provider-initiated co-intervention that could potentially influence the outcome* means any additional intervention that could potentially influence the outcome of interest that can be initiated by healthcare providers, e.g., drug, advice, care, patient-healthcare provider interaction.[25]

  Example: Meningococcal serogroups A, B, C, W, and Y cause nearly all meningococcal disease. In a trial, investigators randomly allocated healthy individuals (age 10-25 years) to one group

receiving 0·5 mL of a MenABCWY vaccine (months 0 and 6) and placebo (month 0), or another group receiving MenB-FHbp vaccine (months 0 and 6) and MenACWY-CRM vaccine (month 0) via intramuscular injection into the upper deltoid.[26] Investigators compared the proportion of participants who achieved at least a four-fold increase in hSBA (serum bactericidal antibody using human complement) titers from baseline for each serogroup. Although the study staff administering the vaccine were unblinded, it is unlikely there are any healthcare provider-initiated co-interventions that could potentially influence the outcome, the risk of bias is 'Probably Low'.

Probably High

Healthcare providers were definitely or probably not blinded, AND unblinding of healthcare providers likely to have influenced the outcome because there are healthcare provider-initiated co-interventions that could potentially influence the outcome.

Explanation:

Probably No in Step 1	No explicit statement about blinding of healthcare providers, and: - it is an active control drug trial (A vs. B) but no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the healthcare providers.
Definitely No in Step 1	• Trial explicitly stated that healthcare providers were not blinded. • Trial was described as "open-label" or "unblinded".

Example: A randomized, open-label trial compared remdesivir and standard care versus standard care alone for the treatment of patients in hospital with COVID-19.[27] One outcome was duration of hospital stay. This is an open-label trial so healthcare providers were definitely not blinded. Unblinding of healthcare providers may possibly have influenced duration of hospital stay because duration of hospital stay is to a large extent under the control of the clinicians.

Definitely High

Healthcare providers were definitely or probably not blinded, AND unblinding of healthcare providers very likely to have influenced the outcome because there are healthcare provider-initiated co-interventions that could potentially influence the outcome and investigators have documented dissimilarity in any of these co-interventions between groups.

Explanation:

Probably No in Step 1	No explicit statement about blinding of healthcare providers, and: - it is an active control drug trial (A vs. B) but no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the healthcare providers.
Definitely No in Step 1	• Trial explicitly stated that healthcare providers were not blinded. • Trial was described as "open-label" or "unblinded".

Example: A randomized trial compared the effectiveness of Lopinavir/ Ritonavir/ Hydroxychloroquine (KH) versus Atazanavir/ Ritonavir/ Dolutegravir/ Hydroxychloroquine (ADH) in COVID-19 patients.[28] Healthcare providers were probably not blinded. Corticosteroid was an

important healthcare provider-initiated co-intervention that could influence the outcome mortality. The mortality rate in ADH group (3/32) was lower than KH group (6/30); while the proportion of corticosteroid administration in the ADH group (9/32) was higher than in the KH group (2/30) (p=0.03). The difference in corticosteroid was unlikely to be a function of the effect of interventions since the KH group with more mortality (worse outcome) received less corticosteroid. Thus, we conclude there is 'Definitely High' risk of bias related to blinding of healthcare providers.

Item 5 Blinding of outcome assessors

Step 1: Were outcome assessors blinded

Explanation: Outcome assessors can be different for different outcomes. Reviewers need to identify, for the outcome of interest, who are the outcome assessors. They could be participants (for participant-reported outcomes), healthcare providers (for healthcare provider-assessed outcomes), or independent assessors.

Definitely Yes

Trial explicitly stated that the outcome assessors or adjudicators (people making the measurement or assessment) were blinded.

Explanation: Reviewers should carefully consider who were the assessors or adjudicators for the outcome of interest, especially for trials with multiple outcomes. Consider, for instance, a trial with multiple outcomes explicitly stated "outcome assessors" were blinded; however, for a patient-reported outcome actually the outcome assessors were not blinded (because in this case patients were outcome assessors and patients were not blinded).

Example: A trial stated "Two clinicians blinded to allocation and to each other's assessments adjudicated all suspected clinical outcomes".[18]

Probably Yes

No explicit statement about blinding of outcome assessors or adjudicators, and:

- **it is a placebo-controlled drug trial; or**
- **it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or**
- **trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgement is that one of the blinded groups is the outcome assessors.**

Example: A trial, although with no explicit statement about blinding of outcome assessors, stated "we conducted a prospective randomized double-blind parallel-group study. Study participants were randomly assigned to receive pantoprazole (40 mg in 10 mL of 0.9% saline IV) or placebo (10 mL of 0.9% saline IV)".[16] For the outcome overt gastrointestinal bleeding, clinicians or nurses were the outcome assessors. Here, double-blind is ambiguous as to who was blinded, but this is a placebo-controlled drug trial and it seems extraordinarily unlikely that clinicians or nurses could distinguish between a bag of IV fluid that does or does not contain pantoprazole.

Probably No

No explicit statement about blinding of outcome assessors or adjudicators, and:

- **it is an active control drug trial (A vs. B) but no mention of "double dummy" or that medications were identical or matched; or**
- **it is a non-drug trial; or**
- **trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the outcome assessors.**

Example: A randomized trial compared the effect of tidal peritoneal dialysis versus continuous renal replacement therapy on critically ill patients with acute kidney injury.[19] For the outcome time to recovery of renal function, clinicians were the outcome assessors. Since there is no explicit statement about blinding of outcome assessors/clinicians and this is a non-drug trial (tidal peritoneal dialysis and continuous renal replacement therapy are easily distinguished), clinicians were probably not blinded.

Definitely No

- **Trial explicitly stated that the outcome assessors or adjudicators were not blinded.**
- **Trial was described as "open-label" or "unblinded".**

Example: "This was a prospective, randomized, open-label, multicenter study".[17]

Step 2: Judge risk of bias related to blinding of outcome assessors

Issues to consider:

i) **Were outcome assessors (people making the measurement or assessment) blinded**
Explanation: Outcome assessors can be different for different outcomes, thus evaluation of blinding of outcome assessors must consider the outcome. Reviewers need to identify, for the outcome of interest, who are the outcome assessors. They could be participants (for participant-reported outcomes), healthcare providers (for healthcare provider-assessed outcomes), or independent assessors.

ii) **If unblinded, how likely unblinding of outcome assessors has influenced the outcome assessment**
   - **Degree of judgement/subjectivity involved in the outcome assessment (more judgment/subjectivity more likelihood of bias)**

Definitely Low

- **Outcome assessors were definitely blinded.**
  Explanation: Trial explicitly stated that the outcome assessors or adjudicators (people making the measurement or assessment) were blinded.
  Example: A trial stated "Two clinicians blinded to allocation and to each other's assessments adjudicated all suspected clinical outcomes".[18]

OR

- **Outcome is all-cause mortality.**
  Explanation: Assessment of all-cause mortality involves no judgement.
  Example: A randomized open-label trial compared the effect of high-dose hemodiafiltration versus continuation of high-flux hemodialysis on patients with kidney failure.[29] For the outcome all-cause mortality, there is 'Definitely Low' risk of bias.

Probably Low

- **Outcome assessors were probably blinded.**
  Explanation:

Probably Yes in Step 1	No explicit statement about blinding of outcome assessors or adjudicators, and: - it is a placebo-controlled drug trial; or - it is an active control drug trial (A vs. B) and mention of "double dummy" or that medications were identical or matched; or - trial was described as "single blinded", "double blinded" or "triple blinded", and the best judgement is that one of the blinded groups is the outcome assessors.

Example: A trial, although with no explicit statement about blinding of outcome assessors,

stated "we conducted a prospective randomized double-blind parallel-group study. Study participants were randomly assigned to receive pantoprazole (40 mg in 10 mL of 0.9% saline IV) or placebo (10 mL of 0.9% saline IV)".[16] For the outcome overt gastrointestinal bleeding, clinicians or nurses were the outcome assessors. Here, double-blind is ambiguous as to who was blinded, but this is a placebo-controlled drug trial and it seems extraordinarily unlikely that clinicians or nurses could distinguish between a bag of IV fluid that does or does not contain pantoprazole.

OR

- **Unblinding of outcome assessors likely could not have influenced the outcome assessment because the outcome assessment involves minimal judgement (other objective outcomes, e.g., laboratory measurement, hospital admission, mechanical ventilation).**
  Example: A randomized trial compared the effect of tidal peritoneal dialysis versus continuous renal replacement therapy on critically ill patients with acute kidney injury.[19] For the outcome length of ICU stay, outcome assessors were probably not blinded (no explicit statement about blinding of outcome assessors and this is a non-drug trial). However, since assessment of length of ICU stay involves minimal judgement, there is 'Probably Low' risk of bias.

Probably High

Outcome assessors were definitely or probably not blinded, AND unblinding of outcome assessors likely could have influenced the outcome assessment because the outcome assessment involves some judgment (e.g., cause-specific mortality).

Explanation:

Probably No in Step 1	No explicit statement about blinding of outcome assessors or adjudicators, and: - it is an active control drug trial (A vs. B) but no mention of "double dummy" or that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the outcome assessors.
Definitely No in Step 1	• Trial explicitly stated that the outcome assessors or adjudicators were not blinded. • Trial was described as "open-label" or "unblinded".

Example: A randomized open-label trial compared the effect of high-dose hemofiltration versus continuation of high-flux hemodialysis on patients with kidney failure.[29] For the outcome death from cardiovascular causes, clinicians were the outcome assessors. Since outcome assessors were not blinded (open-label) and assessment of death from cardiovascular causes involves some judgement, there is 'Probably High' risk of bias.

Definitely High

Outcome assessors were definitely or probably not blinded, AND unblinding of outcome assessors could have influenced the outcome assessment because the outcome assessment involves considerable judgment by participant or adjudicator (e.g., symptoms and symptom scores, quality of life, seizure occurrence).

Explanation:

Probably No in Step 1	No explicit statement about blinding of outcome assessors or adjudicators, and: - it is an active control drug trial (A vs. B) but no mention of "double dummy" or

	that medications were identical or matched; or - it is a non-drug trial; or - trial was described as "single blinded" and the best judgement is that the single blinded group is someone other than the outcome assessors.
Definitely No in Step 1	• Trial explicitly stated that the outcome assessors or adjudicators were not blinded. • Trial was described as "open-label" or "unblinded".

Example: A randomized open-label trial compared the efficacy of ibuprofen plus paracetamol versus ibuprofen alone on patients with acute low back pain.[23] For the outcome pain intensity assessed using a visual analogue scale, patients were outcome assessors. Since patients were not blinded (open-label) and assessment of pain intensity involves considerable judgement, there is 'Definitely High' risk of bias.

Item 6 Outcome data not included in analysis

Two approaches for assessing risk of bias related to item 6

Item 6 addresses the randomized participants whose outcome data were not included in the analysis; that is, the participants whose outcome data were unavailable to systematic reviewers. Outcome data unavailable to reviewers could be a result of outcome data unavailable to trialists (e.g. lost to follow-up), or due to trialists excluded the participants for whom outcome data were available from the analysis because the participants did not adhere to the protocol (i.e., per-protocol analysis).

Rationale for completing only step 1
Step 1 asks the systematic reviewers to extract the number of participants whose outcome data were not included in analysis. The systematic review teams can use these numbers to assess risk of bias associated with missing data across the entire body of evidence at the meta-analysis level. The process of doing so begins with a complete-case analysis followed by an analysis imputing data for participants in each trial who were not included in the analysis.[1,2]

The review team can always use this approach either for binary or continuous outcomes but,[2] because the information needs is seldom available, rarely for time-to-event outcomes.[1]

If systematic review teams decide to use this approach to assess risk of bias associated with this item at the meta-analysis level, they can ask the systematic reviewers assessing individual trials to complete only step 1.

Rationale for completing both steps
Step 1 can serve to inform step 2. Step 1 asks systematic reviewers to extract the number of participants not included in analysis and calculate the overall percentage of participants not included in analysis. This percentage provides the information required for step 2, which asks systematic reviewers to classify trials into definitely low, probably low, probably high and definitely high risk of bias.

Set thresholds for step 2

If systematic reviewers need to complete both steps, systematic review teams need to set a threshold - the missing percentage - for each response option, as example below.

Step 2: Judge risk of bias related to the overall percentage of participants not included in analysis

Issue to consider: Was the overall percentage of participants not included in analysis acceptably low

Note: Systematic review teams need to set and fill in the threshold for each response option. See manual for instructions and example thresholds.

☐ **Definitely Low**	Percentage of participants not included in analysis is < 5 %
☐ **Probably Low**	• Percentage of participants not included in analysis is 5 % to < 10 % • If the trial did not mention whether there were participants not included in analysis, a substantial loss to follow-up is unlikely (e.g., ICU mortality)
☐ **Probably High**	• Percentage of participants not included in analysis is 10 % to < 15 % • If the trial did not mention whether there were participants not included in analysis, a substantial loss to follow-up is likely (e.g., 1-year quality of life)
☐ **Definitely High**	Percentage of participants not included in analysis is ≥ 15 %
Support for judgement:	

Specify post-randomization exclusion without bias (optional)

Review teams aiming at the highest level of rigor will consider one additional issue in addressing item 6. Usually, 'N total' is the number of randomized participants. However, sometimes, post-randomization exclusion does not introduce bias ('N exclusion without bias'). In this case, 'N excluded without bias' should be excluded from the 'N total' and 'N not analyzed'. Review teams will find a full explanation of the issue of exclusion without bias in:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1124168/[30]

Manual for completing step 1

'N total'
Usually, 'N total' is the number of randomized participants.

However, sometimes, post-randomization exclusion may be legitimate and does not introduce bias ('N exclusion without bias'). In this case, 'N exclusion without bias' should be excluded from the 'N total' and 'N not analyzed'.

Systematic reviewers should refer to the instructions that the systematic review team provided regarding in which cases post-randomization exclusion does not introduce bias.

'N not analyzed'
'N not analyzed' is: among the 'N total', how many participants' outcome data were not included in the analysis of the outcome of interest. Analyses of different outcomes may include different numbers of participants.

There are two main reasons for not including in analysis:

1) Missing outcome data

   Participants outcome data were unavailable to the trialists because of loss to follow-up, outcome data cannot be collected because of withdrawal of consent, etc.

   For time-to-event outcomes, missing outcome data refers to the participants who are censored because of missing follow-up data (i.e., informative censoring) rather than because of end of study.[1] Thus, systematic reviewers should count the number of participants who were censored because of missing follow-up data in 'N not analyzed'.

2) Per-protocol analysis

   Trialists may exclude the participants for whom outcome data were available from the analysis because the participants did not adhere to the protocol. For example, excluding because the participants did not receive assigned intervention, not meeting inclusion criteria, etc.

   Per-protocol analysis is a type of deviation from intention-to-treat analysis. Intention-to-treat analysis means that all participants for whom outcome data were available were included in the analysis and analyzed in the group to which they were randomized. Another type of deviation from intention-to-treat analysis is as-treated analysis, which is addressed as an optional item rather than in this item.

If the systematic review team specified in which cases post-randomization exclusion does not introduce bias, these cases should not be counted in 'N not analyzed'.

<u>Example:</u>
In a trial with 323 randomized patients, 35 were subsequently excluded from analysis: 1 patient died suddenly within 2 hours after randomization, 18 underwent mechanical ventilation <48h (did not meet the inclusion criterion of mechanical ventilation >48h), and 16 were not assessable because of missing important data.[15] The outcome of interest is clinically significant stress-related upper gastrointestinal bleeding.

Consider, if the systematic review team provides instructions that when considering non-death

outcomes post-randomization exclusion because of deaths does not introduce bias, then for the combined group 'N total' = 'N randomized' (323) - 'N exclusion without bias' (1 patient who died) = 322. Among the 322 patients, 34 were not included in analysis, thus 'N not analyzed' is 34. Thus, overall '% not analyzed' = 34/322 = 10.6%

Manual for completing step 2

In step 1, systematic reviewers should have calculated the overall percentage of participants not included in analysis. In step 2, systematic reviewers should, based on the thresholds that set by the systematic review team, classify the trial into ''Definitely Low', 'Probably Low', 'Probably High' or 'Definitely High'.

If the trial did not mention whether there were participants not included in analysis, classify the trial as 'Probably Low' or 'Probably High' depending on the nature of the outcome (easy to follow such as death, versus difficult to follow such as quality of life) and the follow-up time (short such as 1 week versus long such as two years).

Example:
A randomized trial with only an abstract compared pantoprazole with famotidine for stress ulcer prevention in ICU.[31] In the results section it stated: "Patients were randomized to IV pantoprazole (n=68) and IV famotidine (n=61) … Stress-related mucosal bleeding occurred in four patients in the pantoprazole group (3 overt bleeds, p=0.36 and 1 CSUGB, p=0.34) and one patient in the famotidine group (overt bleed). Patients given IV pantoprazole had similar mean ICU/CCU LOS (6.7 vs. 6.5 days, p=0.37) and duration on mechanical ventilation (6.5 vs. 6.3 days, p=0.39). Patients on pantoprazole (n=22) experienced more adverse effects than famotidine (n=10); hypomagnesemia (13 vs. 5) and nausea/vomiting (7 vs. 3); C. difficile diarrhea was the same in both groups (2 vs. 2)." This trial is only available as abstract and there is no detail regarding whether there were patients not included in analysis. Since the patients were followed until transferred out of the ICU (follow-up time is short), the trial could be classified as 'Probably Low'.

In addition:
If participants were not included in the trial analysis, but they were followed successfully and their outcome data were reported in trial publication or provided after inquiry, systematic reviewers should include these outcome data in their own analysis (i.e., conduct the intention-to-treat analysis themselves). In this case, these participants should not be counted as "N not analyzed" in this item.

Example: A randomized trial compared pantoprazole and enteral nutrition versus placebo and enteral nutrition for stress ulcer prophylaxis in critically ill patients with mechanical ventilation.[32] Investigators randomly allocated 124 patients. Of the 62 patients in pantoprazole group, 7 were excluded because they were extubated within the first 24h (one inclusion criterion was mechanical ventilation >48h; thus these patients did not adhere to the protocol); thus, the trial analysis included 55 patients. Of the 62 in placebo group, 15 were excluded because they were extubated within the first 24h; thus, the trial analysis included 47 patients. However, since the trial stated that none of these 22 patients developed any primary or secondary outcomes, these patients can be added back into the analysis. For instance, the trial reported one patient in each group (1/55 in

pantoprazole, 1/47 in placebo) developed overt bleeding. Systematic reviewers should extract the data for bleeding as 1/62 in pantoprazole group and 1/62 in pantoprazole group. In this case, all patients were included and analyzed in the group to which they were randomized.

Decide which if any optional items to include

In addition to the six core items (item 1 to 6), we provide eight optional items. Systematic review teams should decide whether to bring them to the attention of the systematic reviewers who assess individual trials (i.e., risk of bias assessors).

We include these eight items as optional rather than core items mainly either because the information required to make judgements is not commonly reported or because problems with these items occurs infrequently, or both.

However, these items may still be worth considering in certain circumstances.

Table 3 presents these eight optional items, reasons why they are not justified as core items, reasons why we include them as optional items, and further considerations regarding reasons systematic reviewers might or might not include these items in the ROBUST-RCT they use in their systematic review.

We strongly recommend that the systematic review teams consider each of the optional items carefully and judge whether they are relevant to their particular reviews.

If systematic review teams decide to include optional item(s), they need to provide the risk of bias assessors the instructions for assessing the items. Please see example instruction for the optional item 1 in Table 4.

Table 3. Optional items

	Reasons for not including as a core item	Reasons for including as an optional item in this instrument	Considerations regarding inclusion of this item in a systematic review
Optional item 1: **Whether the baseline prognostic factors were balanced between groups**	• When considering this item, one has to consider whether a baseline characteristic is an important prognostic factor in the particular context. • We have already included random sequence generation and allocation concealment as core items. Prognostic imbalance due to inappropriate randomization will be at least in part covered by these core items. If randomization is conducted properly and sample size is sufficient, prognostic imbalance would happen rarely and be simply due to chance. • While prognostic imbalance will often happen in small studies, it is much less likely across the entire range of studies (prognostic imbalance in one study is likely to be ameliorated by distribution of prognostic variables in other studies). • Empirical evidence regarding this item actually creating bias is very uncertain.	• When sample size is small, investigators may generate sequence appropriately and conceal and still have imbalance of prognostic factors simply by chance. • When a baseline characteristic with known appreciable prognostic power is imbalanced between groups, it may create serious bias.	• If there is problem with random sequence generation or allocation concealment (thus high risk of bias related to sequence generation or concealment item), no need to include this item. • If random sequence generation and allocation concealment performed well, and there is an important imbalance in important prognostic factors, an extra risk of bias exists. This item captures this problem. • The larger the imbalance in a factor, the stronger the prognostic power of the factor, the larger the number of trials in which this exists and the larger the weight of these trials in the meta-analysis, the more likely one would include this item.
Optional item 2: **Whether the co-interventions were balanced between groups in**	• It is difficult for non-expert systematic reviewers to judge what is a sufficiently important co-intervention and, if it is sufficiently important, when there is enough	• When sample size is small, investigators may generate sequence, conceal and conduct blinding appropriately and still have imbalance	• If there is problem with random sequence generation or allocation concealment (thus high risk of bias), or participants or healthcare providers

blinded trials	imbalance to consider as high risk of bias. • It is difficult to distinguish whether the imbalance in co-interventions is a risk of bias issue or is a function of the effect of intervention (e.g., participants experiencing adverse events receive additional drug to treat the adverse events). • If randomization is conducted properly and participants and healthcare providers are blinded, imbalance in co-interventions happens rarely and is simply due to chance.	of co-interventions simply by chance. • When an important co-intervention is imbalanced between groups and the imbalance in the co-intervention is not a function of the effect of intervention, it may create serious bias.	were unblinded, no need to include this item. • If randomization performed well and participants and healthcare providers were blinded (thus low risk of bias), and there is an important imbalance in important co-interventions, an extra risk of bias exists. This item captures this problem. • If the imbalance in co-intervention is a function of the effect of intervention, this is not a risk of bias problem. • The larger the imbalance in a co-intervention, the stronger the influence of the co-intervention on the outcome, the larger the number of trials in which this exists and the larger the weight of these trials in the meta-analysis, the more likely one would include this item.
Optional item 3: **Whether the outcome assessment or data collection differed between groups**	• Information required to make judgements is not commonly reported. • Even if reported, systematic reviewers will not make the judgement easily. • Different outcome measurement or data collection between groups happens rarely.	• If trials do report the information, and there is a difference between groups in the way of outcome assessment or data collection, it could lead to risk of bias.	• The problem is much more likely to occur in unblinded situations, but it is theoretically possible it could occur in blinded situations as well.

	• No empirical evidence supports this item.		
Optional item 4: **Whether the follow-up time, frequency, or intensity of outcome assessment differed between groups**	• Information required to make judgement is not commonly reported. • Even if reported, systematic reviewers cannot make the judgement easily. • Different follow-up time, frequency or intensity of outcome assessment between groups happens rarely. • No empirical evidence supports this item.	• If trials do report the information, and there is a difference between groups in the follow-up time, frequency or intensity of outcome assessment, it could lead to risk of bias.	• The problem is much more likely to occur in unblinded situations, but it is theoretically possible it could occur in blinded situations as well.
Optional item 5: **Whether the outcome measurement method was valid (i.e., validity of outcome measurement)**	• It is difficult for non-expert systematic reviewers to judge whether the outcome assessment method is valid or not. Even if the method has been validated, it may not be valid in the particular setting of the systematic review. • For many outcomes e.g., mortality, stroke, admission to hospital, this item is not applicable. • No empirical evidence supports this item.	If reviewers have major concerns that the outcome assessment method is not valid, risk of bias is present.	
Optional item 6: **When investigators conducted an as-treated analysis, was the percentage of participants not analyzed in the groups to which they were randomized sufficiently**	• Although "as treated analysis" (crossover occurred and participants were analyzed according to the intervention they received rather than the intervention they allocated to) is a serious problem, it happens very rarely.	• If in a randomized trial, crossover occurred and participants were analyzed according to the intervention they received rather than the intervention they allocated to (commonly referred to as "as-treated analysis"), and the percentage of participants analyzed in wrong group is large, there is serious risk of bias. • If in the context area of interest (e.g. surgical trial), participants were likely to switch to the other intervention group and trialists were likely to analyze these participants	

low	• No empirical evidence supports this item.	according to the intervention they received rather than the intervention they allocated to, systematic reviewers should include this item. • The larger the percentage of patients in all trials included in trials that used an as-treated analysis, and the larger the percentage who crossed in each of the trials, the stronger the case for including this item. • Question for this item could be: Among the participants who were included in analysis, was the percentage of participants analyzed in wrong group acceptably low.
Optional item 7: **Whether there was selective reporting**	• The protocol or analysis plan is not always available and if available it does not always provide sufficient detail to make the judgment regarding whether there are multiple choices regarding relevant analyses plans. • It is difficult for non-expert systematic reviewers to judge whether the selection is or is not based on the apparent magnitude of effect. • Systematic reviewers usually go back to the two-by-two tables and use the raw data to do their analysis, so the problem does not occur frequently. • Empirical evidence is very uncertain.	If the reported result is selected from multiple effect estimates that are available to the trialists (e.g., multiple outcome measurement instruments measuring the same construct, multiple time points, or multiple alternative analytic strategies) and the basis of the choice is the apparent magnitude of effect, or there is inconsistency between the protocol and the publication report, there may be a serious risk of bias.
Optional item 8: **Whether the trial was terminated early for benefit**	• Early termination for benefit is unlikely to be a risk of bias when a meta-analysis includes a large number of trials in which the contribution of the trials that stopped early for benefits is small.	When trials that terminated early for benefit have a substantial weight in the meta-analysis, there will be a serious risk of bias. For binary outcomes, if the trials had modest numbers of events, the problem will be more serious.[33] Systematic reviewers should be alert to this problem and certainly include it when the trials terminated early for benefit contribute substantial weight to the

	• Early termination does not occur frequently in some certain clinical areas.	pooled estimate.

Table 4. Example instruction for optional item 1

Optional item	Instruction	Judgement	Support for judgement
Whether the baseline prognostic factors were balanced between groups	**Definitely Yes/Low:** There was no imbalance in any important prognostic factors. **Probably Yes/Low:** • The imbalance in prognostic factor(s) may not result in important risk of bias because the imbalance may not be large enough and the prognostic power of the factor(s) may not be strong enough. • No problem with random sequence generation and allocation concealment, and the trial did not report prognostic factors. **Probably No/High:** The imbalance in prognostic factor(s) may result in important risk of bias because the imbalance may be large enough and the prognostic power of the factor(s) may be strong enough. **Definitely No/High:** The trial reported important imbalance in important prognostic factor(s). **Not applicable:** There is problem with random sequence generation or allocation concealment (definitely/probably high risk of bias related to sequence generation or concealment item), thus no need to assess this optional item.		

References

1. Goldkuhle M, Bender R, Akl EA, et al. GRADE Guidelines: 29. Rating the certainty in time-to-event outcomes-Study limitations due to censoring of participants with missing data in intervention studies. *J Clin Epidemiol* 2021;129:126-37. doi: 10.1016/j.jclinepi.2020.09.017 [published Online First: 20200930]
2. Guyatt GH, Ebrahim S, Alonso-Coello P, et al. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. *J Clin Epidemiol* 2017;87:14-22. doi: 10.1016/j.jclinepi.2017.05.005 [published Online First: 20170518]
3. Altman DG, Bland JM. Treatment allocation by minimisation. *BMJ* 2005;330(7495):843. doi: 10.1136/bmj.330.7495.843
4. Krag M, Marker S, Perner A, et al. Pantoprazole in Patients at Risk for Gastrointestinal Bleeding in the ICU. *N Engl J Med* 2018;379(23):2199-208. doi: 10.1056/NEJMoa1714919 [published Online First: 20181024]
5. Altman DG, Bland JM. How to randomise. *BMJ* 1999;319(7211):703-4. doi: 10.1136/bmj.319.7211.703
6. Lin CC, Hsu YL, Chung CS, et al. Stress ulcer prophylaxis in patients being weaned from the ventilator in a respiratory care center: A randomized control trial. *J Formos Med Assoc* 2016;115(1):19-24. doi: 10.1016/j.jfma.2014.10.006 [published Online First: 20150210]
7. Ben-Menachem T, Fogel R, Patel RV, et al. Prophylaxis for stress-related gastric hemorrhage in the medical intensive care unit. A randomized, controlled, single-blind study. *Annals of internal medicine* 1994;121(8):568-75. doi: 10.7326/0003-4819-121-8-199410150-00003
8. Powell H, Morgan M, Li SK, et al. Inhibition of gastric acid secretion in the intensive care unit after coronary artery bypass graft. A pilot control study of intravenous omeprazole by bolus and infusion, ranitidine and placebo. *Theor Surg* 1993;8:125-30.
9. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. *Lancet* 2002;359(9305):515-9. doi: 10.1016/S0140-6736(02)07683-3
10. Brophy GM, Brackbill ML, Bidwell KL, et al. Prospective, randomized comparison of lansoprazole suspension, and intermittent intravenous famotidine on gastric pH and acid production in critically ill neurosurgical patients. *Neurocritical care* 2010;13(2):176-81. doi: 10.1007/s12028-010-9397-3
11. Peto R. Failure of randomisation by "sealed" envelope. *Lancet* 1999;354(9172):73. doi: 10.1016/S0140-6736(05)75340-X
12. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869. doi: 10.1136/bmj.c869 [published Online First: 20100323]
13. Krag M, Perner A, Wetterslev J, et al. Stress ulcer prophylaxis in the intensive care unit trial: detailed statistical analysis plan. *Acta anaesthesiologica Scandinavica* 2017;61(7):859-68. doi: 10.1111/aas.12917 [published Online First: 20170612]
14. Setting Up the Randomization Module in REDCap – How-To Guide. https://www.ctsi.ufl.edu/files/2018/12/Setting-Up-the-Randomization-Module-in-REDCap.pdf
15. Kantorova I, Svoboda P, Scheer P, et al. Stress ulcer prophylaxis in critically ill patients: a randomized controlled trial. *Hepato-gastroenterology* 2004;51(57):757-61.
16. Selvanderan SP, Summers MJ, Finnis ME, et al. Pantoprazole or Placebo for Stress Ulcer Prophylaxis (POP-UP): Randomized Double-Blind Exploratory Study. *Crit Care Med* 2016;44(10):1842-50. doi: 10.1097/CCM.0000000000001819
17. Gundogan K, Karakoc E, Teke T, et al. Effects of oral/enteral nutrition alone versus plus pantoprazole on gastrointestinal bleeding in critically ill patients with low risk factor: a multicenter, randomized controlled trial. *Turk J Med Sci* 2020;50(4):776-83. doi: 10.3906/sag-1911-42

[published Online First: 20200623]
18. Alhazzani W, Guyatt G, Alshahrani M, et al. Withholding Pantoprazole for Stress Ulcer Prophylaxis in Critically Ill Patients: A Pilot Randomized Clinical Trial and Meta-Analysis. *Crit Care Med* 2017;45(7):1121-29. doi: 10.1097/CCM.0000000000002461
19. Al-Hwiesh A, Abdul-Rahman I, Finkelstein F, et al. Acute Kidney Injury in Critically Ill Patients: A Prospective Randomized Study of Tidal Peritoneal Dialysis Versus Continuous Renal Replacement Therapy. *Ther Apher Dial* 2018;22(4):371-79. doi: 10.1111/1744-9987.12660 [published Online First: 20180325]
20. Godfrey E, Wileman V, Galea Holmes M, et al. Physical Therapy Informed by Acceptance and Commitment Therapy (PACT) Versus Usual Care Physical Therapy for Adults With Chronic Low Back Pain: A Randomized Controlled Trial. *J Pain* 2020;21(1-2):71-81. doi: 10.1016/j.jpain.2019.05.012 [published Online First: 20190605]
21. Guyatt GH, Rennie D, Meade MO, et al. Users' guide to the medical literature : a manual for evidence-based clinical practice. New York: McGraw-Hill Education 2015.
22. van Workum F, Verstegen MHP, Klarenbeek BR, et al. Intrathoracic vs Cervical Anastomosis After Totally or Hybrid Minimally Invasive Esophagectomy for Esophageal Cancer: A Randomized Clinical Trial. *JAMA Surg* 2021;156(7):601-10. doi: 10.1001/jamasurg.2021.1555
23. Ostojic P, Radunovic G, Lazovic M, et al. Ibuprofen plus paracetamol versus ibuprofen in acute low back pain: a randomized open label multicenter clinical study. *Acta Reumatol Port* 2017;42(1):18-25.
24. Mousavi SA, Heydari K, Mehravaran H, et al. Melatonin effects on sleep quality and outcomes of COVID-19 patients: An open-label, randomized, controlled trial. *J Med Virol* 2022;94(1):263-71. doi: 10.1002/jmv.27312 [published Online First: 20210908]
25. Aggarwal R, Ranganathan P. Study designs: Part 5 - interventional studies (II). *Perspect Clin Res* 2019;10(4):183-86. doi: 10.4103/picr.PICR_138_19
26. Peterson J, Drazan D, Czajka H, et al. Immunogenicity and safety of a pentavalent meningococcal ABCWY vaccine in adolescents and young adults: an observer-blind, active-controlled, randomised trial. *Lancet Infect Dis* 2023;23(12):1370-82. doi: 10.1016/S1473-3099(23)00191-3 [published Online First: 20230811]
27. Ali K, Azher T, Baqi M, et al. Remdesivir for the treatment of patients in hospital with COVID-19 in Canada: a randomized controlled trial. *CMAJ* 2022;194(7):E242-E51. doi: 10.1503/cmaj.211698 [published Online First: 20220119]
28. Kalantari S, Fard SR, Maleki D, et al. Comparing the effectiveness of Atazanavir/Ritonavir/Dolutegravir/Hydroxychloroquine and Lopinavir/Ritonavir/Hydroxychloroquine treatment regimens in COVID-19 patients. *J Med Virol* 2021;93(12):6557-65. doi: 10.1002/jmv.27195 [published Online First: 20210728]
29. Blankestijn PJ, Vernooij RWM, Hockham C, et al. Effect of Hemodiafiltration or Hemodialysis on Mortality in Kidney Failure. *N Engl J Med* 2023;389(8):700-09. doi: 10.1056/NEJMoa2304820 [published Online First: 20230616]
30. Fergusson D, Aaron SD, Guyatt G, et al. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. *BMJ* 2002;325(7365):652-4. doi: 10.1136/bmj.325.7365.652
31. Wee B, Liu CH, Cohen H, et al. 731: IV Famotidine vs. IV Pantoprazole for Stress Ulcer Prevention in the ICU: A Prospective Study. *Critical Care Medicine* 2013;41(12):A181. doi: 10.1097/01.ccm.0000439969.36301.c9
32. El-Kersh K, Jalil B, McClave SA, et al. Enteral nutrition as stress ulcer prophylaxis in critically ill patients: A randomized controlled exploratory study. *J Crit Care* 2018;43:108-13. doi:

10.1016/j.jcrc.2017.08.036 [published Online First: 20170826]
33. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA* 2010;303(12):1180-7. doi: 10.1001/jama.2010.310

Chapter 5: Discussion and Conclusion to This Thesis

We came up with the idea for this thesis while considering that recent risk of bias assessment instruments had paid much attention to the methodological advances while sacrificing practicability and user-friendliness (1-3). This thesis aimed to develop an instrument for rating risk of bias in RCTs that aligns with the as-simple-as possible principle of evidence-based medicine (1). Target users of the instrument are systematic reviewers.

This thesis describes the detailed process for the instrument development. This thesis also includes the preparatory work that we conducted to support the instrument development. This chapter summarizes main findings, discusses strengths and limitations, and explores directions for future studies.

Main Findings

This thesis began with a systematic survey of existing RCT risk of bias instruments published from 2010 to October 2021 that documented their included items (chapter 2). The 17 eligible instruments included over a hundred unique items. More than half of the items were deemed by our expert panel as addressing other issues (e.g., applicability, imprecision, reporting quality) rather than risk of bias. Except for the revised Cochrane instrument (Cochrane RoB 2) (4), all other instruments included items not addressing risk of bias. This indicated that these instruments may not be appropriate for use to address risk of bias in systematic reviews. Since risk of bias is one of the five reasons for rating down the certainty of evidence, this is especially so when review teams apply the GRADE approach (5, 6).

The main objective of chapter 2 is to generate a candidate item list for the new instrument. Through an item classification survey of the panelists, we identified the items that are clearly related to risk of bias (majority agreed addressing risk of bias,

category 1), items that are clearly not related to risk of bias (majority agreed not addressing risk of bias, category 2), and items that panelists disagreed on whether they address risk of bias (category 3).

Another preparatory work is a systematic survey of meta-epidemiological studies evaluating whether and how the possible risk of bias items (items in category 1 and 3) influence estimates of intervention effects in RCTs (chapter 3). We used meta-analytic approach to combine the ratios of odds ratios from the meta-epidemiological studies and applied the GRADE approach (5) and ICEMAN instrument (7) to assess the certainty of inferences. This work demonstrated the importance of random sequence generation and allocation concealment, as well as the importance of patients blinding for patient-reported outcomes and outcome assessors' blinding for subjective outcomes. If investigators fail to ensure these methodological safeguards, trials possibly overestimate the effects of interventions. Empirical evidence for other items remains limited. This study provided empirical evidence that supported the item selection for the new instrument.

Chapter 4 describes the step-by-step process for developing the new instrument, named ROBUST-RCT. Chapter 4 presents the final version of the ROBUST-RCT and the user manual. ROBUST-RCT includes six core items each of which includes two steps: first evaluating what happened in individual trials and second judging the associated risk of bias. ROBUST-RCT provides eight optional items that may be relevant in specific cases. ROBUST-RCT achieved its goal of both methodological rigor and simplicity and user-friendliness.

Strengths and Limitations

This thesis followed a rigorous instrument development process. It started with

assembling a panel of experts with diverse backgrounds and setting up ground rules for instrument development. To support item selection for the new instrument, we conducted two preliminary works. Innovation of the survey of existing instruments (chapter 2) is that we conducted an item classification survey. Through this work, we classified the items into three categories, which became the starting point for selecting items for the new instrument.

Another preliminary work is a survey of meta-epidemiological studies (chapter 3). Strengths of this survey include the restriction to only meta-epidemiological studies that preserved the clustering design (see details in chapter 3). These restrictions facilitated the separate consideration of studies based on between-trial comparisons and within-trial comparisons and thus facilitated assessment of certainty of evidence and increased credibility of the results.

During the panel process of the instrument development (chapter 4), we used six criteria that helped deciding inclusion or exclusion of the candidate items. Through these criteria, we considered the items in a clear and comprehensive way. The criteria ensured the content validity of the instrument and also assured the core items to be easily assessed by junior systematic reviewers. The extensive pre-testing with both junior systematic reviewers and review experts further ensured the user-friendly and applicability of the ROBUST-RCT.

Individual chapters present the detailed limitations. Except for the limitations described in the individual chapters, this thesis does not address how to summarize the overall risk of bias in individual trials and how to use the ROBUST-RCT to inform whether to rate down the GRADE certainty of evidence due to risk of bias. The objective of assessing risk of bias in individual trials is to inform the decision of whether to rate down the certainty

of evidence for risk of bias. To make this decision, review teams do not necessarily need to summarize the overall risk of bias. Indeed, considering the whole picture of the risk of bias for each item in each trial can provide more complete inference. Even if reviewers do need the overall risk of bias results (e.g., they want to conduct subgroup analysis based on the overall risk of bias), summarization of overall risk of bias involves review team's judgement regarding the threshold, that is, failure to how many or which items would lead the reviewers to judge as overall high risk of bias.

In addition, our selection of optional items did not consider their relative importance and if review teams include optional items the relative importance between the optional items and the core items. For optional items, although we provided considerations regarding inclusion of these items in systematic reviews, we did not offer detailed instructions.

Directions for Future Research

We will address the issues of summarizing overall risk of bias and using the ROBUST-RCT to inform whether to rate down the certainty of evidence for risk of bias in a methodological guidance illustrating the essentials of the GRADE approach (i.e., Core GRADE).

Reliability and validity are two important properties of an instrument. We will assess the inter-rater reliability of the ROBUST-RCT in future. However, since there is no gold standard for rating risk of bias in RCTs (otherwise we would not develop the ROBUST-RCT), evaluating criterion validity of the ROBUST-RCT is impossible. If necessary, we will update the ROBUST-RCT based on the study results and user feedback. For optional items, if systematic reviewers find it difficult to construct instructions, we may offer detailed suggestions in future.

The current ROBUST-RCT addresses only risk of bias assessment of individually randomized parallel-group trials. Extensions of the ROBUST-RCT to other trial designs, e.g., cluster trials and crossover trials, is needed. Chapter 2 has identified the items specifically for cluster and crossover trials that included in existing instruments, which may serve as a starting point for item selection for extensions of ROBUST-RCT. The extension will follow the same motivation of maximizing simplicity while keeping methodological rigor. It may result in additional optional items addressing other trial designs and/or revision of the wording for existing items so that people could apply these items to other trial designs. This work will ensure the systematic reviews including any types of RCTs can successfully use the ROBUST-RCT.

To promote the application of the instrument, we plan to generate a website that presents the ROBUST-RCT and its extensions. It will also include any updates about the instrument in future. We will present the ROBUST-RCT in academic conferences. We may translate the ROBUST-RCT to other languages.

Moreover, we may conduct methodological studies about risk of bias. We developed instructions for assessing the allocation concealment status as probably yes or probably no when trialists do not report the concealment method clearly. A validation of our instructions by contacting trial authors for verification may be necessary.

References

1. Kuehn R, Wang Y, Guyatt G. Overly complex methods may impair pragmatic use of core evidence-based medicine principles. BMJ Evid Based Med. 2024.

2. Martimbianco ALC, Sa KMM, Santos GM, Santos EM, Pacheco RL, Riera R. Most Cochrane systematic reviews and protocols did not adhere to the Cochrane's risk of bias 2.0 tool. Rev Assoc Med Bras (1992). 2023;69(3):469-72.

3. 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.

4. Sterne JAC, Savovic 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.

5. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383-94.

6. 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(4):407-15.

7. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ. 2020;192(32):E901-E6.