BMJ Open

Disagreements in risk of bias assessment for randomised controlled trials included in more than one Cochrane systematic reviews

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-028382
Article Type:	Research
Date Submitted by the Author:	05-Dec-2018
Complete List of Authors:	Bertizzolo, Lorenzo; INSERM, U 1153, Equipe Methods Bossuyt, Patrick; Academic Medical Center; University of Amsterdam, Dept. Clinical Epidemiology and Biostatistics Atal, Ignacio; INSERM U1153, Team Methods Ravaud, Philippe; INSERM, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS) Dechartres, Agnes; INSERM U738 H+&#65533;pital H+&#65533;telDieu Universit+&#65533; ParisDescartes, Centre dEpidemiologie Clinique
Keywords:	risk of bias, Cochrane, systematic reviews, interrater agreement, reproducibility

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

Lorenzo Bertizzolo[1], Patrick M Bossuyt[2], Ignacio Atal[1, 5], Philippe Ravaud[1, 3-6], Agnès Dechartres[1, 3-5, 7]

[1] INSERM, U1153 Epidemiology and Biostatistics Sorbonne Paris Cité Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases Team (METHODS), Paris, F-75004 France; Paris Descartes University, Sorbonne Paris Cité, France.

[2] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Netherlands.

[3] Centre d'Épidémiologie Clinique, Hôpital Hôtel Dieu, AP-HP (Assistance Publique des Hôpitaux de Paris), Paris, France.

[4] Faculté de Médecine, Université Paris Descartes, Sorbonne Paris Cité, Paris, France.

[5] Cochrane France, Paris, France

[6] Columbia University, Mailman School of Public Health, Department of Epidemiology, New York, USA

[7] Sorbonne Université, INSERM, Institut Pierre Louis de Santé Publique, Département Biostatistique, Santé Publique et Information Médicale, AP-HP, Hôpitaux Universitaires Pitié Salpêtrière – Charles Foix, Paris, France

Correspondence to: Lorenzo Bertizzolo, M.D., MPH

INSERM, U1153 - Centre d'Epidémiologie Clinique –

Hôpital Hôtel-Dieu,

1, place du parvis Notre Dame, 75004 Paris, France

Tel: +33 (0)1 42 34 78 25

E-mail: lorenzo.bertizzolo@gmail.com

Key-words: risk of bias, Cochrane, Systematic Reviews, Interrater Agreement, Reproducibility

Word Count:

Transparency declaration: The guarantor (Lorenzo Bertizzolo) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethics approval: Not applicable. This is a research on research study.

ABSTRACT (299 words)

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

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

Data sources: 2,796 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 (e.g., 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: 1,604 RCTs were included in more than one review. Proportion of agreement ranged from 57% (770/1,348 trials) for incomplete outcome data to 81% for random sequence generation (1,193/1,466). 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.

Article summary

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 were available to review authors and how they utilized 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 clinical 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), 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 randomised controlled trials 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 2,796 Cochrane reviews, which corresponds to all 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 v3.2.2[21] with the XML package[22]. The vocabulary used for risk of bias items slightly varied across reviews (e.g., 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 standardization was done for a previous publication[23].

Selection of eligible reviews

We excluded withdrawn or "empty" reviews (i.e., 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 situation concerned less than 10% of our included RCTs and was decided because of workload and to facilitate direct comparison of two assessments. 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 (i.e., "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 (i.e., 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 (e.g. "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 (e.g., "randomised, probably done"; "randomised, probably not done"; "study only mention randomization, but does not specify how randomization was performed; unclear"; "study states it is randomized; 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 formalized 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 (e.g., 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 the judgement cites information that is not available in the study report; additional sources are cited (e.g., 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 (e.g.. "Randomization 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 Supplementary Appendix 1, we report a figure synthetizing 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 (e.g., 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 (i.e., 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 13.1[25]. We decided to use simple percent 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 (ICC) 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 2,796 systematic reviews published between March 2011 and September 2014, 2,291 reviews included RCTs only and 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 1,604 single RCTs evaluated for the same risk of bias item in more than one review. The Supplementary Appendix 2 reports the frequency of the different Cochrane groups among those reviews.

Among the 1,604 selected RCTs: 1,603 had duplicate evaluation for allocation concealment, 1,466 for random sequence generation, 375 for blinding of participants and personnel, 583 for blinding of outcome assessment and 1,348 for incomplete outcome data.

Evaluation of agreement and distribution of disagreements

The agreement of risk of bias judgements ranged from 57% (770/1,348 trials) for incomplete outcome data to 81% (1,193/1,466 trials) for random sequence generation (Figure 2). We identified most disagreements for “low” and “unclear” risk of bias judgments, 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 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 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 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 categorization of disagreements and this explain the majority of cases where we were unable to categorize the source of disagreement ("unclear" source in Figure 3).

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 (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 randomization 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 (e.g., 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. 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 ten 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. 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 did not assess non-Cochrane reviews, even if they often use the Cochrane risk of bias tool. Agreement in these reviews is likely worse because of less familiarity and training with the tool. 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 Jorgensen et al.[3] and Savovic et al.[26]. In particular, Savovic et al.[26], surveying users of the risk of bias tool, reported on the possibility of confusion between random sequence generation and allocation 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, Jorgensen et al.[3], evaluating comments on the use of the risk of bias tool, highlighted how authors complained that judgment 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 differences. 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 Jorgensen et al.[3] and Savovic 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 randomization process, especially about what to do in some cases of incomplete information (e.g., randomization 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 1,600 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.

REFERENCES

1. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997;126(5):376-80. [published Online First: 1997/03/01]
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(8):e003342. doi: 10.1136/bmjopen-2013-003342 [published Online First: 2013/08/27]
3. 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: 2016/05/11]
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. doi: 10.1002/14651858.ED000058 [published Online First: 2013/06/04]
5. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343(oct18 2):d5928. doi: 10.1136/bmj.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(5):408-12. [published Online First: 1995/02/01]
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? *The Lancet* 1998;352(9128):609-13. doi: 10.1016/s0140-6736(98)01085-x
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(22):2973-82. [published Online First: 2002/06/08]
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(7):e0159267. doi: 10.1371/journal.pone.0159267 [published Online First: 2016/07/12]
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. doi: 10.1016/j.jclinepi.2016.04.005
11. Higgins JPT, Altman DG, Sterne JAC (editors). 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. Available from 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(oct19 1):b4012. doi: 10.1136/bmj.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(2):e17242. doi: 10.1371/journal.pone.0017242
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(9):973-81. doi: 10.1016/j.jclinepi.2012.07.005
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(5):e96920. doi: 10.1371/journal.pone.0096920 [published Online First: 2014/05/16]
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(1):12-8. doi: 10.1111/j.1365-2753.2010.01516.x
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(1):44. doi: 10.1186/s13643-017-0441-7

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;81:72-76. doi: 10.1016/j.jclinepi.2016.08.012 [published Online First: 2016/09/14]
19. Wilkins AJ. Risk of bias in assessing Risk of Bias. *Ophthalmic Physiol Opt* 2017;37(1):107-09. doi: 10.1111/opo.12333
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 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ* 2017;357:j2490. doi: 10.1136/bmj.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. 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: 2014/04/16]
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* 2018 doi: 10.1016/j.jclinepi.2018.10.002
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(1):301. doi: 10.1186/s13063-016-1408-z
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. doi: 10.1186/2046-4053-1-60
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. doi: 10.1002/14651858.MR000030.pub2
31. Altman DG, Moher D, Schulz KF. Improving the reporting of randomised trials: the CONSORT Statement and beyond. *Stat Med* 2012;31(25):2985-97. doi: 10.1002/sim.5402
32. RoB 2.0 Tool [Available from: https://sites.google.com/site/riskofbiastool//welcome/rob-2-0-tool accessed 22/08/2018.
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(10 (Suppl 1)) doi: dx.doi.org/10.1002/14651858.CD201601

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.

CONTRIBUTIONS:

Lorenzo Bertizzolo was involved in the study conception, selection of trials, data extraction, data analysis, interpretation of results and drafting the manuscript.

Patrick Bossuyt was involved in the study conception, data analysis, interpretation of results and drafting the manuscript.

Ignacio Atal was involved in the study conception, data extraction, and drafting the manuscript.

Philippe Ravaud was involved in the study conception and drafting the manuscript.

Agnès Dechartres was involved in the study conception, selection of trials, data extraction, data analysis, interpretation of results and drafting the manuscript.

Lorenzo Bertizzolo is the guarantor. He had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

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:

The authors declare that they have no competing interests in relation to this study.

DATA SHARING:

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

List of tables

List of figures

List of appendices

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 number list."*; Low risk of bias
	Confusion in the definition of the item	19 (13)	*"Researchers attempted to contact all patients seen by physicians during one 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

[+] 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.

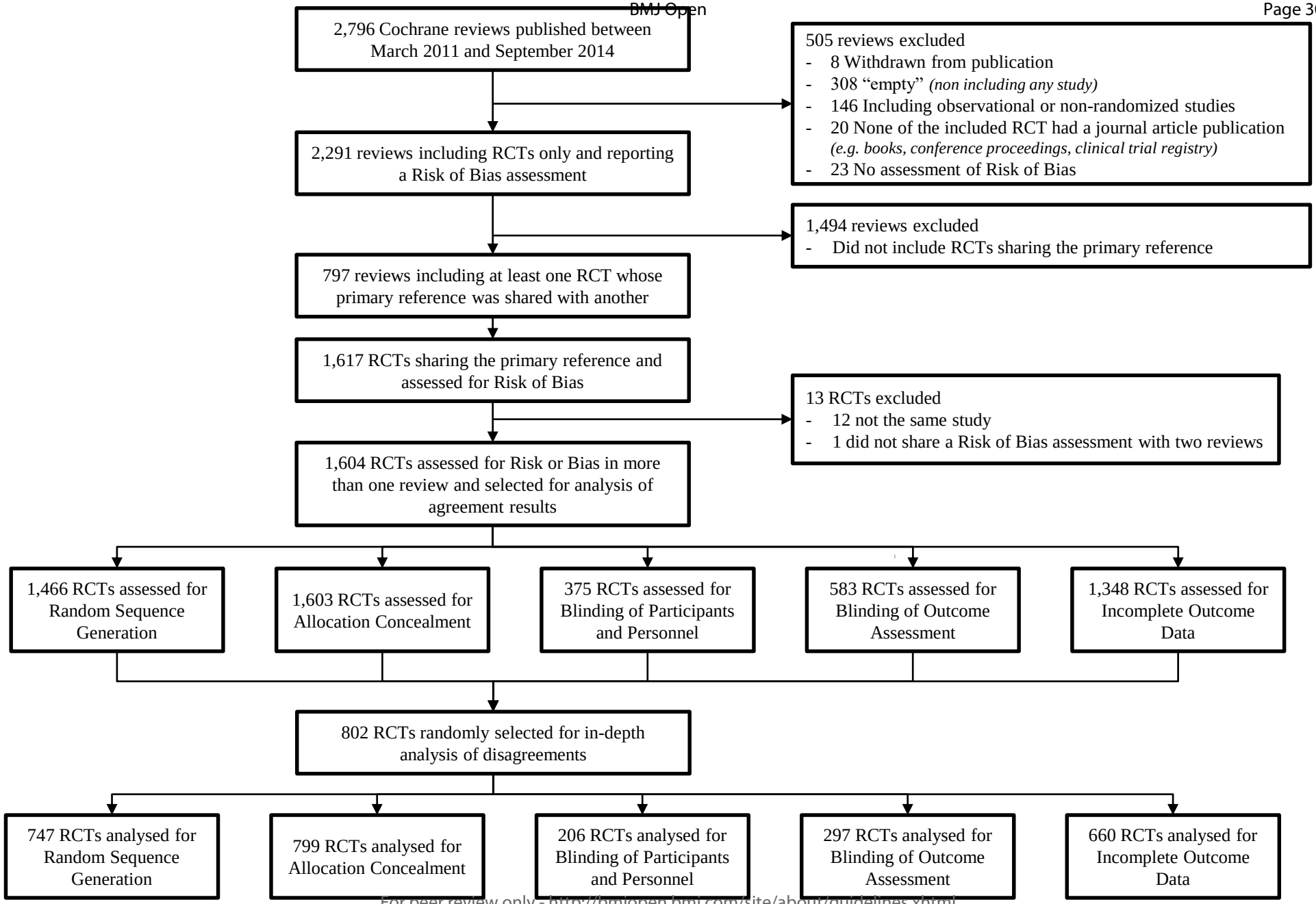
Risk of bias Item	Main reasons for disagreements	N (%)[+]	Examples of support for judgement from the review*
blinding of participants and personnel	Assess risk differently if blinding was not feasible because of the type of intervention	20 (36)	*"Not possible to blind participants"*; Low risk of bias *"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

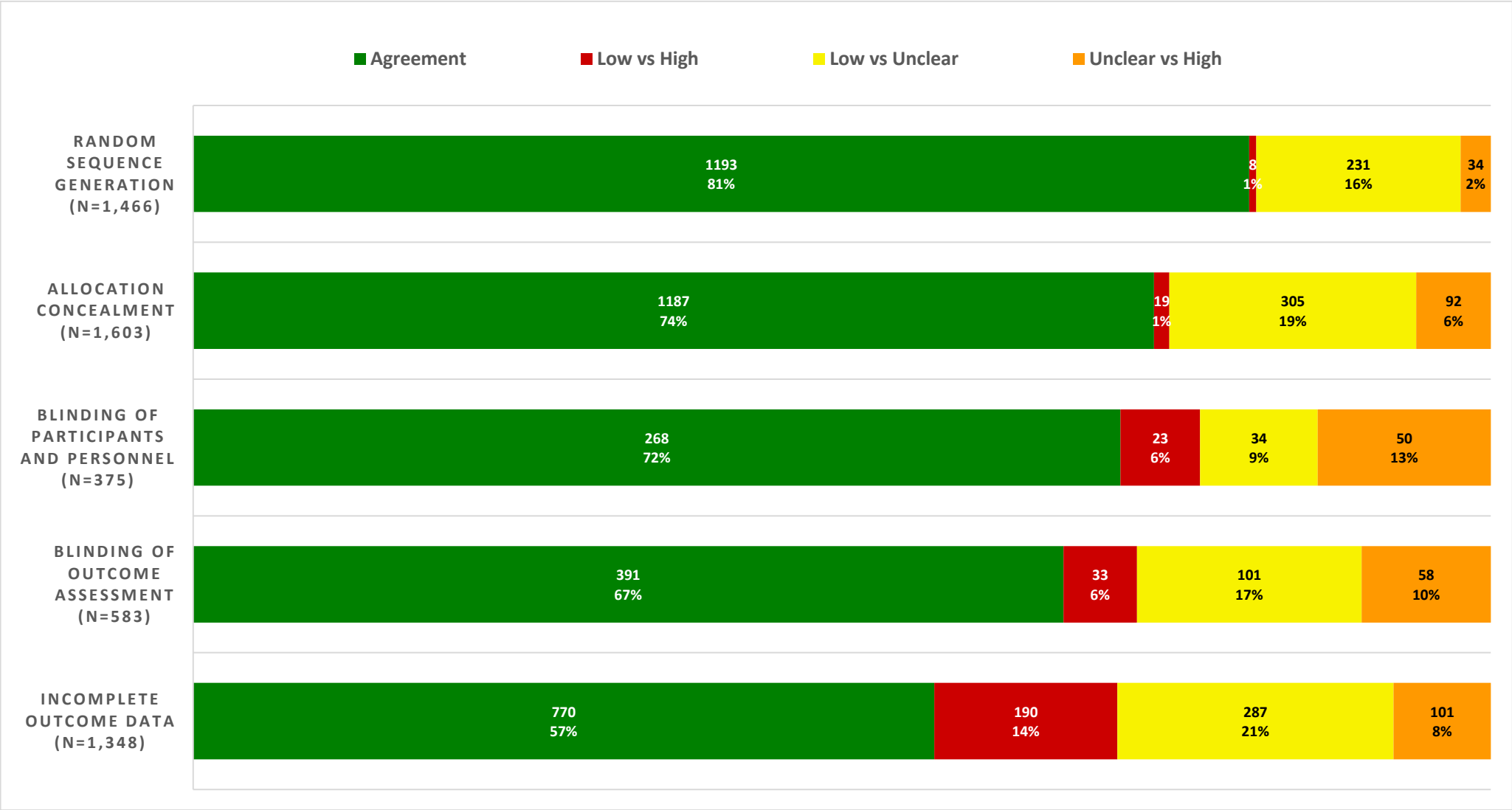
[+] Number of RCTs disagreeing for this reason; percentage over the total of disagreements for different interpretation.

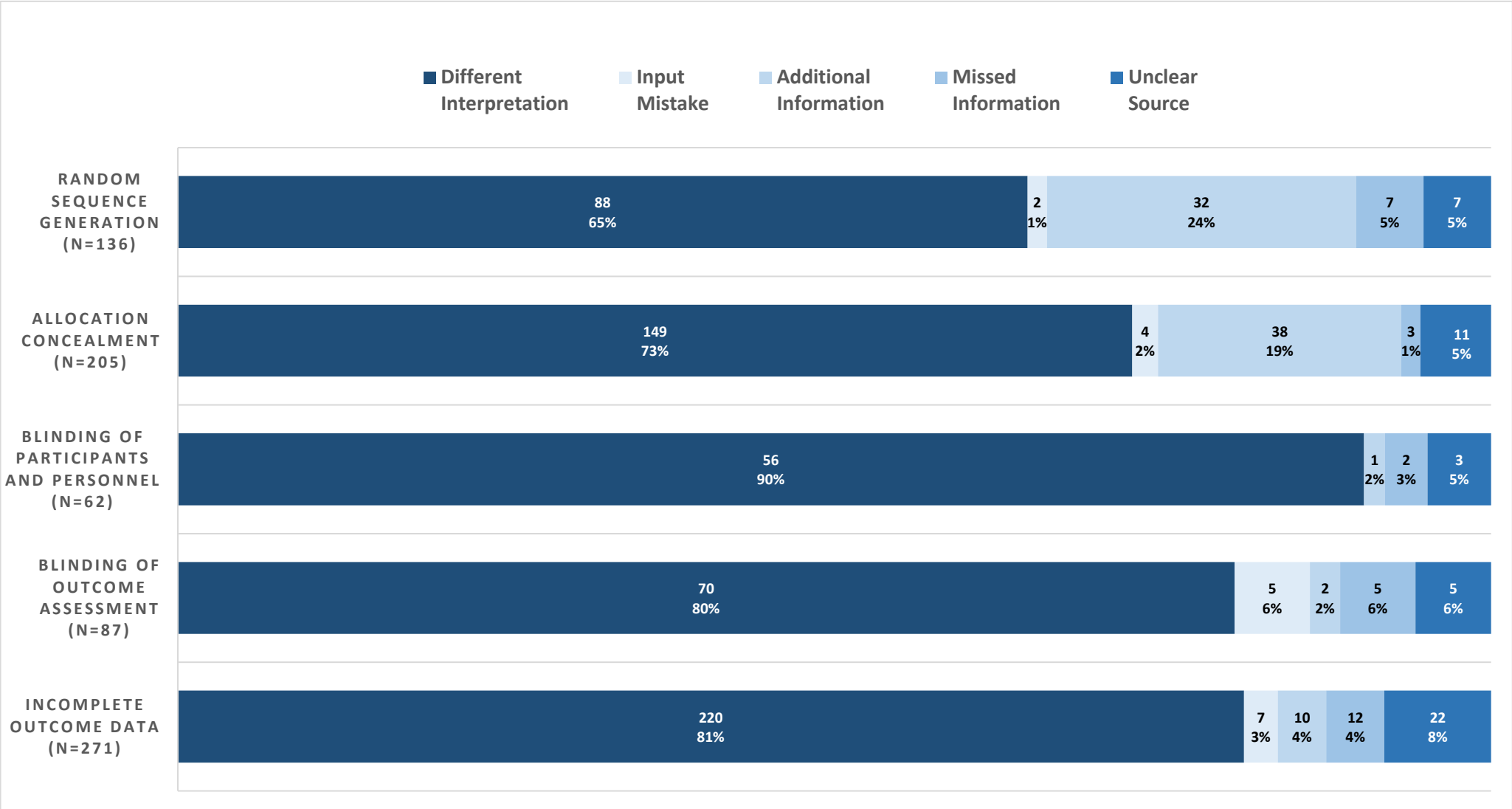
* When two extracts are reported, they refer to the same study.

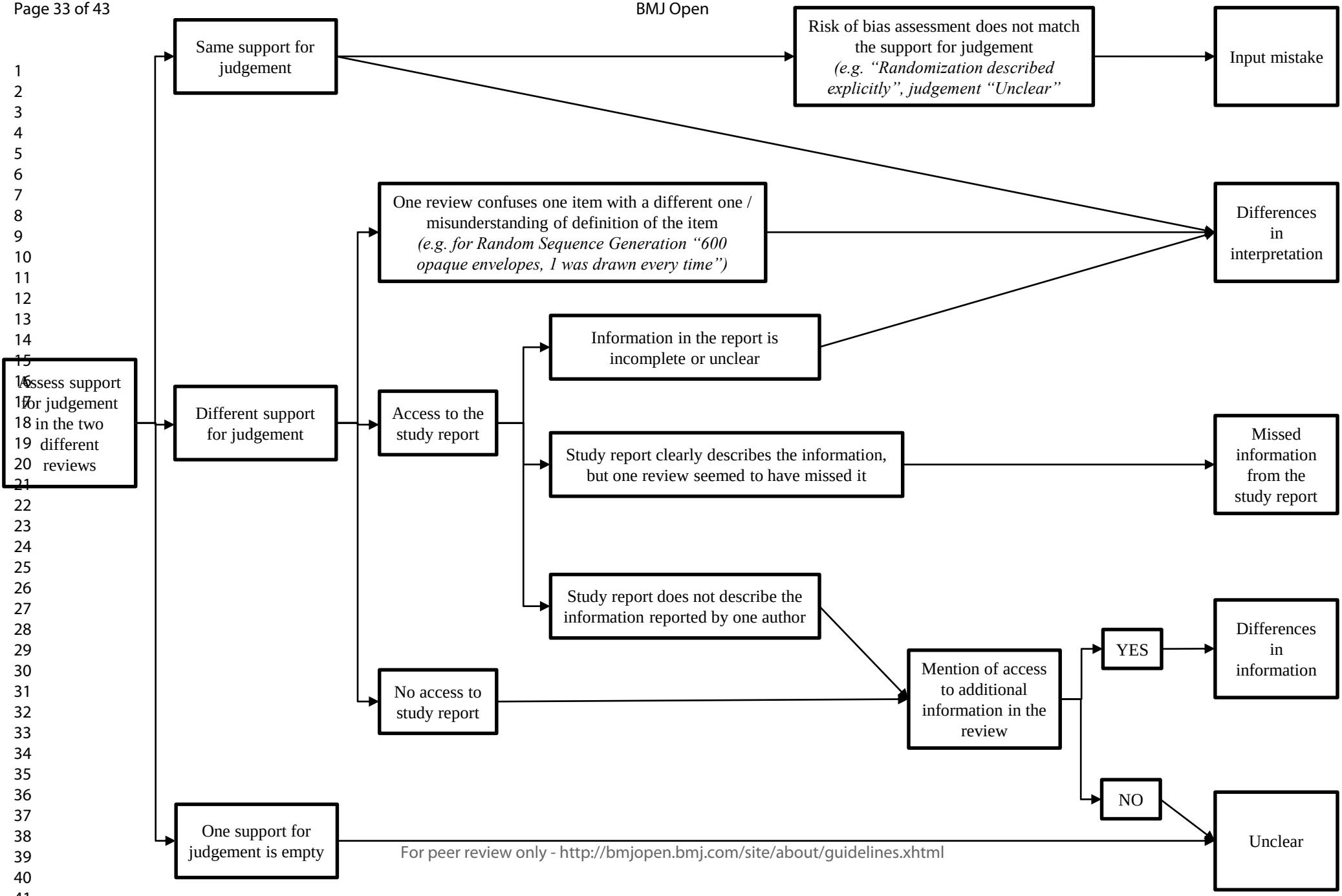
Risk of bias Item	Main reasons for disagreements	N (%)[+]	Examples of support for judgement from the review*
incomplete outcome data	Use different cut-off for the rate of missing data	57 (26)	*"11 withdrawals (10%)."*; Low risk of bias *"Comment: there were post-randomisation drop-outs"*; High risk of bias
	Focus on number vs 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 dropouts 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 (e.g. miscarriage). 1120 in sample. 51/1287 non-responders were included as continuing smokers."* High risk of bias
	Consider differently intention-to-treat analysis	25 (11)	*"147 randomised; 4 in the letrozole group and 3 in the LOD 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 noncompliance"*; 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 dropouts"* Low risk of bias
	Consider differently imputation of missing data	20 (9)	*"Imputation method not described"*; Unclear risk of bias *"Dropout 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)	*"Dropout 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.


2,796 Cochrane reviews published between March 2011 and September 2014
505 reviews excluded
- 8 Withdrawn from publication
- 308 "empty" (non including any study)
- 146 Including observational or non-randomized studies
- 20 None of the included RCT had a journal article publication (e.g. books, conference proceedings, clinical trial registry)
- 23 No assessment of Risk of Bias
2,291 reviews including RCTs only and reporting a Risk of Bias assessment
1,494 reviews excluded
- Did not include RCTs sharing the primary reference
797 reviews including at least one RCT whose primary reference was shared with another
1,617 RCTs sharing the primary reference and assessed for Risk of Bias
13 RCTs excluded
- 12 not the same study
- 1 did not share a Risk of Bias assessment with two reviews
1,604 RCTs assessed for Risk or Bias in more than one review and selected for analysis of agreement results
1,466 RCTs assessed for Random Sequence Generation
1,603 RCTs assessed for Allocation Concealment
375 RCTs assessed for Blinding of Participants and Personnel
583 RCTs assessed for Blinding of Outcome Assessment
1,348 RCTs assessed for Incomplete Outcome Data
802 RCTs randomly selected for in-depth analysis of disagreements
747 RCTs analysed for Random Sequence Generation
799 RCTs analysed for Allocation Concealment
206 RCTs analysed for Blinding of Participants and Personnel
297 RCTs analysed for Blinding of Outcome Assessment
660 RCTs analysed for Incomplete Outcome Data

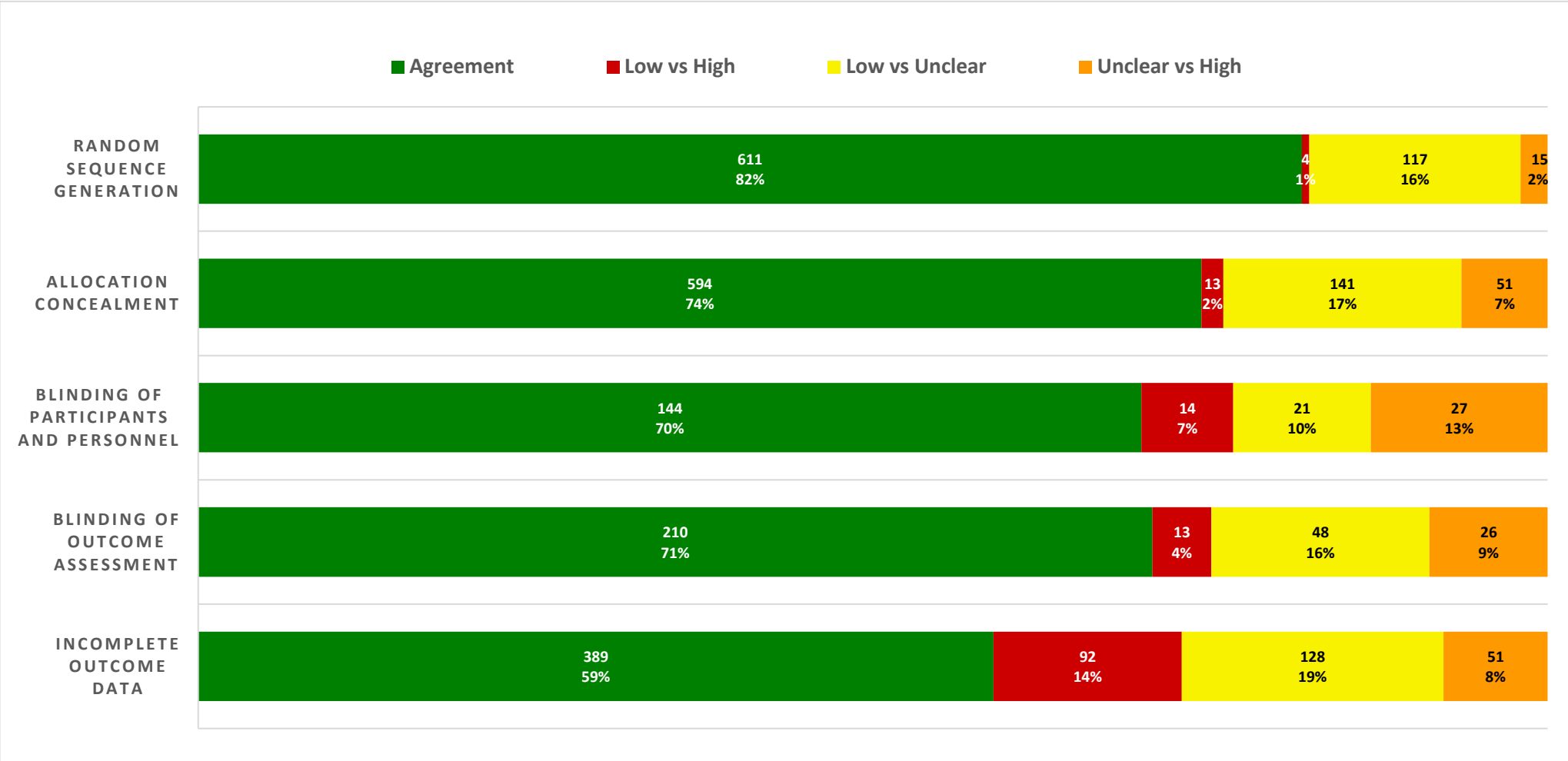

Agreement
Low vs High
Low vs Unclear
Unclear vs High
RANDOM SEQUENCE GENERATION (N=1,466)
1193 81%
8 1%
231 16%
34 2%
ALLOCATION CONCEALMENT (N=1,603)
1187 74%
19 1%
305 19%
92 6%
BLINDING OF PARTICIPANTS AND PERSONNEL (N=375)
268 72%
23 6%
34 9%
50 13%
BLINDING OF OUTCOME ASSESSMENT (N=583)
391 67%
33 6%
101 17%
58 10%
INCOMPLETE OUTCOME DATA (N=1,348)
770 57%
190 14%
287 21%
101 8%


Different Interpretation
Input Mistake
Additional Information
Missed Information
Unclear Source
RANDOM SEQUENCE GENERATION (N=136)
88 65%
2 1%
32 24%
7 5%
7 5%
ALLOCATION CONCEALMENT (N=205)
149 73%
4 2%
38 19%
3 1%
11 5%
BLINDING OF PARTICIPANTS AND PERSONNEL (N=62)
56 90%
1 2%
2 3%
3 5%
BLINDING OF OUTCOME ASSESSMENT (N=87)
70 80%
5 6%
2 2%
5 6%
5 6%
INCOMPLETE OUTCOME DATA (N=271)
220 81%
7 3%
10 4%
12 4%
22 8%


Assess support for judgement in the two different reviews
Same support for judgement
Different support for judgement
One support for judgement is empty
Risk of bias assessment does not match the support for judgement (e.g. "Randomization described explicitly", judgement "Unclear"
Input mistake
One review confuses one item with a different one / misunderstanding of definition of the item (e.g. for Random Sequence Generation "600 opaque envelopes, 1 was drawn every time")
Differences in interpretation
Access to the study report
Information in the report is incomplete or unclear
Study report clearly describes the information, but one review seemed to have missed it
Missed information from the study report
Study report does not describe the information reported by one author
No access to study report
Mention of access to additional information in the review
YES
NO
Differences in information
Unclear

Cochrane Group	Number of reviews	% on the total
Pregrancy and Childbirth	93	11.7%
Airways	48	6.0%
Pain, Palliative and Supportive Care Group	42	5.3%
Acute Respiratory Infections	37	4.6%
Gynaecology and Fertility	29	3.6%
Neonatal	29	3.6%
Tobacco Addiction	27	3.4%
Stroke	25	3.1%
Gynaecological, Neuro-oncology and Orphan Cancer Group	23	2.9%
Wounds	23	2.9%
Hepato-Biliary	22	2.8%
Cystic Fibrosis and Genetic Disorders	21	2.6%
Anaethesia	20	2.5%
Drugs and Alcohol	20	2.5%
Neuromuscolar	19	2.4%
Common Mental Disorders	18	2.3%
Fertility Regulation	18	2.3%
Heart	17	2.1%
Developmental, Psychosocial and Learning Problems	16	2.0%
Incontinence	16	2.0%
Kidney disease	16	2.0%
Schizophrenia	16	2.0%
Infectious Diseases	15	1.9%
Oral Health	14	1.8%
Vascular	14	1.8%
Dementia and Cognitive Improvement	13	1.6%
Musculoskeletal	12	1.5%
Consumers and Communication	10	1.3%
Epilepsy	10	1.3%
Eyes and Vision	9	1.1%
Metabolic and Endocrine Disorders	9	1.1%
Back and Neck	8	1.0%
Hypertension	8	1.0%
Multiple Sclerosis	8	1.0%
Effective Practice and Organisation of Care	7	0.9%
HIV/AIDS	7	0.9%
Infiammatory Bowel Disease	7	0.9%
Injuries	7	0.9%
Bone, Joint and Muscle Trauma Group	6	0.8%
ENT	6	0.8%
Haematological Malignancies	6	0.8%

Cochrane Group	Number of reviews	% on the total
Breast Cancer	5	0.6%
Colorectal Cancer	5	0.6%
Lung Cancer	3	0.4%
Movement Disorders	3	0.4%
Skin	3	0.4%
Occupational Health	2	0.3%
Sexually Transmitted Infections	2	0.3%
Public Health	1	0.1%
Upper GI and Pancreatic Diseases	1	0.1%
Urology	1	0.1%
Total	797	


Agreement
Low vs High
Low vs Unclear
Unclear vs High
RANDOM SEQUENCE GENERATION
611 82%
4 1%
117 16%
15 2%
ALLOCATION CONCEALMENT
594 74%
13 2%
141 17%
51 7%
BLINDING OF PARTICIPANTS AND PERSONNEL
144 70%
14 7%
21 10%
27 13%
BLINDING OF OUTCOME ASSESSMENT
210 71%
13 4%
48 16%
26 9%
INCOMPLETE OUTCOME DATA
389 59%
92 14%
128 19%
51 8%

Supplementary Appendix 4| Examples of in-depth analysis of disagreements conducted with the support of the study report

Risk of bias item	Study Name	Support for judgement*	Information in the study report^	Category of disagreement	Reason of disagreement
Random sequence generation	ABCD 2004	Review 4136: Generated the randomisation list using SAS, stratified by sex and SCr; Low Risk	*"The (...) statistician generated the randomization list using SAS (...) stratified by sex and baseline serum creatinine concentration (...)."*	Missed information from the study report	
		Review 8277: Method not reported; Unclear Risk			
	Cho 2006	Review 7566: Stated that it is a quasi-randomized study but details not given; High Risk	*"... using a quasi-experimental design with a non-equivalent control group."* *"They were randomly assigned to participate in the experimental group (...) or a waiting-list control group (...)."*	Different interpretation	Consider differently incomplete or unclear description
		Review 9553: Participants randomly allocated to treatment or control group; Unclear Risk			
	Petersen 2005	Review 9132: Quote: "[P]atients were randomly assigned..."Quote: "We used an adaptive allocation scheme for the treatment assignment, with the MMSE score, age and APOE e4 status as balancing covariates"; Low Risk	*"We used an adaptive allocation scheme for the treatment assignment, with the MMSE score, age, and APOE e 4 status as balancing covariates."*	Different interpretation	Confusion or misknowledge
		Review 7176: The trial is described as randomised, but the method of sequence generation was not specified. Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Allocation concealment	Burge 2000	Review 2991: Participants were randomly assigned sequentially from a list comprising treatment numbers only; Low Risk	*"We used a computer generated allocation schedule stratified by centre (block size of six). Patients were randomised sequentially from a list comprising treatment numbers only".*	Different interpretation	Consider differently incomplete or unclear description
		Review 10115: Information not available; Unclear Risk			
	McMurdo 1993	Review 4294: Quote: "Randomisation was by opening sealed envelopes supplied in sequence by the study co-ordinator; Low Risk	*"Randomization was by opening sealed envelopes supplied in sequence by the study co-ordinator (...), and prepared from a computer-generated random numbers table."*	Different interpretation	Consider differently envelopes description
		Review 4963: Unclear, insufficient reporting to permit judgement; Unclear Risk			
	Draper 2007	Review 8179: "... and alternating between treatment or wait list control groups."; High Risk	*"On each occasion that a least eight patients had been recruited, their names were selected at random by a blinded investigator to be allocated alternately to the immediate treatment group or a wait-list control group."*	Different interpretation	Consider differently incomplete or unclear description
		Review 1919: "Reported as concealed but specific method for concealment not reported"; Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Blinding of participants and personnel	Nielsen 2006	Review 9672: "Double-blind"; Low Risk	*"This study was a randomized, placebo-controlled, double blind, Danish, multi-center (two centers) study."* *"The treatment was applied by a nasal spray with one puff in each nostril every day either in the morning or evening."*	Different information	One review accessed additional data through another study report
		Review 4143: Although "All treatments were supplied as identical intranasal sprays..." the 2004 publication describes a higher rate of withdrawal due to adverse effects in the intervention groups [11.7% in the placebo group, 21.7% in the 150 gm group and 28.7% in the 300 gm group} which may have affected blinding status; Unclear Risk			
	Gersel 1979	Review 10562: Described as double-blind [presumed participants and personnel/investigators]; Low Risk	*"A double-blind experimental design was used, employing each patient as his own control."*	Different interpretation	Consider differently information of "double blind"
		Review 6968: Not mentioned and no information to suggest this was done.; Unclear Risk			
	Stein 2011	Review 7025: Not possible to blind participants to intervention. Insufficient information to make a judgement about blinding of therapists; High Risk	*"Follow-up assessment was made 3 months after release (research staff conducting assessments were blind to treatment assignment)."* *"Randomization was accomplished via random numbers table in advance and placed in an envelope by the project coordinator. Following baseline assessment, research staff opened the envelope to learn of intervention assignment."*	Different interpretation	Consider differently incomplete or unclear description
		Review 10901: Researchers were blind until after the baseline assessment. Participants were not blinded.; Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Blinding of outcome assessment	Schoen 2007	Review 3603: Outcome assessor was not blinded.; High risk	*"In total 72 patients were screened by a maxillofacial surgeon (PJS) and prosthodontist (HR)." "All clinical assessments were performed by the investigator (PJS) who was not involved in treatment of the patients."*	Different interpretation	Consider differently incomplete or unclear description
		Review 5005: Outcome assessor may have been unaware of allocation: "All clinical assessments were performed by the investigator (PJS) who was not involved in treatment of the patients."; Low risk			
	Geroin 2011	Review 6185: Not done; High risk	*"All patients were evaluated by the same examiner (an experienced internal coworker) who was not aware of the treatment received by the patients."*	Missed information from the study report	
		Review 9645: Quote: "All patients were evaluated by the same examiner (an experienced internal coworker) who was not aware of the treatment received by the patients"; Low Risk			
	McCambridge 2004	Review 8969: As one interventionist was the study PI, a second independent interviewer who was blind to study condition was employed to conduct 3 month follow-ups, and an additional interviewer who was blind to initial group allocation was employed for 12 months follow-ups; Low Risk	*"further area of possible bias was that intervention recipients might report more favourable outcome data to the researcher who had delivered the intervention (J.M.). To study any such bias, a second independent interviewer who was blind to study condition, was employed to interview a sample of participants."*	Different interpretation	Consider differently incomplete or unclear description
		Review 7025: A second independent interviewer who was blind to study condition was employed to interview a sample of participants, though not all participants; Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Incomplete outcome data	Altmaier 1992	Review 1822: All subjects recorded follow-up data; Low Risk	*[From table] "The n = 21 for control group and n = 24 for psychological group on all process measures." [From table] The n = 21 for each group at each assessment.]*	Different interpretation	Consider differently incomplete or unclear description
		Review 7407: Inadequately reported; High Risk			
	Killen 1984	Review 146: 11/75 recruited dropped out before full treatment, and are excluded from analyses.; Low Risk	*"The first 75 were accepted into the study. Seven failed to attend (...) two dropped (...). The final sample (N = 64)."*	Missed information from the study report	
		Review 3999: Losses to follow-up not reported, all participants included; Unclear Risk			
	Creager 2008	Review 986: There was a huge loss to follow up (only 50% completed the 6 month follow up) in this study and therefore there is a high risk of attrition bias; High Risk	*"The remaining 525 patients met the inclusion criteria (...) The remaining 430 patients met their criteria for randomization (...).The ITT population consisted of 370 randomized patients (...). The per-protocol patient population consisted of 214 randomized patients"*	Different interpretation	Consider differently intention-to-treat analysis
		Review 5262: Unclear why of patients stopped medication, unclear whether data presented represents intention-to-treat or per-protocol analysis			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Supplementary Appendix 5| Reasons for disagreements in cases of a different interpretation of the same information; focus on "low" versus "high" disagreements.

Risk of bias item	Main reasons for disagreements	Examples
random sequence generation	Consider differently incomplete or unclear description	*"The names of communities within each group of three were written on individual cards, mixed and selected randomly: the first from each group was assigned to arm A (IEC alone), the second to arm B (IEC and STI management) and the third to arm C";* Low risk of bias *"Names of communities within each triplet were written on separate cards and shuffled.";* High risk of bias
allocation concealment	Consider differently envelopes description	*"Closed envelopes";* Low risk of bias *"Closed envelopes, although not opaque.";* High risk of bias
	Confusion in the definition of the item	*"pg. 2 - Methods - randomisation was done centrally to preserve allocation concealment"*; Low risk of bias *"904 patients were eligible for the study. 446 patients were randomised (49%). Due to the number of patients declining screening, there is an increased risk of inclusion bias."*; High risk of bias
	Confusion with blinding	*"States used "preprogrammed laptop computer". Remote site";* Low risk of bias *"participants were told to which compound they had been allocated."; High risk of bias*
blinding of participants and personnel	Assess risk differently if blinding was not feasible because of the type of intervention	*"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	*"Not possible to blind but most of the outcomes not likely to be influenced by lack of blinding.";* Low risk of bias *"Not blinded due to nature of intervention."*; High risk of bias
	Confusion with allocation concealment	*"participants were randomly allocated to either intervention or control group by an independent party";* Low risk of bias *"Control group did not receive the comparable non-exercise related attention to the intervention group"*; High risk of bias

Risk of bias Item	**Main reasons for disagreements**	**Examples**
blinding of outcome assessment	outcome considered not influenced by blinding	*"No information given about whether patients or assessors were blind to physician allocation but primary outcomes (treatment outcome and patient reported physician cultural competency) judged unlikely to be affected by lack of blinding"*; Low risk of bias *"Unblinded."*; High risk of bias
	Consider differently patient reported outcomes when patients are blinded or not to the intervention	*"Insufficient information available to assess"*; Low risk of bias *"Comment: depression assessed by patient self-report"*; High risk of bias
	Assess risk differently if blinding was not feasible because of the type of intervention	*"Unclear blinding of outcome assessment"*; Low risk of bias *"blinding not possible due to intervention"*; High risk of bias
incomplete outcome data	Use different cut-off for the rate of missing data	*"11 withdrawals (10%)."*; Low risk of bias *"Comment: there were post-randomisation drop-outs"*; High risk of bias
	Focus on number vs reasons/precise report of missing data	*"Numbers and reasons for dropouts and withdrawals in all intervention groups were described."*; Low risk of bias *"20 drop-outs (27.2%) with 4 deaths (3 males, 1 female) from cardiovascular events"*; High risk of bias
	Consider differently incomplete or unclear description	*"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 (e.g. miscarriage). 1120 in sample. 51/1287 non-responders were included as continuing smokers."* High risk of bias
	Use different cut-off for difference in the rate missing data between different arms/comparisons	*"Similar rates of withdrawal between arms. Withdrawals: 36 BUD, 51 placebo"*; Low risk of bias *"Dropout higher in placebo group (35% vs 25% in budesonide group). ITT used."*; High risk of bias

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

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-028382.R1
Article Type:	Research
Date Submitted by the Author:	14-Feb-2019
Complete List of Authors:	Bertizzolo, Lorenzo; INSERM, U 1153, Equipe Methods Bossuyt, Patrick; Academic Medical Center; University of Amsterdam, Dept. Clinical Epidemiology and Biostatistics Atal, Ignacio; INSERM U1153, Team Methods Ravaud, Philippe; INSERM, U1153, Epidemiology and Biostatistics Sorbonne Paris Cite Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases team (METHODS) Dechartres, Agnes; INSERM U738 H+&#65533;pital H+&#65533;telDieu Universit+&#65533; ParisDescartes, Centre dEpidemiologie Clinique
Primary Subject Heading:	Epidemiology
Secondary Subject Heading:	Evidence based practice, Public health, Qualitative research, Research methods
Keywords:	risk of bias, Cochrane, systematic reviews, interrater agreement, reproducibility, Public health < INFECTIOUS DISEASES

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], Patrick M Bossuyt[2], Ignacio Atal[1, 5], Philippe Ravaud[1, 3-6], Agnès Dechartres[7]

[1] INSERM, U1153 Epidemiology and Biostatistics Sorbonne Paris Cité Research Center (CRESS), Methods of therapeutic evaluation of chronic diseases Team (METHODS), Paris, F-75004 France; Paris Descartes University, Sorbonne Paris Cité, France.

[2] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, Netherlands.

[3] Centre d'Épidémiologie Clinique, Hôpital Hôtel Dieu, AP-HP (Assistance Publique des Hôpitaux de Paris), Paris, France.

[4] Faculté de Médecine, Université Paris Descartes, Sorbonne Paris Cité, Paris, France.

[5] Cochrane France, Paris, France

[6] Columbia University, Mailman School of Public Health, Department of Epidemiology, New York, USA

[7] Sorbonne Université, INSERM, Institut Pierre Louis de Santé Publique, Département Biostatistique, Santé Publique et Information Médicale, AP-HP, Hôpitaux Universitaires Pitié Salpêtrière – Charles Foix, Paris, France

Correspondence to: Lorenzo Bertizzolo, M.D., MPH

INSERM, U1153 - Centre d'Epidémiologie Clinique –

Hôpital Hôtel-Dieu,

1, place du parvis Notre Dame, 75004 Paris, France

Tel: +33 (0)1 42 34 78 25

E-mail: lorenzo.bertizzolo@gmail.com

Key-words: risk of bias, Cochrane, Systematic Reviews, Interrater Agreement, Reproducibility

Word Count: 3800

Transparency declaration: The guarantor (Lorenzo Bertizzolo) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Ethics approval: Not applicable. This is a research on research study.

ABSTRACT (299 words)

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

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

Data sources: 2,796 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 (e.g., 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: 1,604 RCTs were included in more than one review. Proportion of agreement ranged from 57% (770/1,348 trials) for incomplete outcome data to 81% for random sequence generation (1,193/1,466). 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.

Article summary

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 were available to review authors and how they utilized 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 clinical 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), 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 randomised controlled trials 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 2,796 Cochrane reviews, which corresponds to all 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 v3.2.2[21] with the XML package[22]. The vocabulary used for risk of bias items slightly varied across reviews (e.g., 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 standardization was done for a previous publication[23].

Selection of eligible reviews

We excluded withdrawn or "empty" reviews (i.e., 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 situation concerned less than 10% of our included RCTs and was decided because of workload and to facilitate direct comparison of two assessments. 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 (i.e., "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 (i.e., 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 (e.g. "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 (e.g., "randomised, probably done"; "randomised, probably not done"; "study only mention randomization, but does not specify how randomization was performed; unclear"; "study states it is randomized; 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 formalized 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 (e.g., 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 the judgement cites information that is not available in the study report; additional sources are cited (e.g., 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 (e.g.. "Randomization 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 Supplementary Appendix 1, we report a figure synthetizing 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 (e.g., 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 (i.e., 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 13.1[25]. We decided to use simple percent 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 (ICC) 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 2,796 systematic reviews published between March 2011 and September 2014, 2,291 reviews included RCTs only and 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 1,604 single RCTs evaluated for the same risk of bias item in more than one review. The Supplementary Appendix 2 reports the frequency of the different Cochrane groups among those reviews.

Among the 1,604 selected RCTs: 1,603 had duplicate evaluation for allocation concealment, 1,466 for random sequence generation, 375 for blinding of participants and personnel, 583 for blinding of outcome assessment and 1,348 for incomplete outcome data.

Evaluation of agreement and distribution of disagreements

The agreement of risk of bias judgements ranged from 57% (770/1,348 trials) for incomplete outcome data to 81% (1,193/1,466 trials) for random sequence generation (Figure 2). We identified most disagreements for “low” and “unclear” risk of bias judgments, 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 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 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 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 categorization of disagreements and this explain the majority of cases where we were unable to categorize the source of disagreement ("unclear" source in Figure 3).

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 (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 randomization 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 (e.g., 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 organization, therefore we expected results and procedures to be more appropriately comparable. Authors compiling Cochrane reviews are members of the organization 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 ten 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 Jorgensen et al.[3] and Savovic et al.[26]. In particular, Savovic et al.[26], surveying users of the risk of bias tool, reported on the possibility of confusion between random sequence generation and allocation 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, Jorgensen et al.[3], evaluating comments on the use of the risk of bias tool, highlighted how authors complained that judgment 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 differences. 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 Jorgensen et al.[3] and Savovic 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 randomization process, especially about what to do in some cases of incomplete information (e.g., randomization 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 1,600 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.

REFERENCES

1. Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997;126(5):376-80. [published Online First: 1997/03/01]
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(8):e003342. doi: 10.1136/bmjopen-2013-003342 [published Online First: 2013/08/27]
3. 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: 2016/05/11]
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. doi: 10.1002/14651858.ED000058 [published Online First: 2013/06/04]
5. Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343(oct18 2):d5928. doi: 10.1136/bmj.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(5):408-12. [published Online First: 1995/02/01]
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? *The Lancet* 1998;352(9128):609-13. doi: 10.1016/s0140-6736(98)01085-x
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(22):2973-82. [published Online First: 2002/06/08]
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(7):e0159267. doi: 10.1371/journal.pone.0159267 [published Online First: 2016/07/12]
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. doi: 10.1016/j.jclinepi.2016.04.005
11. Higgins JPT, Altman DG, Sterne JAC (editors). 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. Available from 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(oct19 1):b4012. doi: 10.1136/bmj.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(2):e17242. doi: 10.1371/journal.pone.0017242
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(9):973-81. doi: 10.1016/j.jclinepi.2012.07.005
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(5):e96920. doi: 10.1371/journal.pone.0096920 [published Online First: 2014/05/16]
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(1):12-8. doi: 10.1111/j.1365-2753.2010.01516.x
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(1):44. doi: 10.1186/s13643-017-0441-7

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;81:72-76. doi: 10.1016/j.jclinepi.2016.08.012 [published Online First: 2016/09/14]
19. Wilkins AJ. Risk of bias in assessing Risk of Bias. *Ophthalmic Physiol Opt* 2017;37(1):107-09. doi: 10.1111/opo.12333
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 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ* 2017;357:j2490. doi: 10.1136/bmj.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. 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: 2014/04/16]
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* 2018 doi: 10.1016/j.jclinepi.2018.10.002
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(1):301. doi: 10.1186/s13063-016-1408-z
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. doi: 10.1186/2046-4053-1-60
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. doi: 10.1002/14651858.MR000030.pub2
31. Altman DG, Moher D, Schulz KF. Improving the reporting of randomised trials: the CONSORT Statement and beyond. *Stat Med* 2012;31(25):2985-97. doi: 10.1002/sim.5402
32. RoB 2.0 Tool [Available from: https://sites.google.com/site/riskofbiastool//welcome/rob-2-0-tool accessed 22/08/2018.
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(10 (Suppl 1)) doi: dx.doi.org/10.1002/14651858.CD201601

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.

CONTRIBUTIONS:

Lorenzo Bertizzolo was involved in the study conception, selection of trials, data extraction, data analysis, interpretation of results and drafting the manuscript.

Patrick Bossuyt was involved in the study conception, data analysis, interpretation of results and drafting the manuscript.

Ignacio Atal was involved in the study conception, data extraction, and drafting the manuscript.

Philippe Ravaud was involved in the study conception and drafting the manuscript.

Agnès Dechartres was involved in the study conception, selection of trials, data extraction, data analysis, interpretation of results and drafting the manuscript.

Lorenzo Bertizzolo is the guarantor. He had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

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:

The authors declare that they have no competing interests in relation to this study.

DATA SHARING:

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

List of tables

List of figures

List of appendices

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 number list."*; Low risk of bias
	Confusion in the definition of the item	19 (13)	*"Researchers attempted to contact all patients seen by physicians during one 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

[+] 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.

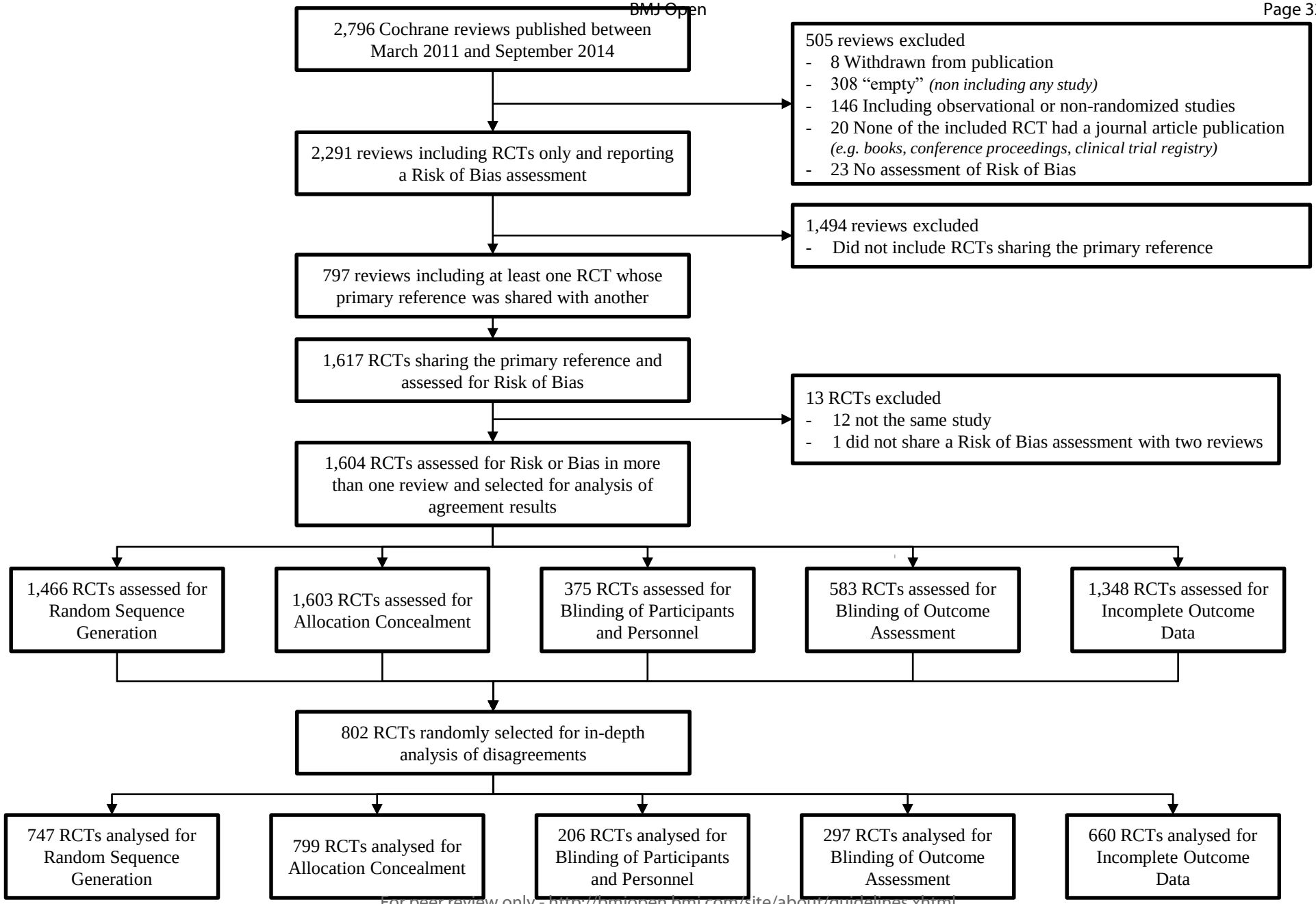
Risk of bias Item	Main reasons for disagreements	N (%)[+]	Examples of support for judgement from the review*
blinding of participants and personnel	Assess risk differently if blinding was not feasible because of the type of intervention	20 (36)	*"Not possible to blind participants"*; Low risk of bias *"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

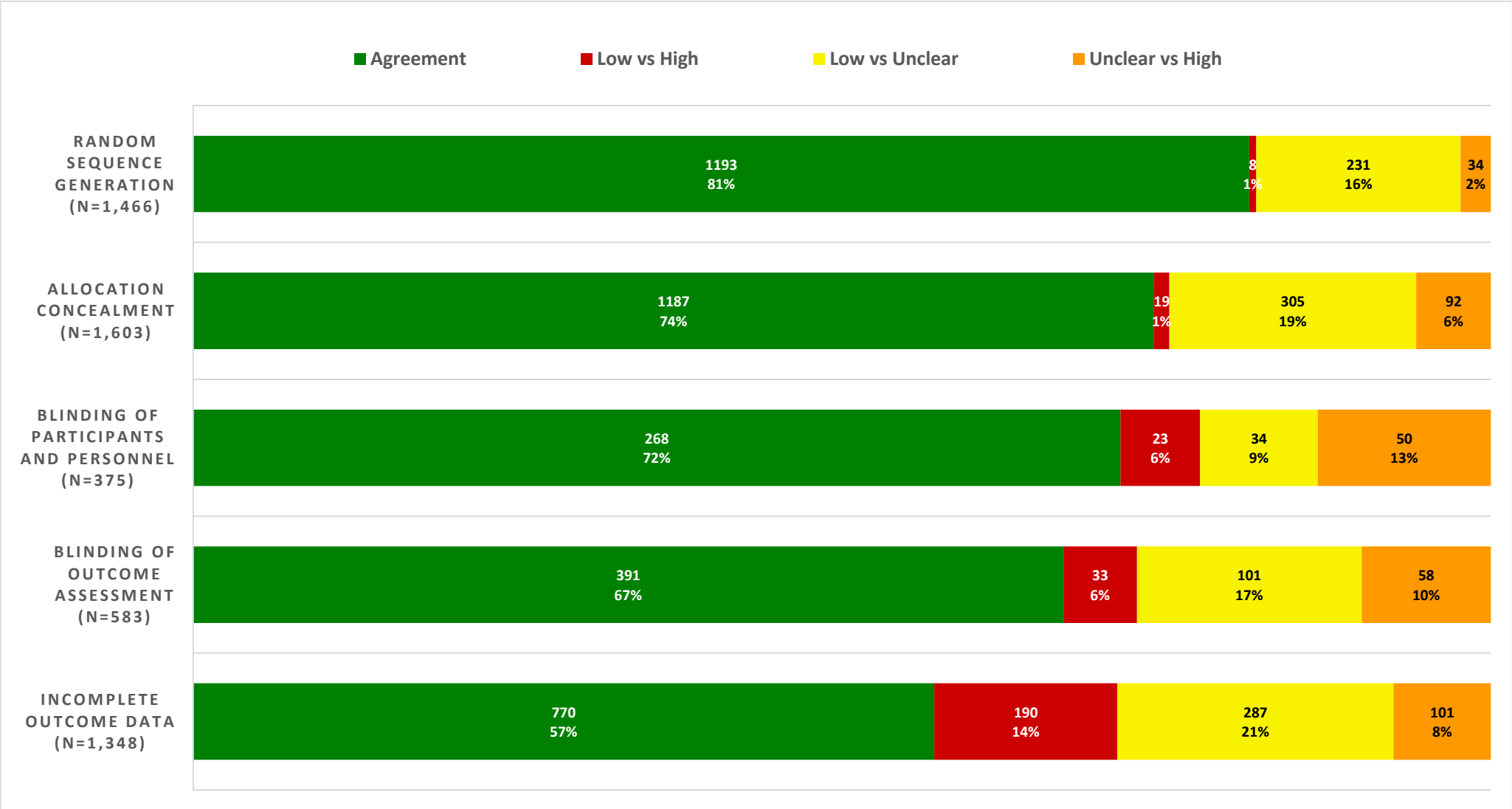
[+] Number of RCTs disagreeing for this reason; percentage over the total of disagreements for different interpretation.
* When two extracts are reported, they refer to the same study.

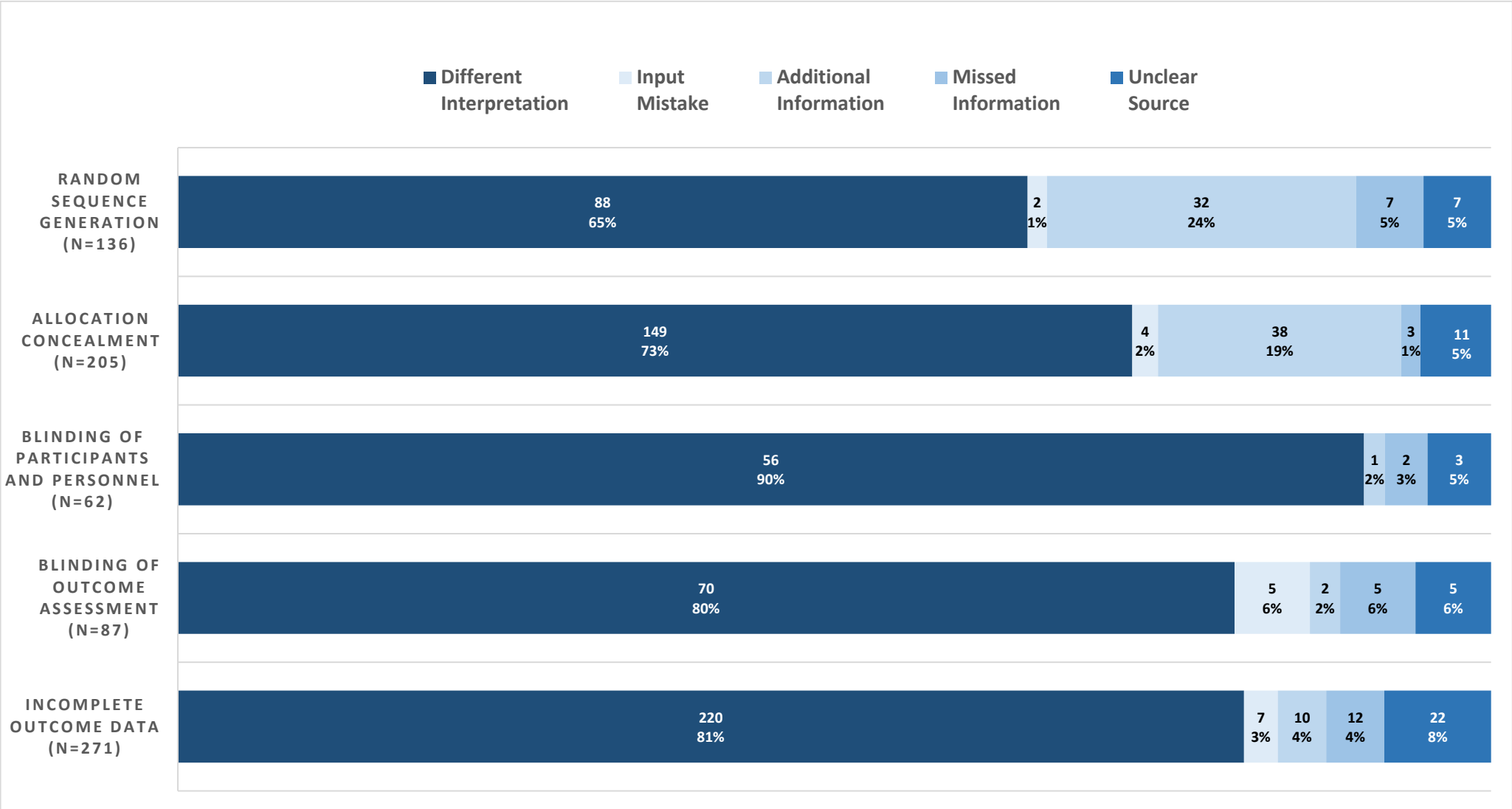
Risk of bias Item	Main reasons for disagreements	N (%)[+]	Examples of support for judgement from the review*
incomplete outcome data	Use different cut-off for the rate of missing data	57 (26)	*"11 withdrawals (10%)."*; Low risk of bias *"Comment: there were post-randomisation drop-outs"*; High risk of bias
	Focus on number vs 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 dropouts 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 (e.g. miscarriage). 1120 in sample. 51/1287 non-responders were included as continuing smokers."* High risk of bias
	Consider differently intention-to-treat analysis	25 (11)	*"147 randomised; 4 in the letrozole group and 3 in the LOD 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 noncompliance"*; 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 dropouts"* Low risk of bias
	Consider differently imputation of missing data	20 (9)	*"Imputation method not described"*; Unclear risk of bias *"Dropout 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)	*"Dropout 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

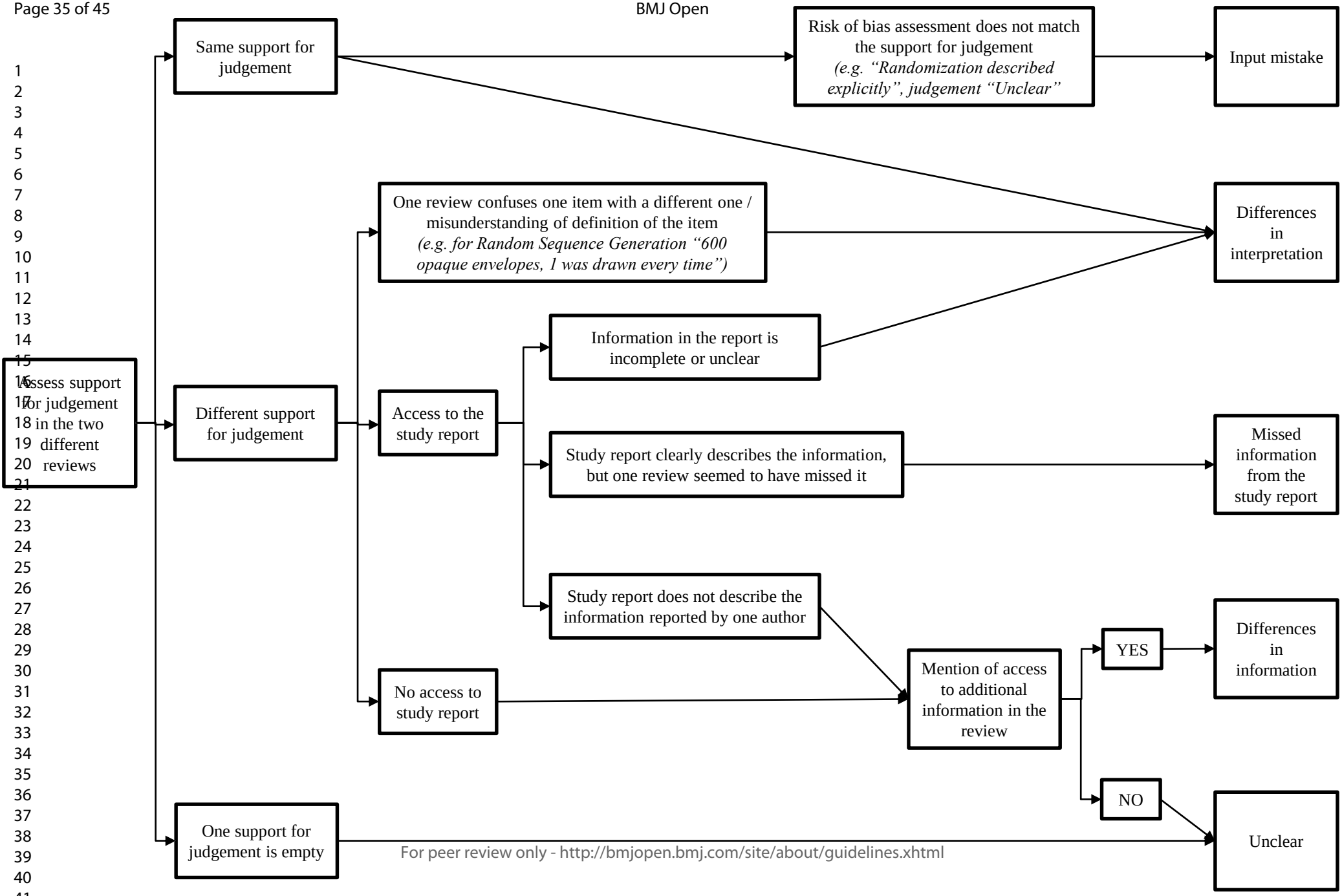
[+] Number of RCTs disagreeing for this reason; percentage over the total of disagreements for different interpretation.

* When two extracts are reported, they refer to the same study.


2,796 Cochrane reviews published between March 2011 and September 2014
505 reviews excluded
- 8 Withdrawn from publication
- 308 "empty" (non including any study)
- 146 Including observational or non-randomized studies
- 20 None of the included RCT had a journal article publication (e.g. books, conference proceedings, clinical trial registry)
- 23 No assessment of Risk of Bias
2,291 reviews including RCTs only and reporting a Risk of Bias assessment
1,494 reviews excluded
- Did not include RCTs sharing the primary reference
797 reviews including at least one RCT whose primary reference was shared with another
1,617 RCTs sharing the primary reference and assessed for Risk of Bias
13 RCTs excluded
- 12 not the same study
- 1 did not share a Risk of Bias assessment with two reviews
1,604 RCTs assessed for Risk or Bias in more than one review and selected for analysis of agreement results
1,466 RCTs assessed for Random Sequence Generation
1,603 RCTs assessed for Allocation Concealment
375 RCTs assessed for Blinding of Participants and Personnel
583 RCTs assessed for Blinding of Outcome Assessment
1,348 RCTs assessed for Incomplete Outcome Data
802 RCTs randomly selected for in-depth analysis of disagreements
747 RCTs analysed for Random Sequence Generation
799 RCTs analysed for Allocation Concealment
206 RCTs analysed for Blinding of Participants and Personnel
297 RCTs analysed for Blinding of Outcome Assessment
660 RCTs analysed for Incomplete Outcome Data

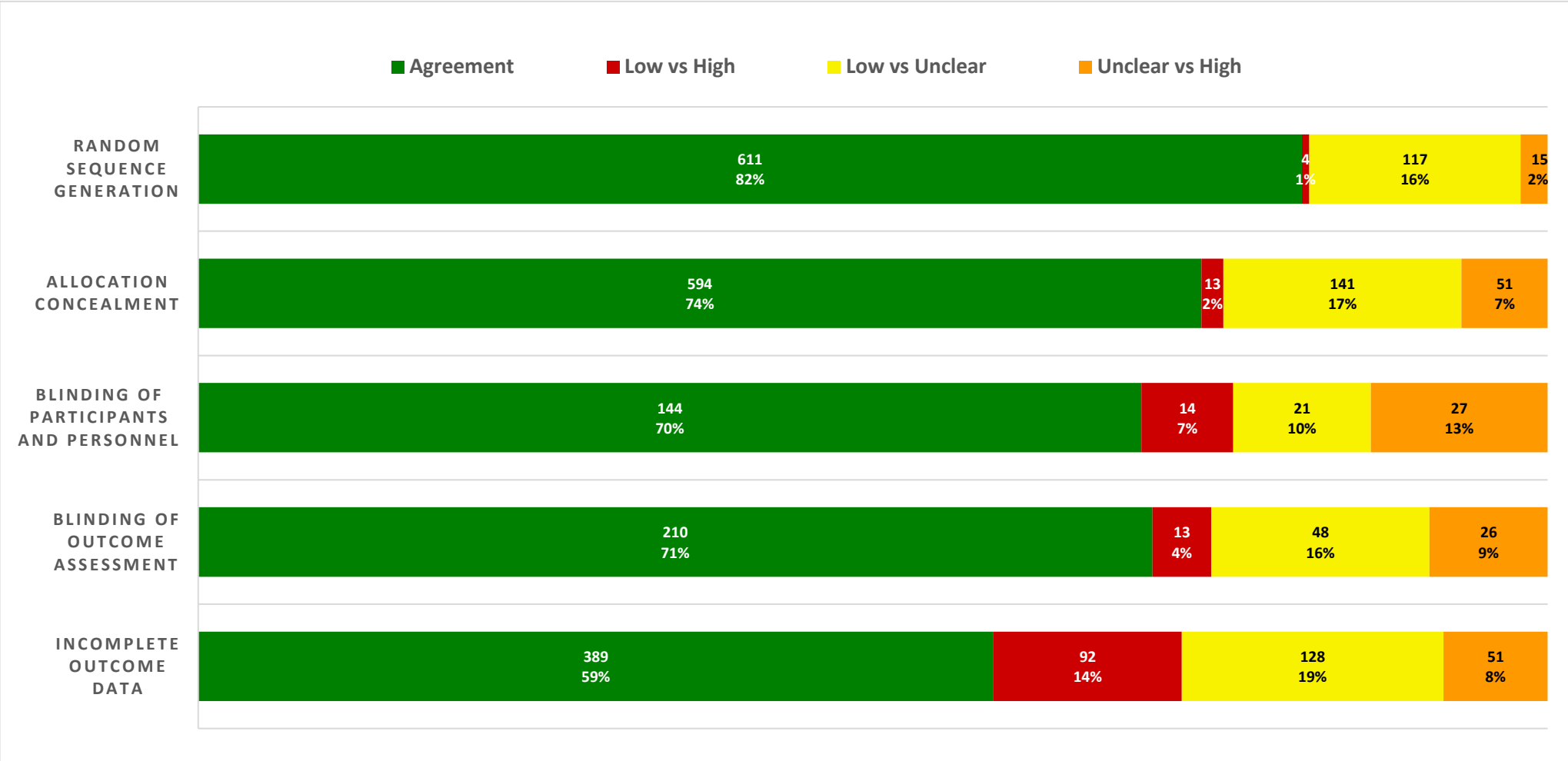

Agreement
Low vs High
Low vs Unclear
Unclear vs High
RANDOM SEQUENCE GENERATION (N=1,466)
1193 81%
8 1%
231 16%
34 2%
ALLOCATION CONCEALMENT (N=1,603)
1187 74%
19 1%
305 19%
92 6%
BLINDING OF PARTICIPANTS AND PERSONNEL (N=375)
268 72%
23 6%
34 9%
50 13%
BLINDING OF OUTCOME ASSESSMENT (N=583)
391 67%
33 6%
101 17%
58 10%
INCOMPLETE OUTCOME DATA (N=1,348)
770 57%
190 14%
287 21%
101 8%


Different Interpretation
Input Mistake
Additional Information
Missed Information
Unclear Source
RANDOM SEQUENCE GENERATION (N=136)
88 65%
2 1%
32 24%
7 5%
7 5%
ALLOCATION CONCEALMENT (N=205)
149 73%
4 2%
38 19%
3 1%
11 5%
BLINDING OF PARTICIPANTS AND PERSONNEL (N=62)
56 90%
1 2%
2 3%
3 5%
BLINDING OF OUTCOME ASSESSMENT (N=87)
70 80%
5 6%
2 2%
5 6%
5 6%
INCOMPLETE OUTCOME DATA (N=271)
220 81%
7 3%
10 4%
12 4%
22 8%


Assess support for judgement in the two different reviews
Same support for judgement
Different support for judgement
One support for judgement is empty
Risk of bias assessment does not match the support for judgement (e.g. "Randomization described explicitly", judgement "Unclear"
Input mistake
One review confuses one item with a different one / misunderstanding of definition of the item (e.g. for Random Sequence Generation "600 opaque envelopes, 1 was drawn every time")
Differences in interpretation
Access to the study report
Information in the report is incomplete or unclear
Study report clearly describes the information, but one review seemed to have missed it
Missed information from the study report
Study report does not describe the information reported by one author
No access to study report
Mention of access to additional information in the review
YES
NO
Differences in information
Unclear

Cochrane Group	Number of reviews	% on the total
Pregrancy and Childbirth	93	11.7%
Airways	48	6.0%
Pain, Palliative and Supportive Care Group	42	5.3%
Acute Respiratory Infections	37	4.6%
Gynaecology and Fertility	29	3.6%
Neonatal	29	3.6%
Tobacco Addiction	27	3.4%
Stroke	25	3.1%
Gynaecological, Neuro-oncology and Orphan Cancer Group	23	2.9%
Wounds	23	2.9%
Hepato-Biliary	22	2.8%
Cystic Fibrosis and Genetic Disorders	21	2.6%
Anaethesia	20	2.5%
Drugs and Alcohol	20	2.5%
Neuromuscolar	19	2.4%
Common Mental Disorders	18	2.3%
Fertility Regulation	18	2.3%
Heart	17	2.1%
Developmental, Psychosocial and Learning Problems	16	2.0%
Incontinence	16	2.0%
Kidney disease	16	2.0%
Schizophrenia	16	2.0%
Infectious Diseases	15	1.9%
Oral Health	14	1.8%
Vascular	14	1.8%
Dementia and Cognitive Improvement	13	1.6%
Musculoskeletal	12	1.5%
Consumers and Communication	10	1.3%
Epilepsy	10	1.3%
Eyes and Vision	9	1.1%
Metabolic and Endocrine Disorders	9	1.1%
Back and Neck	8	1.0%
Hypertension	8	1.0%
Multiple Sclerosis	8	1.0%
Effective Practice and Organisation of Care	7	0.9%
HIV/AIDS	7	0.9%
Infiammatory Bowel Disease	7	0.9%
Injuries	7	0.9%
Bone, Joint and Muscle Trauma Group	6	0.8%
ENT	6	0.8%
Haematological Malignancies	6	0.8%

Cochrane Group	Number of reviews	% on the total
Breast Cancer	5	0.6%
Colorectal Cancer	5	0.6%
Lung Cancer	3	0.4%
Movement Disorders	3	0.4%
Skin	3	0.4%
Occupational Health	2	0.3%
Sexually Transmitted Infections	2	0.3%
Public Health	1	0.1%
Upper GI and Pancreatic Diseases	1	0.1%
Urology	1	0.1%
Total	797	


Agreement
Low vs High
Low vs Unclear
Unclear vs High
RANDOM SEQUENCE GENERATION
611 82%
4 1%
117 16%
15 2%
ALLOCATION CONCEALMENT
594 74%
13 2%
141 17%
51 7%
BLINDING OF PARTICIPANTS AND PERSONNEL
144 70%
14 7%
21 10%
27 13%
BLINDING OF OUTCOME ASSESSMENT
210 71%
13 4%
48 16%
26 9%
INCOMPLETE OUTCOME DATA
389 59%
92 14%
128 19%
51 8%

Supplementary Appendix 4| Examples of in-depth analysis of disagreements conducted with the support of the study report

Risk of bias item	Study Name	Support for judgement*	Information in the study report^	Category of disagreement	Reason of disagreement
Random sequence generation	ABCD 2004	Review 4136: Generated the randomisation list using SAS, stratified by sex and SCr; Low Risk	*"The (...) statistician generated the randomization list using SAS (...) stratified by sex and baseline serum creatinine concentration (...)."*	Missed information from the study report	
		Review 8277: Method not reported; Unclear Risk			
	Cho 2006	Review 7566: Stated that it is a quasi-randomized study but details not given; High Risk	*"... using a quasi-experimental design with a non-equivalent control group." "They were randomly assigned to participate in the experimental group (...) or a waiting-list control group (...)."*	Different interpretation	Consider differently incomplete or unclear description
		Review 9553: Participants randomly allocated to treatment or control group; Unclear Risk			
	Petersen 2005	Review 9132: Quote: "[P]atients were randomly assigned..."Quote: "We used an adaptive allocation scheme for the treatment assignment, with the MMSE score, age and APOE e4 status as balancing covariates"; Low Risk	*"We used an adaptive allocation scheme for the treatment assignment, with the MMSE score, age, and APOE e 4 status as balancing covariates."*	Different interpretation	Confusion or misknowledge
		Review 7176: The trial is described as randomised, but the method of sequence generation was not specified. Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Allocation concealment	Burge 2000	Review 2991: Participants were randomly assigned sequentially from a list comprising treatment numbers only; Low Risk	*"We used a computer generated allocation schedule stratified by centre (block size of six). Patients were randomised sequentially from a list comprising treatment numbers only".*	Different interpretation	Consider differently incomplete or unclear description
		Review 10115: Information not available; Unclear Risk			
	McMurdo 1993	Review 4294: Quote: "Randomisation was by opening sealed envelopes supplied in sequence by the study co-ordinator; Low Risk	*"Randomization was by opening sealed envelopes supplied in sequence by the study co-ordinator (...), and prepared from a computer-generated random numbers table."*	Different interpretation	Consider differently envelopes description
		Review 4963: Unclear, insufficient reporting to permit judgement; Unclear Risk			
	Draper 2007	Review 8179: "... and alternating between treatment or wait list control groups."; High Risk	*"On each occasion that a least eight patients had been recruited, their names were selected at random by a blinded investigator to be allocated alternately to the immediate treatment group or a wait-list control group."*	Different interpretation	Consider differently incomplete or unclear description
		Review 1919: "Reported as concealed but specific method for concealment not reported"; Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Blinding of participants and personnel	Nielsen 2006	Review 9672: "Double-blind"; Low Risk	*"This study was a randomized, placebo-controlled, double blind, Danish, multi-center (two centers) study." "The treatment was applied by a nasal spray with one puff in each nostril every day either in the morning or evening."*	Different information	One review accessed additional data through another study report
		Review 4143: Although "All treatments were supplied as identical intranasal sprays..." the 2004 publication describes a higher rate of withdrawal due to adverse effects in the intervention groups [11.7% in the placebo group, 21.7% in the 150 gm group and 28.7% in the 300 gm group} which may have affected blinding status; Unclear Risk			
	Gersel 1979	Review 10562: Described as double-blind [presumed participants and personnel/investigators]; Low Risk	*"A double-blind experimental design was used, employing each patient as his own control."*	Different interpretation	Consider differently information of "double blind"
		Review 6968: Not mentioned and no information to suggest this was done.; Unclear Risk			
	Stein 2011	Review 7025: Not possible to blind participants to intervention. Insufficient information to make a judgement about blinding of therapists; High Risk	*"Follow-up assessment was made 3 months after release (research staff conducting assessments were blind to treatment assignment)." "Randomization was accomplished via random numbers table in advance and placed in an envelope by the project coordinator. Following baseline assessment, research staff opened the envelope to learn of intervention assignment."*	Different interpretation	Consider differently incomplete or unclear description
		Review 10901: Researchers were blind until after the baseline assessment. Participants were not blinded.; Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Blinding of outcome assessment	Schoen 2007	Review 3603: Outcome assessor was not blinded.; High risk	*"In total 72 patients were screened by a maxillofacial surgeon (PJS) and prosthodontist (HR)."* *"All clinical assessments were performed by the investigator (PJS) who was not involved in treatment of the patients."*	Different interpretation	Consider differently incomplete or unclear description
		Review 5005: Outcome assessor may have been unaware of allocation: "All clinical assessments were performed by the investigator (PJS) who was not involved in treatment of the patients."; Low risk			
	Geroin 2011	Review 6185: Not done; High risk	*"All patients were evaluated by the same examiner (an experienced internal coworker) who was not aware of the treatment received by the patients."*	Missed information from the study report	
		Review 9645: Quote: "All patients were evaluated by the same examiner (an experienced internal coworker) who was not aware of the treatment received by the patients"; Low Risk			
	McCambridge 2004	Review 8969: As one interventionist was the study PI, a second independent interviewer who was blind to study condition was employed to conduct 3 month follow-ups, and an additional interviewer who was blind to initial group allocation was employed for 12 months follow-ups; Low Risk	*"further area of possible bias was that intervention recipients might report more favourable outcome data to the researcher who had delivered the intervention (J.M.). To study any such bias, a second independent interviewer who was blind to study condition, was employed to interview a sample of participants."*	Different interpretation	Consider differently incomplete or unclear description
		Review 7025: A second independent interviewer who was blind to study condition was employed to interview a sample of participants, though not all participants; Unclear Risk			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Risk of bias item	Study Name	Support for judgement*	What is reported in the study report^	Category of disagreement	Reason of disagreement
Incomplete outcome data	Altmaier 1992	Review 1822: All subjects recorded follow-up data; Low Risk	*[From table] "The n = 21 for control group and n = 24 for psychological group on all process measures." [From table] The n = 21 for each group at each assessment.]*	Different interpretation	Consider differently incomplete or unclear description
		Review 7407: Inadequately reported; High Risk			
	Killen 1984	Review 146: 11/75 recruited dropped out before full treatment, and are excluded from analyses.; Low Risk	*"The first 75 were accepted into the study. Seven failed to attend (...) two dropped (...). The final sample (N = 64)."*	Missed information from the study report	
		Review 3999: Losses to follow-up not reported, all participants included; Unclear Risk			
	Creager 2008	Review 986: There was a huge loss to follow up (only 50% completed the 6 month follow up) in this study and therefore there is a high risk of attrition bias; High Risk	*"The remaining 525 patients met the inclusion criteria (...) The remaining 430 patients met their criteria for randomization (...).The ITT population consisted of 370 randomized patients (...). The per-protocol patient population consisted of 214 randomized patients"*	Different interpretation	Consider differently intention-to-treat analysis
		Review 5262: Unclear why of patients stopped medication, unclear whether data presented represents intention-to-treat or per-protocol analysis			

*Supports for judgement and risk of bias assessments for the two reviews compared. The number of the review corresponds to the last 4 digits of the DOI.

^ Information that were highlighted in the study report to support the analysis process.

Supplementary Appendix 5| Reasons for disagreements in cases of a different interpretation of the same information; focus on "low" versus "high" disagreements.

Risk of bias item	Main reasons for disagreements	Examples
random sequence generation	Consider differently incomplete or unclear description	*"The names of communities within each group of three were written on individual cards, mixed and selected randomly: the first from each group was assigned to arm A (IEC alone), the second to arm B (IEC and STI management) and the third to arm C";* Low risk of bias *"Names of communities within each triplet were written on separate cards and shuffled.";* High risk of bias
allocation concealment	Consider differently envelopes description	*"Closed envelopes";* Low risk of bias *"Closed envelopes, although not opaque.";* High risk of bias
	Confusion in the definition of the item	*"pg. 2 - Methods - randomisation was done centrally to preserve allocation concealment"*; Low risk of bias *"904 patients were eligible for the study. 446 patients were randomised (49%). Due to the number of patients declining screening, there is an increased risk of inclusion bias."*; High risk of bias
	Confusion with blinding	*"States used "preprogrammed laptop computer". Remote site";* Low risk of bias *"participants were told to which compound they had been allocated."; High risk of bias*
blinding of participants and personnel	Assess risk differently if blinding was not feasible because of the type of intervention	*"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	*"Not possible to blind but most of the outcomes not likely to be influenced by lack of blinding.";* Low risk of bias *"Not blinded due to nature of intervention."*; High risk of bias
	Confusion with allocation concealment	*"participants were randomly allocated to either intervention or control group by an independent party";* Low risk of bias *"Control group did not receive the comparable non-exercise related attention to the intervention group"*; High risk of bias

Risk of bias Item	**Main reasons for disagreements**	**Examples**
blinding of outcome assessment	outcome considered not influenced by blinding	*"No information given about whether patients or assessors were blind to physician allocation but primary outcomes (treatment outcome and patient reported physician cultural competency) judged unlikely to be affected by lack of blinding"*; Low risk of bias *"Unblinded."*; High risk of bias
	Consider differently patient reported outcomes when patients are blinded or not to the intervention	*"Insufficient information available to assess"*; Low risk of bias *"Comment: depression assessed by patient self-report"*; High risk of bias
	Assess risk differently if blinding was not feasible because of the type of intervention	*"Unclear blinding of outcome assessment"*; Low risk of bias *"blinding not possible due to intervention"*; High risk of bias
incomplete outcome data	Use different cut-off for the rate of missing data	*"11 withdrawals (10%)."*; Low risk of bias *"Comment: there were post-randomisation drop-outs"*; High risk of bias
	Focus on number vs reasons/precise report of missing data	*"Numbers and reasons for dropouts and withdrawals in all intervention groups were described."*; Low risk of bias *"20 drop-outs (27.2%) with 4 deaths (3 males, 1 female) from cardiovascular events"*; High risk of bias
	Consider differently incomplete or unclear description	*"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 (e.g. miscarriage). 1120 in sample. 51/1287 non-responders were included as continuing smokers."* High risk of bias
	Use different cut-off for difference in the rate missing data between different arms/comparisons	*"Similar rates of withdrawal between arms. Withdrawals: 36 BUD, 51 placebo"*; Low risk of bias *"Dropout higher in placebo group (35% vs 25% in budesonide group). ITT used."*; High risk of bias