BMJ Open

Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews: a research on research study using cross-sectional design

Lorenzo Bertizzolo,[1,2] Patrick Bossuyt,[2] Ignacio Atal,[1,3] Philippe Ravaud,[1,3,4,5,6] Agnes Dechartres[7]

To cite: Bertizzolo L, Bossuyt P, Atal I, *et al*. Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews: a research on research study using cross-sectional design. *BMJ Open* 2019;**9**:e028382. doi:10.1136/bmjopen-2018-028382

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-028382).

Received 5 December 2018
Revised 14 February 2019
Accepted 15 February 2019

For numbered affiliations see end of article.

Correspondence to
Dr Lorenzo Bertizzolo;
lorenzo.bertizzolo@gmail.com

ABSTRACT

Objectives Assess the frequency and reasons for disagreements in risk of bias assessments for randomised controlled trials (RCTs) included in more than one Cochrane review.

Design Research on research study, using cross-sectional design.

Data sources 2796 Cochrane reviews published between March 2011 and September 2014.

Data selection RCTs included in more than one review.

Data extraction Risk of bias assessment and support for judgement for five key risk of bias items.

Data synthesis For each item, we compared risk of bias assessment made in each review and calculated proportion of agreement. Two reviewers independently analysed 50% of all disagreements by comparing support for each judgement with information from study report to evaluate whether disagreements were related to a difference in information (eg, contact the study author) or a difference in interpretation (same support for judgement but different interpretation). They also identified main reasons for different interpretation.

Results 1604 RCTs were included in more than one review. Proportion of agreement ranged from 57% (770/1348 trials) for incomplete outcome data to 81% for random sequence generation (1193/1466). Most common source of disagreement was difference in interpretation of the same information, ranging from 65% (88/136) for random sequence generation to 90% (56/62) for blinding of participants and personnel. Access to different information explained 32/136 (24%) disagreements for random sequence generation and 38/205 (19%) for allocation concealment. Disagreements related to difference in interpretation were frequently related to incomplete or unclear reporting in the study report (83% of disagreements related to different interpretation for random sequence generation).

Conclusions Risk of bias judgements of RCTs included in more than one Cochrane review differed substantially. Most disagreements were related to a difference in interpretation of an incomplete or unclear description in the study report. A clearer guidance on common causes of incomplete information may improve agreement.

Strengths and limitations of this study

- ► Use of a very large and comprehensive collection of Cochrane reviews to assess the agreement in risk of bias assessment and to understand reasons of disagreement.
- ► Analysis of the full text of study reports to underline what information was available to review authors and how they used them while assessing risk of bias.
- ► Focus on disagreements only. Possible that a proportion of agreements happened 'by chance'. For example, review authors may express the same risk of bias judgement while using different information or interpreting information differently.
- ► No evaluation of the potential impact of disagreements in conclusion making at the review level.

INTRODUCTION

Systematic reviews aim to synthesise all existing evidence for a research question by the use of a rigorous and reproducible methodology.[1] Because reviews may be affected by bias at the level of individual studies,[2] an assessment of the risk of bias in these studies is a crucial step in conducting a systematic review.[3 4]

Cochrane has developed a tool to provide a standardised approach to the assessment of the risk of bias in randomised controlled trials (RCTs).[5] The risk of bias tool is based on specific characteristics related to study design and conduct, selected on theoretical grounds and on empirical evidence from

meta-epidemiological studies that these characteristics are associated with differences in treatment effect estimates.[6–11] The tool includes seven items (random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective reporting, other source of bias), which the researchers assess and judge as either 'high', 'low' or 'unclear' risk of bias.[11 12]

Although Cochrane provides detailed guidance on how to use the tool and recommends consensus between two independent reviewers,[11] personal judgement is also involved, which may bring variability. Several studies have evaluated the reproducibility of the risk of bias tool, generally shown to be poor.[12–19] However, there is some uncertainty about the main causes of disagreements. For example, some reviewers may search for additional information such as protocols or contact study authors and this difference in available information, rather than a difference in judgement, may explain some of the disagreements.

In this study, we used a large collection of Cochrane reviews to evaluate the reproducibility of risk of bias assessments by identifying RCTs included in more than one Cochrane review and comparing the assessments. In addition, we examined the likely reasons for any disagreements. In particular, we evaluated whether disagreements were related to differences in information available to reviewers or differences in interpreting the same information and what could explain such different interpretation.

METHODS

This is a research on research study on risk of bias assessment, which used a cross-sectional design. We identified RCTs included in more than one reviews included in a large collection of Cochrane reviews. For key risk of bias items, we evaluated agreement between the different systematic reviews; analysed whether disagreements were related to a difference in information available to reviewers or a difference in interpretation of the same information and highlighted the main reasons for disagreements by an in-depth, one-by-one evaluation of disagreements.

Data sources

We obtained data from the 2796 Cochrane reviews, which correspond to all the reviews available in the Cochrane library between March 2011 and September 2014, including updates (March 2011 corresponds to the last update of the risk of bias tool[5]). Data consisted of one XML file per review, each file containing all data entered by review authors in RevMan, the software used for managing Cochrane reviews.[20] All individual XML files were merged in a single database by using R V.3.2.2[21] with the XML package.[22] The vocabulary used for risk of bias items slightly varied across reviews (eg, some reviews could refer to 'allocation concealment' as 'allocation masking'). For this reason, two authors independently evaluated all terms used and classified them according to the vocabulary of the tool. Disagreements were resolved by consensus. This standardisation was done for a previous publication.[23]

Selection of eligible reviews

We excluded withdrawn or 'empty' reviews (ie, systematic reviews not including any study) as well as reviews including observational or non-randomised studies and considered only reviews with an assessment of risk of bias for at least one item of the risk of bias tool.

Selection of eligible RCTs

To identify single RCTs included and assessed for risk of bias in more than one systematic review, we proceeded as follows. For each RCT, we identified the primary reference(s), which was the reference identified by review authors as the main reference(s) for an included study. Then, we used a matching algorithm[24] to identify studies that shared the same primary reference. If several primary references were reported, we considered all of them. We manually checked that the studies sharing the same primary reference in the reviews corresponded to the same RCT.

Extraction of risk of bias assessment

For each eligible RCT, we extracted the risk of bias assessment and the corresponding support for judgement for each risk of bias item in each review. Whenever a single RCT was included in three or more reviews, we considered only the risk of bias assessment from two reviews chosen at random; this was decided because of workload and to facilitate direct comparison of two assessments and concerned less than 10% of our included RCTs. We focused on five risk of bias items: random sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment and incomplete outcome data. We did not consider selective reporting because it is difficult to evaluate in the absence of the study protocol, which is frequently lacking, especially for older studies.[11 12 14] We also did not consider the item other bias because the definition is very wide (ie, 'any important concerns about bias not covered in the other domains in the tool'[11]), so comparisons across reviews are difficult.

Comparison of risk of bias assessment between reviews

For each item, we compared the risk of bias assessment in terms of 'high', 'low' or 'unclear' risk of bias between the two reviews. According to the Cochrane handbook, the items blinding of outcome assessment and incomplete outcome data should be assessed for each outcome. Therefore, when the reviews reported an assessment of these items at the outcome level, we manually checked that outcomes were identical in both reviews and we retained for our analysis only the assessments that focused on the same outcomes. For blinding, we followed the last version of the Cochrane handbook and we retained only assessments of blinding of participants and personnel

and blinding of outcome assessment as two independent items, excluding different types of assessment (ie, blinding as a single item, blinding of only participants or of only personnel).

We calculated the percentage agreement for each risk of bias item, as the proportion of studies with a concordant assessment in both reviews (eg, 'low' risk of bias AND 'low' risk of bias). Not all reviews assessed all five key risk of bias items for each RCT included; consequently, the number of RCTs evaluated for discrepancies varies depending on the item considered.

Selection of studies for in-depth analysis of disagreements

For workload reasons, we in-depth evaluated the reasons for disagreements for 50% of the studies analysed in the previous step. In cases of more than one shared RCTs within a given pair of Cochrane reviews, we selected only one RCT at random. To reach 50% of the total sample, we used a simple random selection in the remaining database.

Classification of disagreements

For the random selection, two reviewers (LB and AD) independently evaluated all disagreements in the risk of bias assessment in the two systematic reviews. They first scrutinised the support for the judgement in each review and evaluated whether it was the same or 'conceptually' the same in both reviews (eg, 'randomised, probably done'; 'randomised, probably not done'; 'study only mentions randomisation, but does not specify how randomisation was performed; unclear'; 'study states it is randomised; low risk'). If the support differed, they assessed any other information regarding the study as reported in both reviews, systematically searching and evaluating the full-text study report indicated in the primary reference. A formalised data extraction process for full texts was not used. Full texts were examined, looking primarily for correspondence between information reported by the reviewers in their support for judgement and the text.

They independently classified each case of disagreement as follows:

- Disagreement related to differences in interpretation:
  - The support for judgement was the same (or 'conceptually' the same) in both reviews, but the interpretation differed.
  - One review clearly confused one item of the risk of bias tool with a different one or the review authors misunderstood the definition of the item (eg, for random sequence generation, support for judgement reports '600 opaque envelopes, 1 was drawn every time').
- Disagreement related to differences in information: the support for judgement cites information that is not available in the study report; additional sources are cited (eg, protocol) or the review authors reported that they had contacted the RCT author for additional data.
- Disagreement related to information missed by the review authors: the study report clearly describes the information, but some review authors seemed to have missed this information in the study report.
- Disagreement related to input mistakes: risk of bias assessment in terms of 'high'/'low'/'unclear' did not match the support for the judgement (eg, 'randomisation described explicitly', judgement 'unclear').
- Unclear: when it was not possible to classify the disagreement because the support for the judgement was empty or because we could not retrieve the full-text study report.

Any disagreements between reviewers were solved by discussion to reach consensus. In the online supplementary appendix 1, we report a figure synthetising how the in-depth analysis process was conducted.

Identification of main reasons for different interpretation

For each disagreement related to a difference in interpretation, we evaluated the probable reason for disagreement. For example, the interpretation could differ because of confusion with another risk of bias item (eg, random sequence generation and allocation concealment) or because the information was unclear or insufficiently detailed in the article. When we were unsure about the reason, we classified the reason as unclear. Two authors (LB and AD) conducted this process in duplicate by using all available information (ie, support for the judgement, characteristics of the study reported in the review, full-text article), with disagreements resolved by discussion.

Statistical analysis

Analysis was descriptive with use of frequencies and percentages for qualitative variables. Statistical analysis was conducted with Stata V.13.1.[25] We decided to use simple per cent agreement because other static approaches were problematic. The Kappa statistic requires having defined reviewers, which is not the case of our approach. Another statistic, the intraclass correlation coefficient is not suitable, because it requires assessments to be in an ordinal order, which is not our case. There is no continuum between the assessments of low, unclear and high risk of bias.

Patient involvement

Patients were not involved in any aspect of the study design, conduct or the development of the research question or outcome measures. This is a research-on-research study, and therefore, there was no active patient recruitment for data collection.

RESULTS

Selection process

Figure 1 shows the selection process. From the 2796 systematic reviews published between March 2011 and September 2014, 2291 reviews included RCTs only and

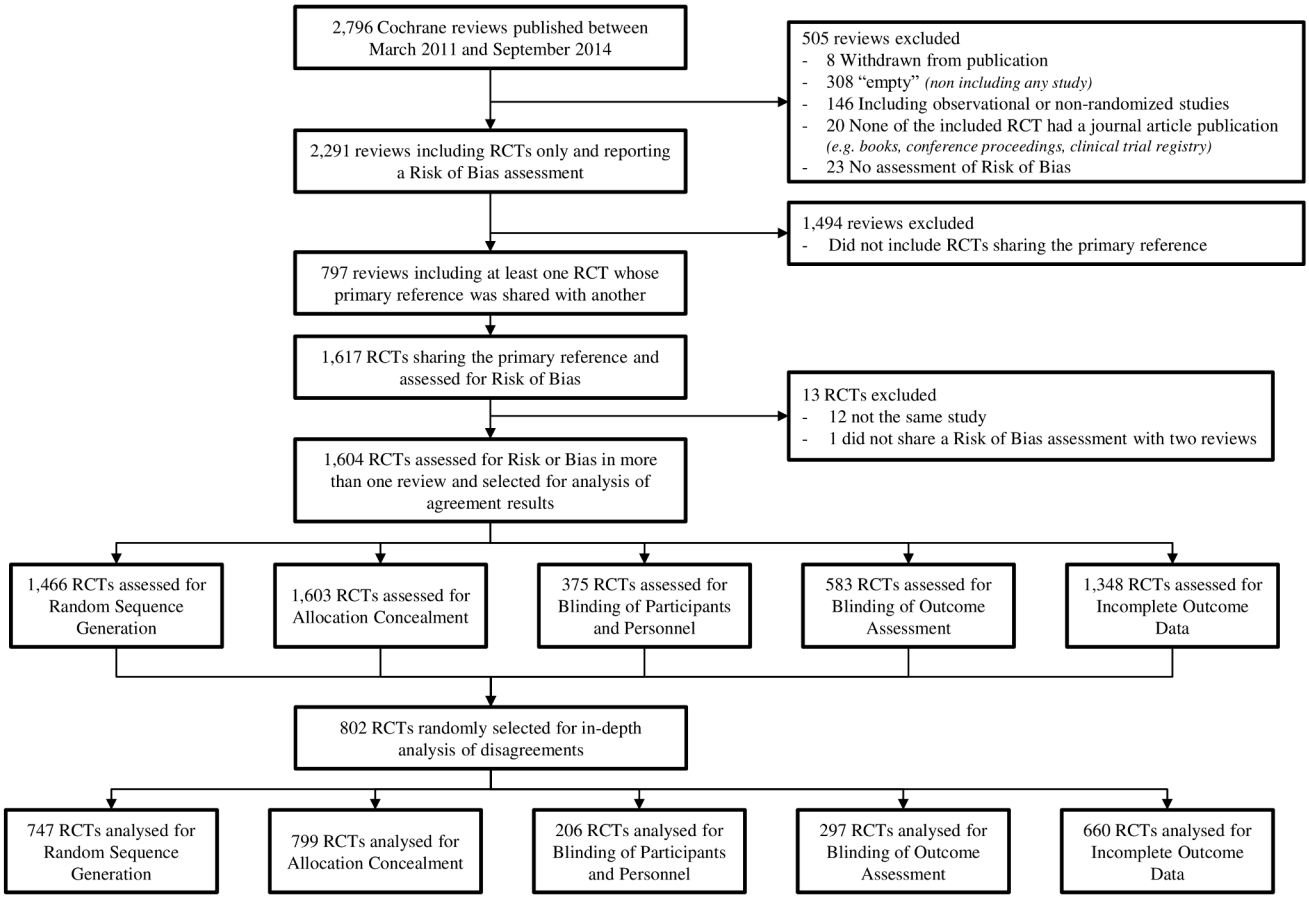

Figure 1 Flow chart of the selection process. RCT, randomised controlled trial.

reported a risk of bias assessment. Of these, 797 included at least one RCT whose primary reference was shared with another review for which a risk of bias assessment was reported. These 797 reviews included 1604 single RCTs evaluated for the same risk of bias item in more than one review. The online supplementary appendix 2 reports the frequency of the different Cochrane groups among those reviews.

Among the 1604 selected RCTs: 1603 had duplicate evaluation for allocation concealment, 1466 for random sequence generation, 375 for blinding of participants and personnel, 583 for blinding of outcome assessment and 1348 for incomplete outcome data.

Evaluation of agreement and distribution of disagreements

The agreement of risk of bias judgements ranged from 57% (770/1348 trials) for incomplete outcome data to 81% (1193/1466 trials) for random sequence generation (figure 2). We identified most disagreements for 'low' and 'unclear' risk of bias judgements, especially for random sequence generation (231/273 trials, 85%). Disagreements between 'low' and 'high' risk of bias were generally rare, for example 8/273 of disagreements (3%) for random sequence generation, with the exception of incomplete outcome data for which they were more frequent (190/578, 33%). For blinding of participants and personnel, the most frequent disagreement was between 'unclear' and 'high' risk of bias (50/107, 47%), then 'low' versus 'unclear' (34/107, 32%), and 'low' versus 'high' (23/107, 21%) (figure 2).

Classification of disagreements

The in-depth analysis of disagreements included 802 studies: 799 for allocation concealment, 747 for random sequence generation, 206 for blinding of participants and personnel, 297 for blinding of outcome assessment and 660 for incomplete outcome data. The agreement results of this sample and the distribution of disagreements are reported in the online supplementary appendix 3.

For all items, the most common source of disagreement was a difference in interpretation, with frequencies ranging from 88/136 (65%) for random sequence generation to 56/62 (90%) for blinding of participants and personnel (figure 3). The access to additional or different information accounted for disagreements in 32/136 (24%) trials for random sequence generation and 38/205 (19%) for allocation concealment. Access to additional information was less common for the remaining items, with proportions ranging from 2% to 4%. In 80% of the cases, the access to additional information was through the contact of the study author.

The other sources of disagreement were less common; input mistake ranged from 1% to 6%, missed information from 1% to 6%. We could not

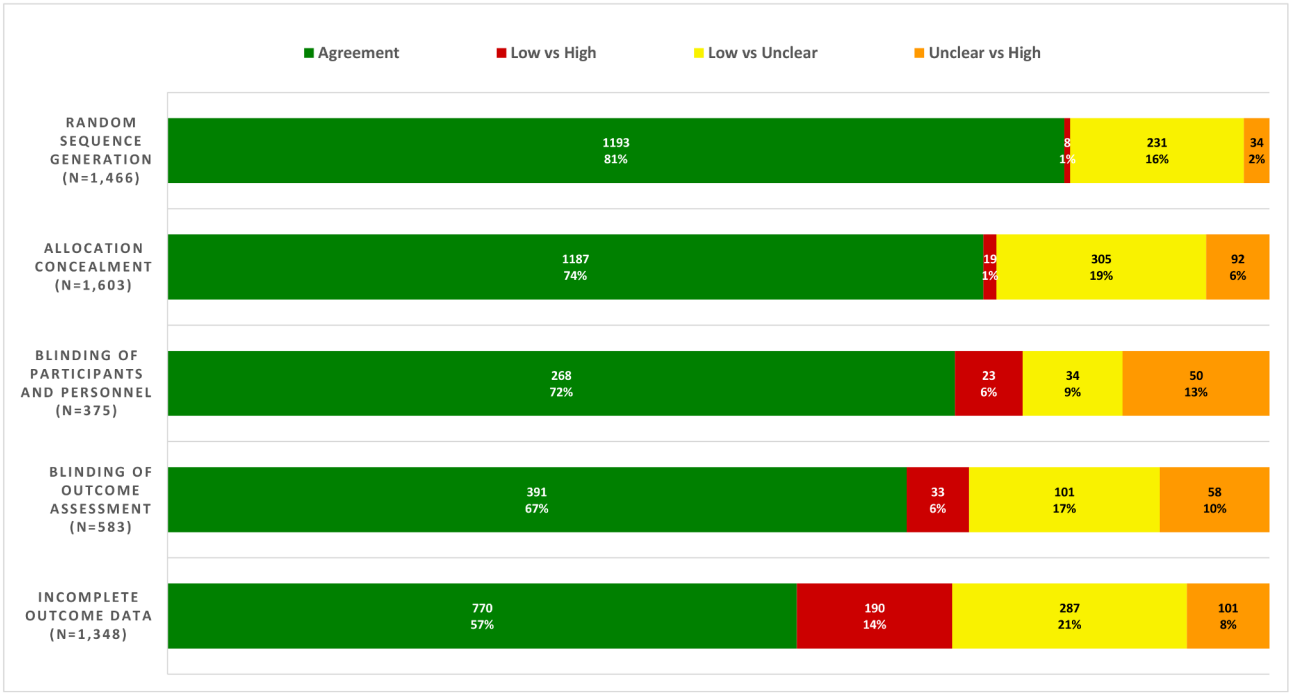

Figure 2 Distribution of agreements and disagreements for the different risk of bias items analysed; raw number and percentage of the total. For disagreements, distribution of the different discrepancies.

determine the source of disagreement in 5% of our disagreements. For this analysis, we accessed the full text of 216 different trials to help us in the process. The online supplementary appendix 4 reports some examples of disagreements in which the access to the study report helped us in the classification and the analysis of reasons of disagreement. We could not retrieve or access 19 full texts we deemed necessary for the categorisation of disagreements and this explains the majority of cases where we were unable to categorise the source of disagreement ('unclear' source in figure 3).

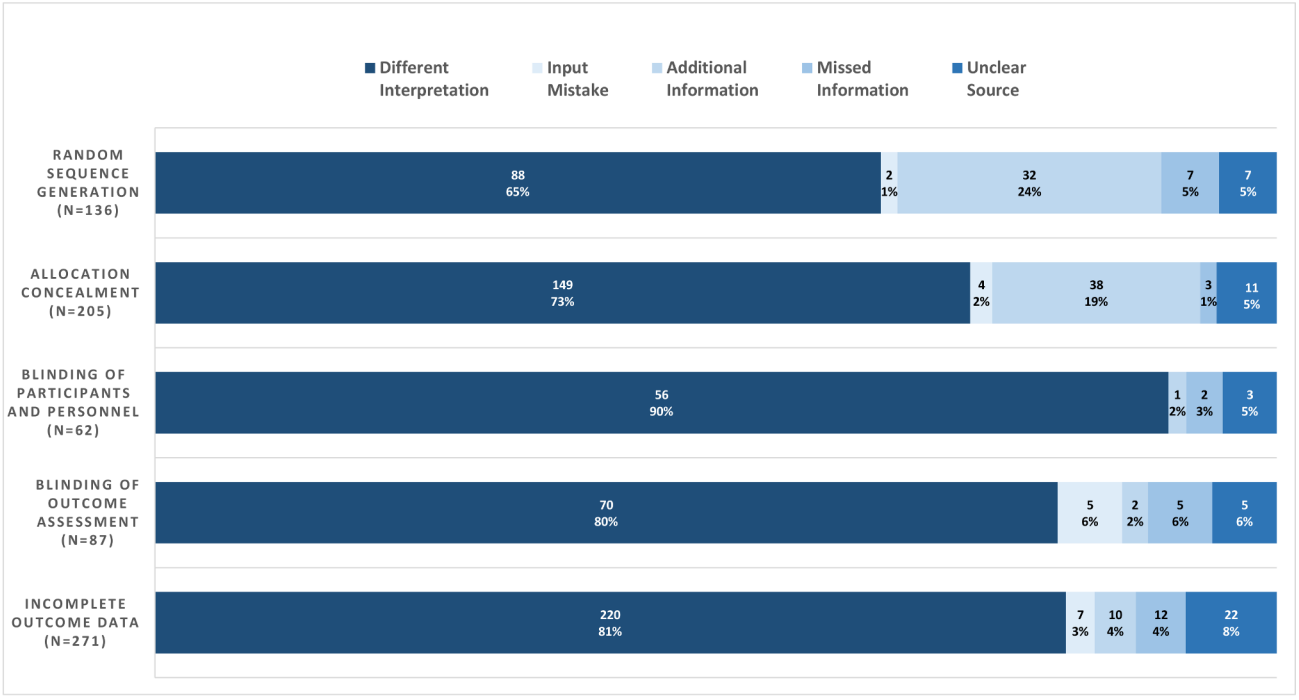

Figure 3 Classification of disagreements for the different items; raw number and percentage of the total.

Main reasons of disagreements for different interpretation

The main reasons for a difference in interpretation for each item are reported in table 1. Additional examples are provided for each item for the high-low disagreements (online supplementary appendix 5). The most common reason across items was related to incomplete or unclear reporting in the RCT. For random sequence generation, disagreements in 73/88 (83%) trials were related to lack of a precise description of the randomisation process with reviewers evaluating 'low', 'high' or 'unclear' risk of bias the reporting of 'randomised' in the text. For allocation concealment, the most common reason for disagreement was a different interpretation of description of the envelopes used to conceal allocation (17%, n=26/149 trials). For the two blinding items, many disagreements occurred when the article mentioned only 'double blind' in RCTs without an additional description (16% of cases, n=9/56 trials for blinding of participants and personnel, 13%, n=9/70 for blinding of outcome assessment). For incomplete outcome data, reviewers assessed differently the statement from the study report of 'no missing data' or 'all data reported' (10%, 22/220 trials). Another common reason for a difference in interpretation was confusion with another item. Allocation concealment was confused with blinding (10%, n=15/149 trials) but also with random sequence generation (4%, n=6/149). For blinding of participants and personnel, the most common cause for disagreement concerned the interpretation of cases when blinding was not feasible (36%, n=20/56 trials), assessed at high risk by some reviewers and low by others. Another common cause of disagreement for the two blinding items related to the assessment of outcomes that should not be affected by blinding (eg, mortality); it explained 21% (n=12 trials) of disagreements for blinding of participants and personnel and 23% (n=16 trials) for blinding of outcome assessment, often low versus high disagreements.

For incomplete outcome data, the use of different cut-offs for the rate of missing data is the most common reason for disagreement (26%, n=57 trials); also common is considering the explanation of reasons for missing data enough to attribute a low risk of bias (13%, n=28 trials).

DISCUSSION

In this study, we took advantage of a very large sample of Cochrane reviews to explore the sources of disagreements in risk of bias assessment for trials included in several reviews. We decided to focus on Cochrane reviews because as these reviews are produced within a single organisation, therefore, we expected results and procedures to be more appropriately comparable. Authors compiling Cochrane reviews are members of the organisation and, in most cases, they underwent a similar training for assessing risk of bias. Our results confirm that the agreement for risk of bias assessments is generally suboptimal, with better agreement for random sequence generation and allocation concealment and less agreement for incomplete outcome data. Access to different sources of information explained why 24% of the trials had disagreements in the assessment of risk of bias for random sequence generation and 19% for allocation concealment. However, the main source of disagreements was a difference in interpretation of the same information, which was frequently related to incomplete or unclear reporting in the study report.

Strengths and weaknesses

Our study goes beyond previous literature on the topic.[3 12–18 26] As compared with most other studies,[12–17] we used real-world data to explore agreement of risk of bias assessments in real scenarios. We evaluated a very large and comprehensive collection of Cochrane reviews that spanned multiple specialties and topics, including a number of trials about 10 times larger than the largest study on the topic.[12] We completed our analysis by searching individual study reports to give support to our comments on reasons for disagreements, which, to our knowledge, has not been done in previous, smaller works that used a similar methodology.[18] While doing this, we developed a suitable classification scheme for sources of disagreements and conducted, in duplicate, an extensive analysis to understand the risk of bias assessment process and explored the most common reasons for disagreements.

Our study has limitations. Whenever a single RCT was included in three reviews or more, we considered only the risk of bias assessment from two reviews chosen at random. Nevertheless, we cannot exclude that different combinations of two chosen evaluations could have produced slightly different results. Although the classification of disagreements was conducted in duplicate following a formalised process, there remains a component of personal judgement. We evaluated only disagreements, but a number of agreements might have occurred 'by chance'. In our analysis of likely reasons for disagreements, some resulted from confusion between risk of bias items. Similar discrepancies might have occurred among agreements; indeed, previous literature on the topic demonstrated that reviewers do not accurately follow the risk of bias tool.[27] We also did not assess the selective reporting item that is frequently judged on incomplete information. We did not evaluate whether disagreements varied depending on the Cochrane review group or year of publication. Finally, we did not evaluate the impact of disagreements and the extent to which the evidence base for making conclusions and providing summary statements of effectiveness may have been affected by changing the rating.

Comparison with other studies

Our findings confirm the importance of issues that were previously identified by Jørgensen *et al*[3] and Savović *et al*.[26] In particular, Savović *et al*,[26] surveying users of the risk of bias tool, reported on the possibility of confusion between random sequence generation and allocation

Table 1 Main reasons for disagreements in cases of a different interpretation of the same information

Risk of bias Item	Main reasons for disagreements	N (%)*	Examples of support for judgement from the review†
Random sequence generation	Consider differently incomplete or unclear description	73 (83)	'States 'cluster randomisation by computer"; Low risk of bias 'Cluster randomisation by computer. No further information provided'; Unclear risk of bias
	Confusion with allocation concealment	9 (10)	'Allocation was done using sealed envelopes containing name of one of the two groups.'; Low risk of bias
Allocation concealment	Consider differently incomplete or unclear description	49 (33)	'Not specified.'; High risk of bias 'Method of concealment not described.'; Unclear risk of bias
	Consider differently envelopes description	26 (17)	'Sequentially numbered sealed envelopes'. Does not state if opaque envelopes.'; Unclear risk of bias 'Sequentially numbered sealed envelopes.'; Low risk of bias
	Random sequence generated by computer or external centre considered enough for low risk	21 (14)	'Treatment was allocated based on the computer-generated no list.'; Low risk of bias
	Confusion in the definition of the item	19 (13)	'Researchers attempted to contact all patients seen by physicians during 1 month'; High risk of bias
	Confusion with blinding	15 (10)	'Participants were told to which compound they had been allocated.'; High risk of bias
	Confusion with random sequence generation	6 (4)	'Computer-generated randomised lists.'; Low risk of bias
Blinding of participants and personnel	Assess risk differently if blinding was not feasible because of the type of intervention	20 (36)	'Not possible to blind participants'; Low risk of bias 'Participants were not blinded for provided treatment. This is inherent to study design'; High risk of bias
	Outcome considered not influenced by blinding	12 (21)	'No information given about whether patients were blind to physician allocation but treatment outcomes judged unlikely to be affected by lack of blinding'; Low risk of bias
	Consider differently information of 'double blind'	9 (16)	'Quote: ' . . . patients were randomised in double-blind conditions . . . 'Comment: probably done'; Low risk of bias 'Quote: 'double blind conditions'. No further details.'; Unclear risk of bias
	Consider differently incomplete or unclear description	7 (12)	'Researchers were blind until after the baseline assessment. participants were not blinded.'; Unclear risk of bias 'Not possible to blind participants to intervention. Insufficient information to make a judgement about blinding of therapists'; High risk of bias
	Confusion in the definition of the item	5 (9)	'Described as an 'open-label' pilot study.'; Low risk of bias
Blinding of outcome assessment	Consider differently incomplete or unclear description	24 (34)	'Not explicitly discussed in the publish study, it was assumed to be open label'; High risk of bias 'Not described in published study'; Unclear risk of bias
	Outcome considered not influenced by blinding	16 (23)	'Not stated, but it was unlikely that the outcome was influenced by lack of blinding'; Low risk of bias
	Consider differently patient-reported outcomes when patients are blinded or not to the intervention	9 (13)	'Comment: depression assessed by patient self-report'; High risk of bias 'Insufficient information available to assess'; Low risk of bias
	Consider differently information of 'double blind'	9 (13)	'Quote: ' . . . double blind' Comment: probably done'; Low risk of bias 'Quote: 'double blind conditions'. No further details.'; Unclear risk of bias
	Assess risk differently if blinding was not feasible because of the type of intervention	6 (9)	'Blinding not possible due to intervention'; High risk of bias 'Unclear blinding of outcome assessment'; Low risk of bias

Continued

Table 1 Continued

Risk of bias Item	Main reasons for disagreements	N (%)*	Examples of support for judgement from the review†
Incomplete outcome data	Use different cut-off for the rate of missing data	57 (26)	'11 withdrawals (10%).'; Low risk of bias 'Comment: there were post-randomisation drop-outs'; High risk of bias
	Focus on no versus reasons/precise report of missing data	28 (13)	'20 drop-outs (27.2%) with 4 deaths (3 males, 1 female) from cardiovascular events'; High risk of bias 'Numbers and reasons for drop-outs and withdrawals in all intervention groups were described.'; Low risk of bias
	Consider differently incomplete or unclear description	27 (12)	'Women who were untraceable or unsuitable for follow-up were excluded, other losses included as smokers'; Low risk of bias '167/1287 (12.9%) (C=83, I=84) excluded from analysis due to moving away, being untraceable or deemed unsuitable for follow-up (eg, miscarriage). 1120 in sample. 51/1287 non-responders were included as continuing smokers.' High risk of bias
	Consider differently intention-to-treat (ITT) analysis	25 (11)	'147 randomised; r4 in the letrozole group and 3 in the laparoscopic ovarian drilling dropped out of the trial, all for non-compliance. However, ITT analysis was not conducted.'; Unclear risk of bias '7 women lost to follow-up, but similar (3 vs 4) in both groups; losses due to non-compliance'; Low risk of bias
	Consider differently report of 'no missing data'	22 (10)	'Did not report number of withdrawals. Comment: all patients who were randomised were included in the final analysis. ITT analysis was conducted.'; Unclear risk of bias 'It does not appear that there were any withdrawals or drop-outs' Low risk of bias
	Consider differently imputation of missing data	20 (9)	'Imputation method not described'; Unclear risk of bias 'Drop-out rate was not significant'; Low risk of bias
	Use different cut-off for difference in the rate missing data between different arms/comparisons	13 (6)	'Drop-out higher in placebo group (35% vs 25% in budesonide group). ITT used.'; High risk of bias 'Similar rates of withdrawal between arms. Withdrawals: 36 BUD, 51 placebo'; Low risk of bias

*Number of RCTs disagreeing for this reason; percentage over the total of disagreements for different interpretation.
†When two extracts are reported, they refer to the same study.
RCT, randomised controlled trial.

concealment and between allocation concealment and blinding; the uncertainty on how to address unfeasibility of blinding; and the difficulties in assessing incomplete outcome data especially regarding the acceptable rate of missing data. More recently, Jørgensen *et al*,[3] evaluating comments on the use of the risk of bias tool, highlighted how authors complained that judgement often originates from incomplete or missing information.

A previous study identified 46 RCTs included in different systematic reviews in the field of fertility and evaluated the percentage agreement in risk of bias assessment. That analysis showed generally worse agreement than in our study, with percentage agreement ranging from 35% to 71%. Differences in sample size and the particular topic may explain these discrepancies. In addition, although the authors had compared supports for judgement between reviews, this evaluation may have been incomplete, because they did not evaluate the primary study reports.[18]

Implications

Our results confirm that the agreement in risk of bias assessment would be enhanced by more detailed guidance in use of the risk of bias tool with particular focus on common causes of disagreements. We showed that in many cases, the unclear reporting from source material allows reviewers ample space for personal judgement and differences in judgement.

The scientific community continues to stress the importance of improving the reporting of trials,[28–31] which may limit disagreements when assessing risk of bias. In parallel, we could also work on restricting the space for personal interpretation when assessing risk of bias. A suggestion could be to give clearer instruction on how to evaluate common cases, for example, when confronted with nothing more than the term 'randomised' or 'double blind' in the study report. Similarly, a threshold could be set on the quota for missing data and indications on which imputation methods are appropriate and in which situations.

To minimise research waste, it could be interesting to have access to risk of bias assessments from other Cochrane groups and the supports they used, including information from authors or from protocols to help reviewers in their assessments. This process would imply having a unique study identification number across reviews and a central shared repository for all studies included in any Cochrane reviews.

Following the suggestions based on the findings and comments provided by Jørgensen *et al*[3] and Savović *et al*,[26] Cochrane has been working on a new version of the risk of bias tool, which has recently been released.[32 33] The new version has a different approach to the risk of bias assessment, guiding reviewers through the process with the use of 'signalling questions', which might leave less room for subjectivity. In addition, there is more guidance in assessing some items. For example, the new tool better clarifies some aspects of the randomisation process, especially about what to do in some cases of incomplete information (eg, randomisation list created by an external centre with no other indication). The new tool also has a different approach to the blinding aspect, oriented to the implications of the masking process. However, the new tool does not cover some of our concerns, especially those related to incomplete outcome data: quota for missing data that are considered acceptable, and whether reviewers should focus more on the reasons for the missing data or their magnitude. It also does not address the common case of authors reporting 'no missing data'. Research-on-research studies are needed to evaluate whether this new version of the tool results in improved reproducibility.

CONCLUSION

This analysis of risk of bias assessment for more than 1600 trials included in more than one reviews showed that agreement remains suboptimal. Most disagreements come from a difference in interpretation of an incomplete or unclear description in the study report. In some cases, the difference in the assessment was due to some but not all review authors obtaining additional information, from a protocol or from contacting study author.

Author affiliations
[1]U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS), INSERM, Paris, Île-de-France, France
[2]Departmentof Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
[3]Cochrane France, Paris, France
[4]Centre d'Épidémiologie Clinique, INSERM U738 Hôpital Hôtel Dieu AP-HP (Assistance Publique des Hôpitaux de Paris), Paris, France
[5]Faculté de Médecine, Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[6]Department of Epidemiology, Columbia University, Mailman School of Public Health, New York, USA
[7]Sorbonne Université, INSERM, Institut Pierre Louis d'Epidémiologie et de Santé Publique, AP-HP, Hôpitaux Universitaires Pitié Salpêtrière - Charles Foix, Département Biostatistique Santé Publique et Information Médicale, F75013, Paris, France

Acknowledgements We thank Camila Olarte Parra for her help during the data management phase and her comments on this manuscript. We thank David Tovey, editor in chief of the Cochrane Library, for agreeing to share data from Cochrane reviews; Javier Mayoral Campos, system administrator; the Cochrane Central Executive for preparing files; and all Cochrane reviewers who collected data. We also thank Laura Smales for English revision of the manuscript.

Contributors LB was involved in the study conception, selection of trials, data extraction, data analysis, interpretation of results and drafting the manuscript. PB was involved in the study conception, data analysis, interpretation of results and drafting the manuscript. IA was involved in the study conception, data extraction and drafting the manuscript. PR was involved in the study conception and drafting the manuscript. AD was involved in the study conception, selection of trials, data extraction, data analysis, interpretation of results and drafting the manuscript.

Funding This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no 676207.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Raw data and analyses are available on request from the authors.

REFERENCES

1. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997;126:376–80.
2. Hopewell S, Boutron I, Altman DG, *et al*. Incorporation of assessments of risk of bias of primary studies in systematic reviews of randomised trials: a cross-sectional study. *BMJ Open* 2013;3:e003342.
3. Jørgensen L, Paludan-Müller 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.
4. Hrobjartsson A, Boutron I, Turner L, *et al*. Assessing risk of bias in randomised clinical trials included in Cochrane Reviews: the why is easy, the how is a challenge. *Cochrane Database Syst Rev* 2013;4:ED000058.
5. Higgins JP, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
6. Schulz KF, Chalmers I, Hayes RJ, *et al*. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;273:408–12.
7. Moher D, Pham B, Jones A, *et al*. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998;352:609–13.
8. Balk EM, Bonis PA, Moskowitz H, *et al*. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. *JAMA* 2002;287:2973–82.
9. 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:e0159267.
10. 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.
11. Higgins JPT, Altman DG, Sterne JAC. Chapter 8: Assessing risk of bias in included studies. In: Higgins JP, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 (updated March 2011) ed*: The Cochrane Collaboration, 2011. www.handbook.cochrane.org.
12. Hartling L, Ospina M, Liang Y, *et al*. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009;339:b4012.
13. Hartling L, Bond K, Vandermeer B, *et al*. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011;6:e17242.
14. Hartling L, Hamm MP, Milne A, *et al*. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *J Clin Epidemiol* 2013;66:973–81.
15. Armijo-Olivo S, Ospina M, da Costa BR, *et al*. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. *PLoS One* 2014;9:e96920.
16. Armijo-Olivo S, Stiles CR, Hagen NA, *et al*. Assessment of study quality for systematic reviews: A comparison of the cochrane collaboration risk of bias tool and the effective public health practice project quality assessment tool: Methodological research. *J Eval Clin Pract* 2012;18:12–18.
17. da Costa BR, Beckett B, Diaz A, *et al*. Effect of standardized training on the reliability of the Cochrane risk of bias assessment tool: a prospective study. *Syst Rev* 2017;6:44.
18. Jordan VM, Lensen SF, Farquhar CM. There were large discrepancies in risk of bias tool judgments when a randomized controlled trial appeared in more than one systematic review. *J Clin Epidemiol* 2017;81:72–6.
19. Wilkins AJ. Risk of bias in assessing Risk of Bias. *Ophthalmic Physiol Opt* 2017;37:107–9.
20. *Review Manager (RevMan) [Computer program] [program]. Version 5.3 version*. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.
21. *R: A language and environment for statistical computing. [program]*. Vien: R Foundation for Statistical Computing, 2013.
22. *XML: Tools for Parsing and Generating XML Within R and S-Plus. [program]. 3.2.2 version*, 2017.
23. Dechartres A, Trinquart L, Atal I, *et al*. Evolution of poor reporting and inadequate methods over time in 20920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ* 2017;357:j2490.
24. deMelo VV. *Conference: Advances in Logic Based Intelligent Systems*, 2005.
25. *Stata Statistical Software: Release 13 [program*. College Station, TX: StataCorp LP, 2013.
26. Savović 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.
27. Propadalo I, Tranfic M, Vuka I, *et al*. In Cochrane reviews, risk of bias assessments for allocation concealment were frequently not in line with Cochrane's Handbook guidance. *J Clin Epidemiol* 2019;106.
28. Shamseer L, Hopewell S, Altman DG, *et al*. Update on the endorsement of CONSORT by high impact factor journals: a survey of journal "Instructions to Authors" in 2014. *Trials* 2016;17:301.
29. Turner L, Shamseer L, Altman DG, *et al*. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Syst Rev* 2012;1:60.
30. Turner L, Shamseer L, Altman DG, *et al*. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev* 2012;11:MR000030.
31. Altman DG, Moher D, Schulz KF. Improving the reporting of randomised trials: the CONSORT Statement and beyond. *Stat Med* 2012;31:2985–97.
32. RoB 2.0 Tool. https://sites.google.com/site/riskofbiastool//welcome/rob-2-0-tool (Accessed 22 Aug 2018).
33. Higgins JP, Sterne JA, Savovic J, *et al*. A revised tool for assessing risk of bias in randomized trials. *Cochrane Database of Systematic Reviews* 2016.