Revised Cochrane risk-of-bias tool for randomized trials (RoB 2)

Edited by Julian PT Higgins, Jelena Savović, Matthew J Page, Jonathan AC Sterne
on behalf of the ROB2 Development Group

9 October 2018

Dedicated to Professor Douglas G Altman, whose contributions were of fundamental importance to development of risk of bias assessment in systematic reviews

Contents

1 Introduction

The RoB 2 tool provides a framework for considering the risk of bias in the findings of any type of randomized trial. The assessment is specific to a single trial result that is an estimate of the relative effect of two interventions or intervention strategies on a particular outcome. We refer to the interventions as the **experimental intervention** and the **comparator intervention**, although we recognize that the result may sometimes refer to a comparison of two active interventions.

The tool is structured into five domains through which bias might be introduced into the result. These were identified based on both empirical evidence (see Box 1) and theoretical considerations. Because the domains cover all types of bias that can affect results of randomized trials, each is mandatory, and no further domains should be added. The five domains for individually randomized trials (including cross-over trials) are:

(1) bias arising from the randomization process;
(2) bias due to deviations from intended interventions;
(3) bias due to missing outcome data;
(4) bias in measurement of the outcome;
(5) bias in selection of the reported result.

The domain names are direct descriptions of the causes of bias addressed in the domain. We have avoided many of the terms used in version 1 of the tool (e.g. selection bias, performance bias, attrition bias, detection bias) because they do not describe the specific issues addressed and so cause confusion (1).

We offer several templates for addressing these domains, tailored to the following study designs:

(1) randomized parallel-group trials;
(2) cluster-randomized parallel-group trials (including stepped-wedge designs);
(3) randomized cross-over trials and other matched designs.

For cluster-randomized trials, an additional domain is included ((1b) Bias arising from identification or recruitment of individual participants within clusters).

This document describes the main features of the RoB 2 tool and provides guidance for its application to individually randomized parallel-group trials. Supplementary documents address additional considerations for cluster-randomized parallel-group trials and individually-randomized cross-over trials. We have not yet developed a version appropriate for cluster cross-over trials.

Box 1: Empirical evidence of bias in randomized trials: the role of meta-epidemiology

Empirical evidence of bias in randomized trials comes from a field known as **meta-epidemiology** (2). A meta-epidemiological study analyses the results of a large collection of previous studies to understand how methodological characteristics of the studies are associated with their results. The first well-known meta-epidemiological study examined 33 meta-analyses containing 250 clinical trials (3). Each trial was categorized on the basis of four characteristics: whether sequence generation was reported to have a random component; whether allocation was reported to be adequately concealed, whether the trial was described as double-blind, and whether the trial reported exclusion of participants from its analysis. For each of the four characteristics separately, trials were compared within each meta-analysis to estimate a ratio of odds ratios among the 'better' versus the 'worse' trials, and these ratios of odds ratios were combined across the 33 meta-analyses. Numerous similar studies have been undertaken since, examining many study characteristics that are potentially associated with biases in results. More recent analyses address both the average and the variability in bias associated with the characteristic under investigation. Specifically, they examine also the extent to which the characteristic is associated with increased between-trial heterogeneity and to which the average bias varies between meta-analyses (4). In several places in this document we refer to empirical evidence from meta-epidemiological studies, or systematic reviews of them, to support the selection of domains and signalling questions.

1.1 Signalling questions

Inclusion of signalling questions within each domain of bias is a key feature of RoB 2. Signalling questions aim to elicit information relevant to an assessment of risk of bias. They seek to be reasonably factual in nature. Responses to these questions feed into algorithms we have developed to guide users of the tool to judgements about the risk of bias.

The **response options for the signalling questions** are:

(1) Yes;
(2) Probably yes;
(3) Probably no;
(4) No;
(5) No information;

To maximize the signalling questions' simplicity and clarity, they are phrased such that a response of 'Yes' may be indicative of either a low or high risk of bias, depending on the most natural way to ask the question.

Responses of 'Yes' and 'Probably yes' have the same implications for risk of bias, as do responses of 'No' and 'Probably no'. The definitive versions ('Yes' and 'No') would typically imply that firm evidence is available in relation to the signalling question; the 'Probably' versions would typically imply that a judgement has been made. If review authors calculate measures of agreement (e.g. kappa statistics) for the answers to the signalling questions, we recommend treating 'Yes' and 'Probably yes' as the same response and 'No' and 'Probably no' as the same response.

The 'No information' response should be used only when both (i) insufficient details are reported to permit a response of 'Probably yes' or 'Probably no', and (ii) in the absence of these details it would be unreasonable to respond 'Probably yes' or 'Probably no' in the circumstances of the trial. For example, in the context of a large trial run by an experienced clinical trials unit, absence of specific information about the randomization methods is likely to result in a response of 'Probably yes' rather than 'No information' to the signalling question about allocation concealment. The implications for risk of bias judgements of a 'No information' response to a signalling question differ according to the purpose of the question. If the question seeks to identify evidence of a problem, then 'No information' corresponds to no evidence of that problem. If the question relates to an item that is expected to be reported (such as whether any participants were lost to follow up), then the absence of information leads to concerns about there being a problem.

For signalling questions that are answered only if the response to a previous question implies that they are required, a response option "Not applicable" is available. Signalling questions should be answered independently: the answer to one question should not affect answers to other questions in the same or other domains other than through determining which subsequent questions are answered.

1.1.1 Free-text boxes alongside signalling questions

The tool provides space for free text alongside the signalling question. In some instances, when the same information is likely to be used to answer more than one question, one text box covers more than one question. These boxes should be used to provide support for the answer to each signalling question. Brief **direct quotations** from the text of the study report should be used whenever possible.

1.2 Risk-of-bias judgements

1.2.1 Domain-level judgements about risk of bias

RoB 2 is conceived hierarchically: responses to signalling questions elicit what happened and provide the basis for domain-level judgements about the risk of bias. In turn, these domain-level judgements provide the basis for an overall risk-of-bias judgement for the specific trial result being assessed.

The tool includes algorithms that map responses to signalling questions onto a proposed risk-of-bias judgement for each domain. The possible **risk-of-bias judgements** are:

(1) Low risk of bias;
(2) Some concerns; and
(3) High risk of bias.

Use of the word "judgement" is important for the risk-of-bias assessment. In particular, the algorithms provide proposed judgements, but users should verify these and change them if they feel this is appropriate. In reaching final judgements, the following considerations apply:

- "Risk of bias" is to be interpreted as "**risk of material bias**". That is, concerns should be expressed only about issues that are likely to affect the ability to draw reliable conclusions from the study.
- Domain-level judgements about risk of bias should have the same implication for each of the six domains with respect to concern about the impact of bias on the trustworthiness of the result. A judgement of 'High' risk of bias for any individual domain will lead to the result being at 'High' risk of bias overall, and a judgement of 'Some concerns' for any individual domain will lead to the result being at 'Some concerns', or 'High' risk, overall (see 1.2.3).

1.2.2 Direction of bias

The tool includes optional judgements of the direction of the bias for each domain and overall. For some domains, the bias is most easily thought of as being towards or away from the null. For example, high levels of switching of participants from their assigned intervention to the other intervention would be likely to lead to the estimated effect of adhering to intervention being biased towards the null. For other domains, the bias is likely to favour one of the interventions being compared, implying an increase or decrease in the effect estimate depending on which intervention is favoured. Examples include manipulation of the randomization process, awareness of interventions received influencing the outcome assessment and selective reporting of results. **If review authors do not have a clear rationale for judging the likely direction of the bias, they should not guess it.**

1.2.3 Reaching an overall judgement about risk of bias

The response options for an overall risk-of-bias judgement are the same as for individual domains. Table 1 shows the basic approach to mapping risk-of-bias judgements within domains to an overall judgement across domains for the outcome.

Table 1. Reaching an overall risk-of-bias judgement for a specific outcome.

Overall risk-of-bias judgement	Criteria
Low risk of bias	The study is judged to be at **low risk of bias for all domains** for this result.
Some concerns	The study is judged to raise **some concerns** in at least one domain for this result, but not to be at high risk of bias for any domain.
High risk of bias	The study is judged to be at **high risk of bias** in at least one domain for this result. Or The study is judged to have **some concerns** for **multiple domains** in a way that substantially lowers confidence in the result.

Judging a result to be at a particular level of risk of bias for an individual domain implies that the result has an overall risk of bias at least this severe. Therefore, a judgement of 'High' risk of bias within any domain should have similar implications for the result as a whole, irrespective of which domain is being assessed. 'Some concerns' in multiple domains may lead the review authors to decide on an overall judgement of 'High' risk of bias for that outcome or group of outcomes.

1.2.4 Free-text boxes alongside risk-of-bias judgements

There is space for free text alongside each risk-of-bias judgement to explain the reasoning that underpins the judgement. It is particularly important that reasons are provided for any judgements that do not follow the proposed algorithms.

1.3 Specifying the nature of the effect of interest

Assessments for the domain 'Bias due to deviations from intended interventions' vary according to whether review authors are interested in quantifying:

(1) the effect of **assignment** to the interventions at baseline (regardless of whether the interventions are received during follow-up, sometimes known as the 'intention-to-treat effect'); or
(2) the effect of **adhering to** the interventions as specified in the trial protocol (sometimes known as the 'per-protocol effect').

These two effects of intervention will differ if some patients do not receive their assigned intervention or deviate from the assigned intervention after baseline. They are each of interest, in both pragmatic and explanatory trials. For example, the estimated effect of assignment to intervention would be the most appropriate to inform a health policy question about whether to recommend an intervention in a particular health system (e.g. whether to instigate a screening programme, or whether to prescribe a new cholesterol-lowering drug), whereas the estimated effect of adhering to the intervention as specified in the trial protocol would be the most appropriate to inform a care decision by an individual patient (e.g. whether to be screened, or whether to take the new drug). Review authors need to define the intervention effect in which they are interested, and apply the risk-of-bias tool appropriately to this effect.

The effect of principal interest should be specified in the review protocol. On occasion, review authors may be interested in both effects of interest. Note that specification of the "effect of interest" in RoB 2 does not relate to the choice of treatment effect metric (odds ratio, risk difference etc.).

1.3.1 Estimating the effect of interest

Authors of randomized trials can use several different analysis approaches to estimate interventions effects. They may not explain the reasons for their choice of analysis approach, or whether their aim is to estimate the effect of assignment or adherence to intervention. We discuss different approaches to analysis and their implications for bias. Because multiple analyses can be reported, we also suggest an order of preference in which estimated effects of intervention should be chosen when review authors are interested in the effect of assignment to intervention.

The effect of **assignment** to intervention should be estimated by an **intention-to-treat (ITT) analysis** that includes all randomized participants (5). The principles of ITT analyses are (6, 7):

(1) analyse participants in the intervention groups to which they were randomized, regardless of the intervention they actually received;
(2) include all randomized participants in the analysis; and
(3) measure outcome data on all participants.

An ITT analysis maintains the benefit of randomization: that the intervention groups do not differ systematically with respect to measured or unmeasured prognostic factors. Note that the term 'intention-to-treat' does not have a consistent definition, and is used inconsistently in study reports (8-10).

In a placebo-controlled trial in which there is non-adherence to randomized interventions, an ITT analysis will usually underestimate the intervention effect that would have been seen if all participants had adhered to the intervention (because we expect effect no or less effect of the experimental intervention in participants who do not adhere). Although ITT effects are often conservative with regard to desired effects of interventions estimated in placebo-controlled trials, they may not be conservative in trials comparing two or more active interventions (11, 12), and are problematic for non-inferiority or equivalence studies (11). Underestimation of the effect if all participants had adhered to the intervention may be particularly problematic when examining harms (adverse effects) of the experimental intervention. Variable rates of non-adherence to randomized allocation may also be a source of heterogeneity in ITT estimates of intervention effects: we expect greater effectiveness in a trial with perfect adherence than in a trial in which a substantial proportion of participants do not adhere (13).

Patients and other stakeholders are often interested in the effect of adhering to the intervention as described in the trial protocol (the 'per protocol effect'). However, the approaches to estimation of per-protocol effects that are commonly used in reports of randomized trials are problematic and may be seriously biased. Two particular types of analysis are:

- 'as-treated' analyses in which participants are analysed according to the intervention they actually received, even if their randomized allocation was to a different treatment group; and

- naïve 'per protocol' analyses restricted to individuals in each intervention group who adhered to the interventions.

Each of these analyses is problematic because prognostic factors may influence whether individuals receive their allocated intervention.

Trial authors often estimate the effect of intervention using more than one approach. **We recommend that when the effect of interest is that of assignment to intervention, the trial result included in meta-analyses, and assessed for risk of bias, should be chosen according to the following order of preference:**

(1) The result corresponding to a full ITT analysis, as defined above;
(2) The result corresponding to an analysis (sometimes described as a 'modified intention-to-treat' (mITT) analysis) that adheres to ITT principles except that participants with missing outcome data are excluded (see section 5.3.1). Such an analysis does not prevent bias due to missing outcome data, which is addressed in the corresponding domain;
(3) A result corresponding to an 'as treated' or naïve 'per-protocol' analysis, or an analysis from which eligible trial participants were excluded.

It is possible, using modern statistical methods, to use data from randomized trials to estimate the effect of adhering to intervention (13) although applications of such methods are relatively rare to date. The most accessible approach is to use randomization status as an instrumental variable in order to estimate the per-protocol effect in trials in which a single intervention, administered at baseline, is compared with standard of care. If a substantial proportion of patients allocated to intervention do not receive it, an ITT analysis will underestimate the per protocol effect. Instrumental variable analyses overcome the problems of traditional 'per-protocol' analyses because they bypass the need to adjust for prognostic factors that predict receipt of intervention. The assumptions required for their validity are described by Hernán and Robins (13).

Estimation of per-protocol effects for trials comparing interventions that are sustained over time is more difficult, because deviations from intervention after baseline necessitate statistical adjustment for post-randomization values of prognostic factors. Conventional statistical methods, such as standard regression models, are not valid if these post-randomization prognostic factors are affected by prior intervention. Valid methods ('g-methods') include inverse probability weighting and the g-formula, but require data on both the pre-randomization and post-randomization prognostic factors that predict deviations from intended intervention (11-13).

For each effect of interest, a signalling question in the domain 'Bias due to deviations from intended interventions' asks whether appropriate statistical methods were used to estimate that effect.

2 Issues in implementation of RoB 2

2.1 Multiple assessments

Trials usually contribute multiple results to a systematic review, mainly through contributing to multiple outcomes. Therefore, several risk-of-bias assessments may be needed for each study. We have not yet formulated recommendations on which results should be targeted with an assessment, or how many results should be assessed. However, these decisions are likely to align with the outcomes included in a Summary of Findings table.

2.2 The data collection process

Assessment of risk of bias is specific to a particular result, for a particular outcome measured at a particular time, from the study. However, some causes of bias (such as biases arising from the randomization process) apply generally to the whole study; some (such as bias due to deviations from intended intervention) apply mainly to the outcome being measured; some (such as bias in measurement of outcomes) apply mainly to the outcome measurement method used; and some (such as bias in selection of the reported result) apply to the specific result. This has implications for how review authors can most efficiently extract information relevant to risk of bias from study reports.

2.3 Presentation of risk-of-bias assessments

We suggest that RoB 2 assessments are presented as follows. More work is required in this area.

- For full transparency of the process, review authors may wish to present the answers, free-text supports and judgements for each assessor separately. Since these may be confusing to the reader, we recommend that they are not presented prominently, so might be included in an appendix or supplementary document.
- Present the domain-level judgements in the main review document (e.g. as a table, or a figure, or within a forest plot of the results). Only consensus judgements across multiple assessors should be presented. If space permits, abridged free-text justifications for each judgement would be an attractive supplement to this within the main review document.
- Provide answers for each signalling question and the free text support for each of these answers in an appendix or supplementary document. Only consensus answers across multiple assessors should be presented.

2.4 Rapid assessments

Because the default overall judgement for the result will be 'High' risk of bias if one of the domains is judged at 'High' risk of bias, users of the tool may be tempted to stop their assessment as soon as one domain is judged as 'High'. We discourage this when the tool is used in the context of a full systematic review, for several reasons. First, many readers of systematic reviews prefer to see full and consistent evaluations of the included evidence, in the interests of transparency. Second, full evaluations of the limitations of existing randomized trials are likely to be useful in the design and conduct of future trials of the intervention(s) in question. Third, there is a drive from the research community to make risk-of-bias assessments of trials produced by review authors publicly available alongside trial results; a fully documented domain-level assessment is needed for this. A more minor consideration for review authors is that meta-epidemiological studies, which re-analyse multiple meta-analyses to learn about the impact of trial design features, and are invaluable sources of information about the size and direction of biases introduced by study limitations, require full assessments for each domain of the tool (14).

We recognize that some users of the tool may need to introduce "stopping rules" into their assessment when the sole purpose is to reach a rapid judgement about whether the trial is at 'High' risk of bias. We recommend that this be done only when it has been pre-specified in the protocol that trials judged to be at 'High' risk of bias will play no role in the synthesis of evidence. If trials are to be included in sensitivity analyses or subgroup analyses, then we recommend that full assessments be made so that the study can be appropriately characterized.

3 Detailed guidance: preliminary considerations

Before completing the risk-of-bias assessment, it is helpful to document important characteristics of the assessment, such as the design of the trial, the outcome being assessed (as well as the specific result being assessed), and whether interest focusses on the effect of *assignment to intervention* or the effect of *adhering to intervention*. Review authors should document the sources that are used to complete the assessment (as many sources as possible should be used in practice). The RoB 2 standard template includes questions to capture these details (Box 2).

Box 2. The RoB 2 tool (part 1): Preliminary considerations

Study design

☐ Individually-randomized parallel-group trial

☐ Cluster-randomized parallel-group trial

☐ Individually randomized cross-over (or other matched) trial

Specify which outcome is being assessed for risk of bias	
Specify the numerical result being assessed. In case of multiple alternative analyses being presented, specify the numeric result (e.g. RR = 1.52 (95% CI 0.83 to 2.77) and/or a reference (e.g. to a table, figure or paragraph) that uniquely defines the result being assessed.	

Is the review team's aim for this result…?

☐ to assess the effect of *assignment to intervention* (the 'intention-to-treat' effect)

☐ to assess the effect of *adhering to intervention* (the 'per-protocol' effect)

Which of the following sources were <u>obtained</u> to help inform the risk-of-bias assessment? (tick as many as apply)

☐ Journal article(s) with results of the trial
☐ Trial protocol
☐ Statistical analysis plan (SAP)
☐ Non-commercial trial registry record (e.g. ClinicalTrials.gov record)
☐ Company-owned trial registry record (e.g. GSK Clinical Study Register record)
☐ "Grey literature" (e.g. unpublished thesis)
☐ Conference abstract(s) about the trial
☐ Regulatory document (e.g. Clinical Study Report, Drug Approval Package)
☐ Research ethics application
☐ Grant database summary (e.g. NIH RePORTER or Research Councils UK Gateway to Research)
☐ Personal communication with trialist
☐ Personal communication with the sponsor

4 Detailed guidance: bias arising from the randomization process

4.1 Background

If successfully accomplished, randomization avoids an influence of either known or unknown prognostic factors (factors that predict the outcome, such as severity of illness or presence of comorbidities) on intervention group assignment. This means that, on average, the intervention groups have the same prognosis before the start of intervention. If prognostic factors influence the intervention group to which participants are assigned then the estimated effect of intervention will be biased by 'confounding', which occurs when there are common causes of intervention group assignment and outcome. Confounding is an important potential cause of bias in intervention effect estimates from observational studies, because treatment decisions in routine care are often influenced by prognostic factors.

To randomize participants into a trial a rule for allocating interventions to participants must be specified, based on some chance (random) process. We call this **allocation sequence generation**. Subsequently, steps must be taken to prevent participants or trial personnel from knowing the forthcoming allocations until after recruitment hs confirmed. This process is often called **allocation sequence concealment**.

Knowledge of the next assignment (e.g. if the sequence is openly posted on a bulletin board) can enable selective enrolment of participants on the basis of prognostic factors. Participants who would have been assigned to an intervention deemed to be "inappropriate" may be rejected. In epidemiological terms this is a type of selection bias. Other participants may be directed to the "appropriate" intervention, which can be accomplished by delaying their entry into the trial until the desired allocation appears. In epidemiological terms, such manipulation of the assigned intervention may introduce confounding. For this reason, successful allocation sequence concealment is a vital part of randomization.

Some review authors confuse allocation concealment with blinding of assigned interventions during the trial. Allocation concealment seeks to prevent bias in intervention assignment by preventing trial personnel and participants from knowing the allocation sequence before and until assignment. It can always be successfully implemented, regardless of the study design or clinical area (15, 16). In contrast, blinding seeks to prevent bias after assignment (16, 17), and cannot always be implemented. This is often the situation, for example, in trials comparing surgical with non-surgical interventions. Thus, allocation concealment up to the point of assignment of the intervention and blinding after that point address different sources of bias and differ in their feasibility.

4.1.1 Approaches to sequence generation

Randomization with no constraints is called **simple randomization** or **unrestricted randomization**. Sometimes **blocked randomization (restricted randomization)** is used to generate a sequence to ensure that the desired ratio of participants in the experimental and comparator intervention groups (e.g. 1:1) is achieved (18, 19). This is done by ensuring that the numbers of participants assigned to each intervention group is balanced within blocks of specified size (for example, for every 10 consecutively entered participants): the specified number of allocations to experimental and comparator intervention groups is assigned in random order within each block. If the block size is known to trial personnel, then the last allocation within each block can always be predicted. To avoid this problem the block size may be randomly varied (random permuted blocks).

Stratified randomization, in which restricted randomization is performed separately within subsets of participants defined by potentially important prognostic factors, such as disease severity and study centres, is also common. If simple (rather than restricted) randomization is used in each stratum, then stratification offers no benefit, but the randomization is still valid.

Another approach that incorporates both general concepts of stratification and restricted randomization is **minimization**, which can be used to make intervention groups closely similar with respect to several prognostic factors. Minimization generally includes a random element (at least for participants enrolled when the groups are balanced with respect to the prognostic factors included in the algorithm). Some methodologists are cautious about the acceptability of minimization, while others consider it to be an attractive approach (20, 21).

Other adequate types of randomization that are sometimes used are biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization (18, 22, 23). If these or other approaches are encountered, consultation with a methodologist may be necessary.

4.1.2 Allocation concealment and failures of randomization

If future assignments can be anticipated, leading to a failure of allocation concealment, then bias can arise through selective enrolment of participants into a study, depending on their prognostic factors, in the light of the upcoming intervention assignment. Ways in which this can happen include:

(1) knowledge of a deterministic assignment rule, such as by alternation, date of birth or day of admission;
(2) knowledge of the sequence of assignments, whether randomized or not (e.g. if a sequence of random assignments is posted on the wall);
(3) ability to predict assignments successfully, based on previous assignments.

The last of these can occur when blocked randomization is used, and when assignments are known to the recruiter after each participant is enrolled into the trial. It may then be possible to predict future assignments, particularly when blocks are of a fixed size and are not divided across multiple recruitment centres (24).

Attempts to achieve allocation concealment may be undermined in practice. For example, unsealed allocation envelopes may be opened, while translucent envelopes may be held against a bright light to reveal the contents (3, 25, 26). Personal accounts suggest that many allocation schemes have been deciphered by investigators because the methods of concealment were inadequate (25).

Information about methods for sequence generation and allocation concealment can usually be found in trial protocols for randomized trials, but unfortunately is often not fully reported in publications (21). For example, a Cochrane review on the completeness of reporting of randomized trials found allocation concealment reported adequately in only 45% (393/876) of randomized trials in CONSORT-endorsing journals and in 22% (329/1520) of randomized trials in non-endorsing journals (27). This can sometimes be due to limited word counts in journals, highlighting the importance of looking at multiple information sources. Lack of description of methods of randomization and allocation concealment in a journal article does not necessarily mean that the methods used were inappropriate (28).

The success of randomization in producing comparable groups is often examined by comparing baseline values of important prognostic factors between intervention groups. In contrast to the under-reporting of randomization methods, baseline characteristics are reported in 95% of RCTs published in CONSORT-endorsing journals and in 87% of RCTs in non-endorsing journals (27). Corbett et al have argued that risk-of-bias assessments should consider whether participant characteristics are balanced between intervention groups (29). RoB 2 includes a signalling question requiring a judgement about whether baseline imbalances suggest that there was a problem with the randomization process (see 4.3.3).

4.2 Empirical evidence of bias arising from the randomization process

A recent meta-analysis of seven meta-epidemiological studies found that an inadequate or unclear (versus adequate) method of **sequence generation** was associated with a small (7%) exaggeration of intervention effect estimates (30). Unexpectedly, the bias was greater in trials reporting subjective outcomes: there was little evidence for bias in trials assessing all-cause mortality and other objective outcomes.

Similarly, a modest (10%) exaggeration of intervention effect estimates was observed for trials with inadequate/unclear (versus adequate) **concealment of allocation**. The average bias associated with inadequate allocation concealment was greatest in trials reporting subjective outcomes and in trials of complementary and alternative medicine, with no evidence of bias in trials of mortality or other objective outcomes. Evidence on baseline imbalances is scarcer. Although three empirical studies found no evidence that baseline imbalances inflate intervention effect estimates (31-33), all estimates were imprecise and these studies also found no evidence that randomization methods were associated with inflated intervention effect estimates. There was little evidence that intervention effect estimates were exaggerated in trials without adjustment for confounders (31) or in unblinded trials with block randomization, in which the last allocation in a block might be predictable (32). However, each characteristic was only examined in a single small study.

4.3 Using this domain of the tool

4.3.1 Assessing random sequence generation

The use of a random component should be sufficient for adequate sequence generation.

In principle, simple randomization can be achieved by allocating interventions using methods such as repeated coin-tossing, throwing dice, dealing previously shuffled cards, or by referring to a published list of random numbers (18, 19). More usually a list of random assignments is generated by a computer. Risk of bias may be judged in the same way whether or not a trial claims to have stratified its randomization.

> *Example of random sequence generation*: "We generated the two comparison groups using simple randomization, with an equal allocation ratio, by referring to a table of random numbers."
>
> *Example of random sequence generation*: "We used blocked randomization to form the allocation list for the two comparison groups. We used a computer random number generator to select random permuted blocks with a block size of eight and an equal allocation ratio."

Systematic methods, such as alternation, assignment based on date of birth, case record number and date of presentation, which are sometimes referred to as "quasi-random", are inadequate methods of sequence generation. Alternation (or rotation, for more than two intervention groups) might in principle result in similar groups, but many other systematic methods of sequence generation may not. For example, the day on which a patient is admitted to hospital is not solely a matter of chance. An important weakness with all systematic methods is that concealing the allocation schedule is usually impossible, which allows foreknowledge of intervention assignment among those recruiting participants to the study, and biased allocations.

> *Example of non-random sequence generation*: "Patients were randomized by the first letter of the last name of their primary resident (34)."
>
> *Example of non-random sequence generation*: "Those born on even dates were resuscitated with room air (room air group), and those born on odd dates were resuscitated with 100% oxygen (oxygen group) (35)."

4.3.1.1 Assessing sequence generation when insufficient information is provided about the methods used

A simple statement such as "we randomly allocated" or "using a randomized design" is often insufficient to be confident that the allocation sequence was genuinely randomized. Indeed, it is common for authors to use the term "randomized" even when it is not justified: many trials with declared systematic allocation have been described by the authors as "randomized". In some situations, a reasonable judgement may be made about whether a random sequence was used. For example, if the study was large, conducted by an independent trials unit or carried out for regulatory purposes, it may be reasonable to assume that the sequence was random and to answer 'Probably yes' to the signalling question. Alternatively, if other (contemporary) trials by the same investigator team have clearly used non-random sequences, it might be reasonable to assume that the current study was done using similar methods, and answer 'Probably no' to the signalling question. If users of the tool are not able (or insufficiently confident) to make such judgements, an answer of 'No information' should be provided.

Trial authors may describe their approach to sequence generation incompletely, without confirming that there was a random component. For example, authors may state that blocked allocation was used without describing the process of selecting the order of allocation within the blocks. In such instances, an answer of 'No information' should generally be provided.

4.3.2 Assessing concealment of allocation sequence

Among the methods used to conceal allocation, central randomization by a third party is the most desirable. Methods using envelopes are more susceptible to manipulation than other approaches (15, 21). If investigators use envelopes, they should develop and monitor the allocation process to preserve concealment. In addition to use of sequentially numbered, opaque, sealed envelopes, they should ensure that the envelopes are opened sequentially, and only after the envelope has been irreversibly assigned to the participant. When blocking is used, it be may be possible to predict the last intervention assignments within each block. This will be a problem when the person recruiting participants knows the start and end of each block and the allocations are revealed after assignment. The problem is likely to be more serious if block sizes are small and of equal sizes. In such situations, an answer of 'No' or 'Probably no' should be provided for the signalling question concerning whether allocations were concealed.

Table 2 provides minimal criteria for a judgement of adequate concealment of allocation sequence and extended criteria, which provide additional assurance that concealment of the allocation sequence was indeed adequate. Some examples of adequate approaches are provided in Box 3.

Table 2. Minimal and extended criteria for judging of allocation sequence to be concealed

Minimal criteria for a judgement of adequate concealment of the allocation sequence	Extended criteria providing additional assurance
Central randomization.	The central randomization office was remote from patient recruitment centres. Participant details were provided, for example, by phone (including interactive voice response systems), email or an interactive online system, and the allocation sequence was concealed to individuals staffing the randomization office until a participant was irreversibly registered.
Sequentially numbered drug containers.	Drug containers prepared by an independent pharmacy were sequentially numbered and opened sequentially. Containers were of identical appearance, tamper-proof and equal in weight.
Sequentially numbered, opaque, sealed envelopes.	Envelopes were sequentially numbered and opened sequentially only after participant details were written on the envelope. Pressure-sensitive or carbon paper inside the envelope transferred the participant's details to the assignment card. Cardboard or aluminium foil inside the envelope rendered the envelope impermeable to intense light. Envelopes were sealed using tamper-proof security tape.

Box 3. Examples of adequate allocation sequence concealment (as compiled by Schulz and Grimes (36))

"… that combined coded numbers with drug allocation. Each block of ten numbers was transmitted from the central office to a person who acted as the randomization authority in each centre. This individual (a pharmacist or a nurse not involved in care of the trial patients and independent of the site investigator) was responsible for allocation, preparation, and accounting of trial infusion. The trial infusion was prepared at a separate site, then taken to the bedside nurse every 24 h. The nurse infused it into the patient at the appropriate rate. The randomization schedule was thus concealed from all care providers, ward physicians, and other research personnel." (37).

"… concealed in sequentially numbered, sealed, opaque envelopes, and kept by the hospital pharmacist of the two centres." (38).

"Treatments were centrally assigned on telephone verification of the correctness of inclusion criteria…" (39).

"Glenfield Hospital Pharmacy Department did the randomization, distributed the study agents, and held the trial codes, which were disclosed after the study." (40).

4.3.3 Using baseline imbalance to identify problems with the randomization process

Baseline imbalances may be due to problems with the randomization process or due to chance (41). The RoB 2 tool includes consideration of situations in which baseline characteristics indicate that something may have gone wrong with the randomization process. It is important that baseline imbalances that are consistent with chance are not interpreted as evidence of risk of bias: see section 4.3.3.1.

Severe baseline imbalances may arise as a result of deliberate attempts to subvert the randomization process (42). They may also occur because of unintentional actions or errors that occurred due to insufficient safeguards: for example an error in a minimization programme such as writing a "plus" instead of a "minus", leading to maximizing instead of minimizing differences in one or more prognostic factors between groups.

Assessment of baseline imbalance should be based on data for all randomized participants. If baseline data are presented only for participants who completed the trial (or some other subset of randomized participants) then it is more difficult to assess baseline imbalance, and the proportion of and reasons for missing data need to be considered. The practice of reporting baseline characteristics of analysed participants only is not common in healthcare trials but may be common in other areas such as social care.

4.3.3.1 Chance imbalances at baseline

In trials using large samples (usually meaning at least 100 in each randomized group (18, 19, 22)), simple randomization generates intervention groups of relatively similar sizes. In trials using small samples, simple randomization will sometimes lead to groups that differ substantially, by chance, in size or in the distribution of prognostic factors (43).

Chance imbalances are not a source of systematic bias, and the **RoB 2 tool does not aim to identify imbalances in baseline variables that have arisen due to chance**. The 95% confidence interval for the effect of intervention incorporates the uncertainty arising from the potential for imbalances in prognostic factors between intervention groups (44). When chance baseline imbalances in prognostic factors occur, it is preferable to adjust for them in a pre-planned way (e.g. based on a rule specified in a trial analysis plan that is published before unblinded data are available to the investigators) (44).

Similarly, the average effect of chance imbalances across the trials included in a meta-analysis will be zero, and the confidence interval for the result incorporates their effect. The possible impact on a synthesis of studies with important chance imbalances across rather than within the studies needs to be considered outside of the study-specific risk of bias assessment.

4.3.3.2 Indications from baseline imbalance that there were problems with the randomization process

(1) Substantial differences between intervention group sizes, compared with the intended allocation ratio

One example is a 1948 trial comparing anticoagulation medication to conventional treatment for myocardial infarction (45). Anticoagulants were administered to patients admitted on odd admission dates (n=589) and conventional therapy to patients admitted on even admission dates (n=442). Such a large difference in numbers is unlikely given the expected 1:1 allocation ratio (P = 0.001), raising suspicion that investigators manipulated the allocation so that more patients were admitted on odd dates, when they would receive the new anticoagulant (45).

(2) A substantial excess in statistically significant differences in baseline characteristics between intervention groups, beyond that expected by chance

It is widely understood that statistical tests for differences in baseline characteristics should not be used in truly randomized trials, because the null hypothesis (that there are no systematic differences between the intervention groups) is known to be true. However, such tests can in principle be used to examine whether randomization was implemented successfully. *It is important that such evidence is interpreted appropriately*. Under randomization, one in 20 tests for baseline imbalance are expected to be statistically significant at a 5% level. If a substantially greater proportion of tests for baseline imbalance provide evidence of differences between intervention groups, or if P values are extremely small, this may suggest problems with the randomization process. However, it is possible that trial authors select the tests for baseline imbalance that are reported, either because they are statistically significant or because they are not statistically significant. Further, different prognostic factors may be correlated (for example a chance imbalance in age may lead to imbalance in other prognostic factors that are influenced by age). Therefore, review authors should be cautious in concluding that there is an excess of statistically significant differences between baseline characteristics.

(3) Imbalance in key prognostic factors, or baseline measures of outcome variables, that are unlikely to be due to chance

These are the factors that might influence those recruiting participants into the study, and therefore have most potential to be manipulated by investigators who (consciously or unconsciously) want to influence the trial results. The review team should, where possible, identify in advance the key prognostic factors that may influence the outcome of interest, for example through the knowledge of subject matter experts who are members of the review group, through initial (scoping) literature reviews, or through discussions with health professionals who make intervention decisions for the target patient or population groups. Based on this knowledge, imbalances in one or more key prognostic factors should be considered to place the study at high risk of bias if the P value for the between-intervention group difference is small enough that they are unlikely to

be due to chance (for example, <0.001) and the difference is big enough for the resulting confounding to bias the intervention effect estimate.

Plotting difference in baseline characteristics between intervention arms on a forest plot can be helpful way of visualizing baseline differences between intervention groups across studies. A methodological case study demonstrated that an apparent treatment effect was in fact due to baseline imbalances between intervention groups (46).

(4) Excessive similarity in baseline characteristics that is not compatible with chance

Excessive similarity across intervention groups may provide evidence of flawed or absent methods of randomization, if it is not compatible with the chance differences that arise through randomization. In an examination of baseline data from 5087 randomized trials, Carlisle observed more instances of baseline similarity than would be expected by chance, which could be explained by data fabrication among other reasons (47). Carlisle also observed that the proportion of trials with excessive similarity was higher among trials that had subsequently been retracted. Note that restricted randomization methods (see section 4.3.1) tend to give rise to groups that are more similar at baseline than simple randomization methods.

(5) Surprising absence of one or more key characteristics that would be expected to be reported

Lack of availability of baseline data for a key prognostic factor or baseline measures of the outcome may be a cause for concern. For example, a trial of nebulized magnesium sulphate versus placebo for the treatment of asthma exacerbations included a table of baseline characteristics which suggested that treatment groups were comparable based on a number of variables (48). However, the table did not include asthma severity, which is the most important prognostic factor for the treatment of asthma and it would be very unusual not to measure this at baseline. This may lead us to suspect that disease severity differed between groups at baseline in a manner that was not compatible with successful randomization.

4.3.4 Analyses that adjust for baseline imbalances

If trialists observe baseline imbalances between intervention groups, they may undertake analyses that attempt to control for these imbalances. For example, they may use a regression model to adjust for baseline values of prognostic variables or baseline values of the outcome variable. To remove the risk of bias caused by problems in the randomization process, it would be necessary to know, and measure, all the prognostic factors that were imbalanced at baseline (i.e. all the confounders). It is unlikely that all confounders are known and measured, so such analyses will at best reduce the risk of bias. If review authors wish to assess the risk of bias in a trial that controlled for baseline imbalances in order to mitigate failures of randomization, the study should be treated as non-randomized and assessed using the ROBINS-I tool (Risk of Bias in Non-randomized Studies of Interventions) (49).

4.4 Signalling questions and criteria for judging risk of bias

Signalling questions for this domain are provided in Box 4. Note that the answer to one signalling question should not affect answers to other questions. For example, if the trial has large baseline imbalances, but authors report adequate randomization methods, then sequence generation and allocation concealment should still be assessed on the basis of the reported adequate methods. Concerns about the imbalances should be reflected in the answer to the question about the baseline imbalance and reflected in the domain-level judgement.

Criteria for reaching risk-of-bias judgements are given in Table 3, and an algorithm for implementing these is provided in Table 4 and Figure 1. Suggested risk of bias judgements can be overridden if review authors believe this is justified: for example the importance of allocation concealment may depend on the extent to which potential participants in the study have different prognoses, whether strong beliefs exist among investigators and participants regarding the benefits or harms of assigned interventions, and whether uncertainty about the interventions is accepted by all people involved (42).

Box 4. The RoB 2 tool (part 2): Risk of bias arising from the randomization process

Signalling questions	Elaboration	Response options
1.1 Was the allocation sequence random?	Answer 'Yes' if a random component was used in the sequence generation process. Examples include computer-generated random numbers; reference to a random number table; coin tossing; shuffling cards or envelopes; throwing dice; or drawing lots. Minimization is generally implemented with a random element (at least when the scores are equal), so an allocation sequence that is generated using minimization should generally be considered to be random. Answer 'No' if no random element was used in generating the allocation sequence or the sequence is predictable. Examples include alternation; methods based on dates (of birth or admission); patient record numbers; allocation decisions made by clinicians or participants; allocation based on the availability of the intervention; or any other systematic or haphazard method. Answer 'No information' if the only information about randomization methods is a statement that the study is randomized. In some situations a judgement may be made to answer 'Probably no' or 'Probably yes'. For example, if the study was large, conducted by an independent trials unit or carried out for regulatory purposes, it may be reasonable to assume that the sequence was random. Alternatively, if other (contemporary) trials by the same investigator team have clearly used non-random sequences, it might be reasonable to assume that the current study was done using similar methods.	Y / PY / PN / N / NI
1.2 Was the allocation sequence concealed until participants were enrolled and assigned to interventions?	Answer 'Yes' if the trial used any form of remote or centrally administered method to allocate interventions to participants , where the process of allocation is controlled by an external unit or organization, independent of the enrolment personnel (e.g. independent central pharmacy, telephone or internet-based randomization service providers). Answer 'Yes' if envelopes or drug containers were used appropriately. Envelopes should be sequentially numbered, sealed with a tamper-proof seal and opaque. Drug containers should be sequentially numbered and of identical appearance. This level of detail is rarely provided in reports, and a judgement may be required to justify an answer of 'Probably yes' or 'Probably no'. Answer 'No' if there is reason to suspect that the enrolling investigator or the participant had knowledge of the forthcoming allocation.	Y / PY / PN / N / NI
1.3 Did baseline differences between intervention groups suggest a problem with the randomization process?	*Note that differences that are compatible with chance do not lead to a risk of bias.* Answer 'No' if no imbalances are apparent or if any observed imbalances are compatible with chance Answer 'Yes' if there are imbalances that indicate problems with the randomization process, including: (1) substantial differences between intervention group sizes, compared with the intended allocation ratio; or (2) a substantial excess in statistically significant differences in baseline characteristics between intervention groups, beyond that expected by chance; or (3) imbalance in one or more key prognostic factors, or baseline measures of outcome variables, that is very unlikely to be due to chance and for which the between-group difference is big enough to result in bias in the intervention effect estimate.	Y / PY / PN / N / NI

	Also answer 'Yes' if there are other reasons to suspect that the randomization process was problematic: (4) excessive similarity in baseline characteristics that is not compatible with chance; or (5) surprising absence of one or more key characteristics that would be expected to be reported. Answer 'No information' when there is no *useful* baseline information available (e.g. abstracts, or studies that reported only baseline characteristics of participants in the final analysis). The answer to this question should not influence answers to questions 1.1 or 1.2. For example, if the trial has large baseline imbalances, but authors report adequate randomization methods, questions 1.1 and 1.2 should still be answered on the basis of the reported adequate methods, and any concerns about the imbalance should be raised in the answer to the question 1.3 and reflected in the domain-level risk-of-bias judgement. Trialists may undertake analyses that attempt to deal with flawed randomization by controlling for imbalances in prognostic factors at baseline. To remove the risk of bias caused by problems in the randomization process, it would be necessary to know, and measure, all the prognostic factors that were imbalanced at baseline. It is unlikely that all important prognostic factors are known and measured, so such analyses will at best reduce the risk of bias. If review authors wish to assess the risk of bias in a trial that controlled for baseline imbalances in order to mitigate failures of randomization, the study should be assessed using the ROBINS-I tool.	
Risk-of-bias judgement	See Table 3, Table 4 and Figure 1.	Low / High / Some concerns
Optional: What is the predicted direction of bias arising from the randomization process?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

Table 3. Reaching risk-of-bias judgements for bias arising from the randomization process

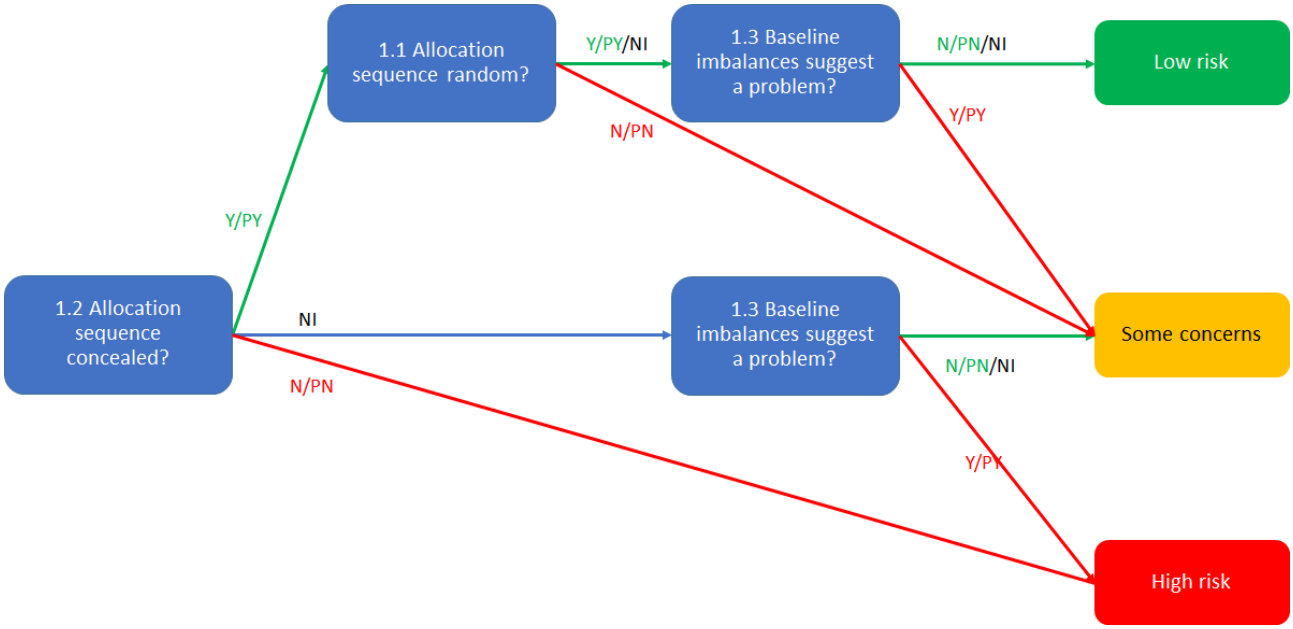
Low risk of bias	(i) The allocation sequence was adequately concealed AND (ii.1) Baseline differences between intervention groups appear to be compatible with chance OR (ii.2) There is no information about baseline imbalances AND (iii.1) The allocation sequence was random OR (iii.2) There is no information about concealment of the allocation sequence
Some concerns	(i.1) The allocation sequence was adequately concealed AND (i.2.1) The allocation sequence was <u>not</u> random OR (i.2.2) Baseline differences between intervention groups suggest a <u>problem</u> with the randomization process OR (ii.1) There is no information about concealment of the allocation sequence AND (ii.2) Baseline differences between intervention groups appear to be compatible with chance OR (iii) There is no information to answer any of the signalling questions
High risk of bias	(i) The allocation sequence was <u>not</u> adequately concealed OR (ii.1) There is no information about concealment of the allocation sequence AND (ii.2) Baseline differences between intervention groups suggest a <u>problem</u> with the randomization process

Table 4 Suggested mapping of signalling questions to risk-of-bias judgements for bias arising from the randomization process. This is only a suggested decision tree: all default judgements can be overridden by assessors.

Signalling question			Domain-level judgement	
1.1 Sequence random?	1.2 Allocation concealed?	1.3 Imbalance suggest problem?	Default risk of bias	Remarks
Y/PY/NI	Y/PY	NI/N/PN	Low	
Y/PY	Y/PY	Y/PY	Some concerns	There is considerable room for judgement here. Substantial baseline imbalance despite apparently sound randomization methods should be investigated carefully, and a judgement of 'Low' risk of bias or 'High' risk of bias might be reached.
N/PN/NI	Y/PY	Y/PY	Some concerns	Substantial baseline imbalance may lead to a judgement of 'High' risk of bias, especially if the method of sequence generation is also inappropriate.
Any response	NI	N/PN/NI	Some concerns	
Any response	NI	Y/PY	High	
Any response	N/PN	Any response	High	

Y/PY = 'Yes' or 'Probably yes'; N/PN = 'No' or 'Probably no'; NI = 'No information'

Figure 1. Suggested algorithm for reaching risk-of-bias judgements for bias arising from the randomization process.



5 Detailed guidance: bias due to deviations from intended interventions

5.1 Background

This domain relates to biases that arise when there are systematic differences, representing deviations from the intended interventions, between the care provided to experimental and comparator intervention groups. Such differences could reflect either additional aspects of care that were not intended or intended aspects of care that were not delivered. Biases that arise due to deviations from intended interventions are sometimes referred to as performance biases.

The intended interventions are those specified in the trial protocol. It is often intended that interventions should change or evolve in response to the health of, or events experienced by, trial participants. For example, the investigators may intend that:

- in a trial of a new drug to control symptoms of rheumatoid arthritis, participants experiencing severe toxicities should receive additional care and/or switch to an alternative drug;
- in a trial of a specified cancer drug regimen, participants whose cancer progresses should switch to a second-line intervention; or
- in a trial comparing surgical intervention with conservative management of stable angina, participants who progress to unstable angina receive surgical intervention.

Unfortunately, trial protocols may not fully specify or articulate the circumstances in which deviations from the initial intervention should occur, or distinguish changes to intervention that are consistent with the intentions of the investigators from those that should be considered as deviations from the intended intervention. For example, a cancer trial protocol may not define progression, or specify the second-line drug that should be used in patients who progress (50). It may therefore be necessary for users of RoB 2 to document changes to intervention that they do and do not consider to be deviations from intended intervention. Similarly, for trials in which the comparator intervention is "usual care", the protocol may not specify the interventions consistent with usual care or whether they are expected to be used alongside the experimental intervention. Users of the RoB 2 tool may therefore need to document what departures from usual care will be considered as deviations from intended intervention.

5.1.1 Important co-interventions

Important co-interventions (i.e. those relevant to this domain) are non-protocol interventions or exposures that trial participants might receive with or after starting the intervention of interest, that are affected by the assigned intervention and that are prognostic for the outcome of interest. If possible, likely co-interventions should be specified in advance (at review protocol writing stage). They may be identified through the expert knowledge of members of the review group, via initial (scoping) reviews of the literature, and through discussions with health professionals. **Co-interventions whose use is specified in the trial protocol, or which are part of routine care, should not be considered to be deviations from the intended intervention.**

5.1.2 The role of the effect of interest

As described in Section 1.3, assessments for this domain depend on whether the intervention effect of interest to the review authors is

(1) the effect of **assignment** to the interventions at baseline (regardless of whether the interventions are received or adhered to during follow-up, sometimes known as the 'intention-to-treat effect'); or
(2) the effect of **adhering to** the interventions as specified in the trial protocol (sometimes known as the 'per-protocol effect').

When the effect of interest is that of *assignment to the intervention*, deviations from intended interventions that reflect the natural course of events (for example, a deviation from intervention that was clinically necessary because of a sudden worsening of the patient's condition) do not lead to bias. However, bias will occur if there are deviations from the intended intervention that:

(1) do not reflect usual practice;
(2) are not balanced between the intervention groups; and
(3) influence the outcome.

For example, monitoring patients randomized to a novel intervention more closely than those randomized to standard care would increase the risk of bias, unless such monitoring was an intended part of the novel intervention. Although some might argue that this is an issue of generalizability of the result rather than bias, we regard the distortion in the particular context of a clinical trial as equivalent to bias.

To examine the effect of *adhering to the interventions* as specified in the trial protocol, we need to consider: (1) how well the intervention was implemented; (2) how well participants adhered to it (without discontinuing or switching to another intervention); (3) whether unintended co-interventions were received alongside the intended intervention and (if so) whether they were balanced across intervention groups; and (4) if such deviations are present, whether appropriate statistical methods were used to adjust for their effects when estimating the effect of adhering to intervention.

Some examples of studies in which there were deviations from the intended interventions are provided in Box 5.

5.1.3 The role of blinding

Bias due to deviations from intended interventions can sometimes be reduced or avoided by implementing mechanisms that ensure the participants, carers and trial personnel (i.e. people delivering the interventions) are unaware of the interventions received. This is commonly referred to as 'blinding', although in some areas (including eye health) the term 'masking' is preferred. Blinding, if successful, should prevent knowledge of the intervention assignment from influencing co-interventions, contamination (application of one of the interventions in participants intended to receive the other), switches from the assigned interventions to interventions that were not specified in the trial protocol, or failure to implement the intervention as intended.

For some questions, blinding is essential for meaningful evaluation of the experimental intervention. For example, trials of acupuncture to treat pain tend to find benefit when it is compared with no treatment, but no important benefit when the comparison is with sham acupuncture and the participants and carers (other than those delivering the intervention) are blinded (51). In situations such as this, meaningful comparisons should use blinding to eliminate placebo effects and isolate the specific hypothesised effect of the intervention. Similarly, a blinded comparison of a drug with placebo allows estimation of the pharmacological effect of the compound concerned, whereas the effect estimated from a comparison of the drug with no intervention combines the effect of the compound with that of the whole treatment process.

If assignment of intervention was not concealed at the time of randomization (see section 4.1.2), then knowledge of the allocation may also be available during the conduct of the trial, so that carers and people delivering the interventions are not fully blinded.

Blinding during a trial can be impossible in some contexts, for example in a trial comparing a surgical with a non-surgical intervention, and is not appropriate when the goal is to compare treatment strategies in individuals who are fully aware of their care, as is often the case in pragmatic trials. Studies of these sorts of interventions might take other measures to reduce the risk of bias, such as treating patients according to a strict protocol to reduce the risk that co-interventions (beyond those specified in the trial protocol) are not balanced between intervention groups. **When interest is in the effect of assignment to intervention, absence of blinding need not lead to bias, providing that all deviations from the intended intervention reflect the care that would routinely be received outside the context of a trial**.

Lack of blinding of participants, carers or people delivering the interventions may cause bias if it leads to unbalanced co-interventions that affect trial outcomes, or switches to alternative interventions. For example, low expectations of improvement among participants in the comparator group may lead them to seek and receive the experimental intervention. Such deviations from intended intervention that arise specifically due to the experimental context lead to bias in the estimated effects of both assignment to intervention and of adhering to intervention, although bias in the latter effect may be ameliorated through the statistical methods used in its estimation (see section 1.3).

Blinding of outcome assessors, to avoid bias in *measuring* the outcome, is considered separately, in the 'Bias in measurement of outcomes' domain. Bias due to differential rates of drop out (withdrawal from the study) is considered in the 'Bias due to missing outcome data' domain.

An attempt to blind participants, carers and people delivering the interventions to intervention group does not ensure successful blinding in practice. For many blinded drug trials, the side effects of the drugs allow the possible detection of the intervention being received for some participants, unless the study compares similar interventions, e.g. drugs with similar side effects, or uses an active placebo (52-54).

Several groups have suggested that it would be sensible to ask trial participants at the end of the trial to guess which intervention they had been receiving (55, 56), and some reviews of such reports have been published (55, 57). Evidence of correct guesses exceeding 50% can simply reflect patients' experiences in the trial: a good outcome, or a marked side effect, will tend to be more often attributed to an active intervention, and a poor outcome to a placebo (58). It follows that we would expect to see some successful "guessing" when there is a difference in either efficacy or adverse effects, but none when the interventions have very similar effects, even when the blinding has been preserved.

Deducing the intervention received, for example among participants experiencing side effects that are specific to the experimental intervention, does not in itself lead to a risk of bias. As discussed above, cessation of a drug intervention because of toxicity will usually not be considered a deviation from intended intervention. Please see the elaborations that accompany the signalling questions for further discussion of this issue.

Study reports often describe blinding in broad terms, such as "double blind". This term makes it difficult to know who was blinded (17). Such terms are also used very inconsistently (59-61), and the frequency of explicit reporting of the blinding status of study participants and trial personnel remains low even in trials published in top journals (62), despite recommendations in the CONSORT Statement to be explicit (63). A review of methods used for blinding highlights the variety of methods used in practice (52).

Box 5. Examples of studies with deviations from the intended interventions

Example 1: substantial numbers of patients not treated as randomized
To determine the efficacy of surgery for lumbar intervertebral disc herniation, the SPORT trial (Spine Patient Outcomes Research Trial) randomized patients with lumbar disc herniation to receive surgical treatment (discectomy) or non-operative care (encompassing a variety of interventions including analgesics, education, physiotherapy and acupuncture) (64). An ITT analysis found no evidence that the primary outcome (Short Form-36 bodily pain and physical function scales) differed between intervention groups two years after randomization. However, by that time only 60% of patients assigned to surgical treatment had undergone the procedure and 45% of those in the non-operative group had been treated surgically, and the authors performed an 'as treated' analysis which, in contrast to ITT, showed advantages for surgery. This leads to a biased estimate of the effect of adhering to intervention (the 'per-protocol effect') because baseline characteristics of participants who underwent surgery (irrespective of their assigned intervention) differed substantially from characteristics of those who did not.

Example 2: crossover from comparator group to experimental intervention group
To determine the efficacy of percutaneous coronary interventions, the FAME 2 trial (Fractional Flow Reserve versus Angiography for Multivessel Evaluation 2 trial) randomized patients with stable, functionally-significant, coronary artery disease to a percutaneous coronary intervention (PCI) with implantation of drug-eluting stents and optimal medical therapy (OMT) or to OMT alone (65). The trial was stopped early because of clear evidence from an ITT analysis that participants assigned to PCI were at reduced risk of the primary composite outcome of death from any cause, nonfatal myocardial infarction, or urgent revascularization (hazard ratio [HR] 0.39; 95% CI 0.26 to 0.57; p<0.001). This was driven by a pronounced 77% reduction in urgent revascularization in the PCI group (HR 0.23; 95% CI 0.14 to 0.38; P<0.001). In contrast, there was only a 21% reduction in the composite of death or myocardial infarction (HR 0.79; 95% CI 0.49 to 1.29). However, 41% of patients allocated to OMT had 'crossed over' to PCI: 40% of these PCI procedures were performed because they were clinically indicated (following a myocardial infarction or unstable angina) but the majority were deviations from the trial protocol. Therefore, it is likely that the effect of assignment to PCI versus OMT (the ITT effect) on death or myocardial infarction was closer to the null than the effect of adhering to the intervention (the 'per-protocol effect'). To estimate this effect, it would be at minimum be necessary to censor follow up when non-clinically-indicated PCIs occurred in participants allocated to OMT, and adjust the analysis (for example, through inverse probability weighting) to account for the censoring.

Example 3: co-interventions not balanced between intervention groups and likely to affect the outcome
An open-label study compared respiratory tract infection (RTI) rates after minimally invasive or open surgery for oesophageal cancer (66). There were two important differences between intervention groups in the delivery of co-interventions. First, one-lung mechanical ventilation (which is thought to increase respiratory complications, including RTIs) was used in the open surgery group, whereas the minimally invasive group underwent two-lung ventilation (note that both types of mechanical ventilation could have been used for either intervention). Second, epidural analgesia was used more frequently in the open surgery group: patients with epidurals are generally less mobile and thus at increased risk of developing an RTI. If some of these co-interventions were administered because of the experimental context then the estimated effect of assignment to intervention was at risk of bias, because the co-interventions were not balanced between intervention groups and likely to impact on the outcome.

Example 4: interventions not implemented successfully
Patients with acute appendicitis were randomized to undergo standard laparoscopic appendicectomy or single incision laparoscopic appendicectomy (67). Standard laparoscopic surgery involves three separate incisions (through which two laparoscopic instruments and the laparoscopic camera are placed), whereas single incision surgery involves just one incision through which the camera and one instrument are placed. However, fewer than 50% of the patients randomized to undergo single incision surgery received this procedure as intended: additional instruments and incisions were required in 51% and 10% of procedures, respectively. ITT analyses found no strong evidence of between-intervention differences in the risk of wound infection (primary outcome), time to return to normal diet, length of hospital stay, time to return to normal activity, or duration of post-operative analgesia. Such analyses are appropriate to estimate the effect of assignment to intervention (the ITT effect), but this effect may be closer to the null than the effect of adhering to intervention (the per-protocol effect) because of the failure to implement the intervention successfully.

5.2 Empirical evidence of bias due to deviations from intended interventions

Empirical evidence of bias due to deviations from the intended interventions largely comes from studies exploring whether reporting of "double blinding" is associated with intervention effects. In the largest meta-epidemiological study conducted to date, lack of or unclear double blinding (versus double blinding) in trials with subjectively assessed outcomes was associated with a 23% exaggeration of odds ratio (68). Lack of or unclear double blinding was also associated with increased between-trial heterogeneity. By comparison, there was little evidence of such bias in trials of mortality or other objectively assessed outcomes, in a meta-analysis of meta-epidemiological studies (30). Two other studies examining subjectively measured continuous outcomes (e.g. patient-rated questionnaires) found that standardized mean differences tended to be exaggerated in trials with lack of or unclear blinding of participants (versus blinding of participants) (69, 70). Because existing empirical evidence does not distinguish blinding of participants and trial personnel from blinding of outcome assessors, it not clear which of these aspects of blinding is most important in preventing bias.

Naïve 'per-protocol' analyses have been found to exaggerate estimates of the effect of assignment to intervention compared with 'intention-to-treat' analyses (71, 72). Reporting the use of a 'modified ITT' analysis (versus ITT) has been associated with exaggerated effects (73). Tierney et al. observed a tendency for analyses conducted after trial authors excluded participants to favour the experimental intervention, compared with analyses including all participants (74).

Interpretation of empirical studies is difficult because exclusions are often poorly reported, particularly in the pre-CONSORT era before 1996. For example, Schulz observed that the *apparent* lack of exclusions was associated with more "beneficial" effect sizes as well as with less likelihood of adequate allocation concealment (15). Hence, failure to report exclusions in trials in Schulz's study may have been a marker of poor trial conduct rather than true absence of any exclusions.

5.3 Using this domain of the tool

5.3.1 The effect of assignment to intervention (the 'intention to treat effect')

The signalling questions address:

(1) whether participants, carers and people delivering the interventions were blinded;
(2) if some of these groups were not blinded, whether deviations from intended intervention that do not reflect usual practice are likely to have biased the intervention effect; and
(3) whether an appropriate analysis was used to estimate the effect of assignment to intervention.

An appropriate analysis should follow the principles of ITT that participants are analysed in the intervention groups to which they were randomized (regardless of the intervention they actually received) and that all randomized participants in the analysis (regardless of whether the interventions were implemented as intended and regardless of adherence of the participants). Some authors may report a 'modified intention-to-treat' (mITT) analysis in which participants with missing outcome data are excluded. Such an analysis may be biased because of the missing outcome data: this is addressed in the corresponding domain. Note that the phrase 'modified intention-to-treat' is used in different ways, and may refer to inclusion of participants who received at least one dose of treatment (75); our use of the term refers to missing data rather than to adherence to intervention.

Inappropriate analyses include 'as-treated' analyses, naïve 'per-protocol' analyses, and other analyses based on excluding eligible trial participants post-randomization (11) (see also section 1.3.1). Other inappropriate reasons for excluding eligible trial participants post-randomization include that they experienced toxicities resulting from intervention or even "lack of efficacy". However, post-randomization exclusions of ineligible participants (when eligibility was not confirmed until after randomization and could not have been influenced by intervention group assignment) can be considered appropriate.

Note that it might be possible to conduct individual participant data meta-analyses that include participants who were excluded by the study authors ("re-inclusions"). Review authors are encouraged to do this when possible. In this situation, the risk of bias assessment should apply to the result of the trial as it is included in the synthesis, rather than the result as it is reported by the trialists.

It should be straightforward to answer the signalling questions about analysis, providing that the trial is reported in accordance with the CONSORT statement, and includes a CONSORT flow chart (63). Reports of 'as-treated' analyses are uncommon. Reports of naïve 'per-protocol' analyses are more common, although they are often reported in addition to, rather than instead of, an ITT or mITT analysis. Results of ITT analyses should be

preferred for inclusion in risk of bias assessments and meta-analyses, when a choice is available (see section 1.3.1). When the result being assessed is based on an 'as treated' or naïve 'per-protocol' analysis, review authors should assess whether there is potential for a substantial impact on the estimated effect of intervention. A 'per protocol' analysis should be selected in preference to an 'as treated' analysis.

Risk of bias in this domain may be differ between outcomes, even if the same people were aware of intervention assignments during the trial. For example, knowledge of the assigned intervention may impact on behaviour (such as number of clinic visits), while not impacting importantly on physiology (including risk of mortality).

5.3.2 The effect of adhering to intervention (the 'per-protocol effect')

The signalling questions address:

(1) whether participants, carers and people delivering the interventions were blinded;
(2) if participants, carers or people delivering the interventions were not blinded, whether important co-interventions were balanced across intervention groups;
(3) whether the intervention was implemented successfully, and whether study participants adhered to the assigned intervention;
(4) if deviations from intended intervention arising from points 2 and 3 above occurred, whether an appropriate analysis was used.

Appropriate analysis approaches are described by Hernán and Robins (13): for example instrumental variable approaches can be used in some circumstances to estimate the effect of intervention among participants who received the assigned intervention (see also section 1.3.1).

5.4 Signalling questions and criteria for judging risk of bias

Signalling questions for the effect of assignment to intervention are provided in Box 6. Criteria for reaching risk-of-bias judgements are given in Table 5, and an algorithm for implementing these is provided in Table 6 and Figure 2. Signalling questions for the effect of adhering to intervention are provided in Box 7. Criteria for reaching risk-of-bias judgements are given in Table 7, and an algorithm for implementing these is provided in Table 8 and Figure 3.

Box 6. The RoB 2 tool (part 3): Risk of bias due to deviations from the intended interventions (*effect of assignment to intervention*)

Signalling questions	Elaboration	Response options
2.1. Were participants aware of their assigned intervention during the trial?	If participants are aware of their group assignment, it is more likely that additional health-related behaviours will differ between the assigned intervention groups. Blinding participants, which is most commonly achieved through use of a placebo or sham intervention, may prevent such differences. If participants experienced side effects or toxicities that they knew to be specific to one of the interventions, the answer to this question should be 'Yes' or 'Probably yes'.	Y / PY / PN / N / NI
2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	If those involved in caring for participants or making decisions about their health care are aware of the assigned intervention, then implementation of the intended intervention, or administration of additional co-interventions, may differ between the assigned intervention groups. Blinding carers and people delivering the interventions, which is most commonly achieved through use of a placebo, may prevent such differences. If participants experienced side effects or toxicities that carers or people delivering the interventions knew to be specific to one of the interventions, the answer to this question should be 'Yes' or 'Probably yes'. If randomized allocation was not concealed, then it is likely that carers and people delivering the interventions were aware of participants' assigned intervention during the trial.	Y / PY / PN / N / NI
2.3. <u>If Y/PY/NI to 2.1 or 2.2</u>: Were there deviations from the intended intervention that arose because of the experimental context?	When interest focusses on the effect of assignment to intervention, bias (in this domain) only arises when there are deviations that arose because of the experimental context, i.e. due to expectations of a difference between experimental and comparator intervention. For example, participants may feel unlucky to have been assigned to the comparator group and therefore seek the experimental intervention, or other interventions. For such deviations, answer 'Yes' or 'Probably yes'. Because deviations that arise due to expectations of a difference between experimental and comparator intervention are not part of usual practice, they may lead to effect estimates that do not reflect what would happen to participants assigned to the interventions in practice. Answer 'No' or 'Probably no' for deviations that are expected to arise in usual care, for example: (1) Cessation of a drug intervention because of acute toxicity; (2) Non-adherence to intervention, or changes to intervention, that are typical of routine care so unrelated to the experimental context; (3) Co-interventions whose aim is to treat consequences of one of the interventions. In some trials blinding is compromised because participants report side effects or toxicities that are specific to one of the interventions. In these situations the answer to this question should be 'No' or 'Probably no' unless there were deviations from intended intervention that arose because of the experimental context. Cessation of or changes to intervention in response to side effects or toxicities will not usually be considered to be a deviation from intended intervention. Trialists do not always report (and do not necessarily know) whether deviations arose because of the experimental context. Therefore, the answer 'No information' may be appropriate. However, if such deviations *probably* occurred the answer should be 'Probably yes'.	NA / Y / PY / PN / N / NI
2.4. <u>If Y/PY to 2.3</u>: Were these deviations from intended intervention	Deviations from intended interventions that do not reflect usual practice (i.e. were due to the experimental context) will be important if there is imbalance in the deviations across the two groups, but not otherwise.	NA / Y / PY / PN / N / NI

balanced between groups?		
2.5 <u>If N/PN/NI to 2.4</u>: Were these deviations likely to have affected the outcome?	Deviations from intended interventions that do not reflect usual practice (i.e. were due to the experimental context) will be important if they affect the outcome, but not otherwise.	NA / Y / PY / PN / N / NI
2.6 Was an appropriate analysis used to estimate the effect of assignment to intervention?	Both intention-to-treat (ITT) analyses and modified intention to treat (mITT) analyses excluding participants with missing outcome data should be considered appropriate. (Missing outcome data are addressed in a separate domain). Both 'as treated' analyses (in which trial participants are grouped according to the intervention that they received, rather than according to their assigned intervention) and naïve 'per-protocol' analyses (excluding trial participants who did not receive their assigned intervention) should be considered inappropriate. Analyses excluding eligible trial participants post-randomization should also be considered inappropriate, but post-randomization exclusions of ineligible participants (when eligibility was not confirmed until after randomization, and could not have been influenced by intervention group assignment) can be considered appropriate.	Y / PY / PN / N / NI
2.7 <u>If N/PN/NI to 2.6:</u> Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomized?	This question addresses whether the number of participants who were analysed in the wrong intervention group, or excluded from the analysis, was sufficient that there could have been a substantial impact on the result. It is not possible to specify a precise rule: there may be potential for substantial impact even if fewer than 5% of participants were analysed in the wrong group or excluded, if the outcome is rare or if exclusions are strongly related to prognostic factors.	NA / Y / PY / PN / N / NI
Risk-of-bias judgement	See Table 5, Table 6 and Figure 2.	Low / High / Some concerns
Optional: What is the predicted direction of bias due to deviations from intended interventions?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

Table 5. Reaching risk-of-bias judgements for bias due to deviations from intended intervention (*effect of assignment to intervention*)

Because the domain addresses two somewhat distinct issues, we separate the algorithm into two parts and combine them to reach the judgement.

	Part 1: criteria for questions 2.1 to 2.5	Part 2: criteria for questions 2.6 and 2.7	Criteria for the domain
Low risk of bias	(i) Participants, carers and people delivering the interventions were unaware of intervention groups during the trial OR (ii.1) Participants, carers or people delivering the interventions were aware of intervention groups during the trial AND (ii.2)No deviations from intended intervention arose because of the experimental context.	An appropriate analysis was used to estimate the effect of assignment to intervention	'Low' risk of bias for Part 1 AND 'Low' risk of bias for Part 2
Some concerns	(i) Participants, carers or people delivering the interventions were aware of intervention groups during the trial AND (ii.1) There is no information on whether there were deviations from intended intervention because of the experimental context OR (ii.1.1) There were deviations from intended interventions that arose because of the experimental context AND (ii.1.1.1) These deviations were balanced between the intervention groups	(i) An appropriate analysis was not used to estimate the effect of assignment to intervention AND (ii) The potential impact (on the estimated effect of intervention) of the failure to analyse participants in the group to which they were randomized was not substantial	'Low' risk of bias or 'Some concerns' for Part 1 AND 'Low' risk of bias or 'Some concerns' for Part 2

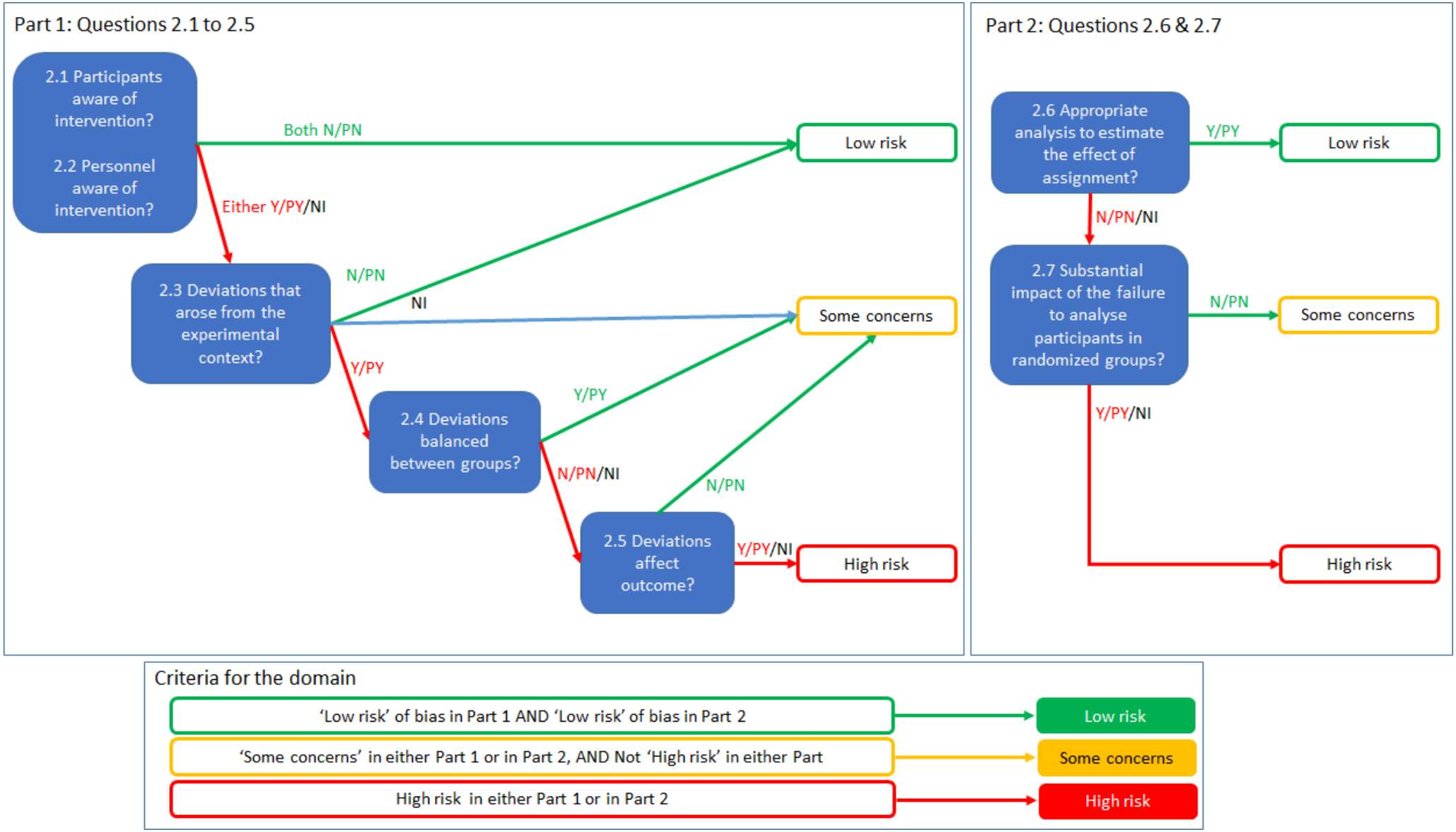
	OR (ii.1.1.2) These deviations were not likely to have affected the outcome		
High risk of bias	(i) Participants, carers or people delivering the interventions were aware of intervention groups during the trial AND (ii) There were deviations from intended interventions that arose because of the experimental context AND (iii) These deviations were unbalanced between the intervention groups AND (iv) These deviations were likely to have affected the outcome	(i) An appropriate analysis was not used to estimate the effect of assignment to intervention AND (ii) The potential impact (on the estimated effect of intervention) of the failure to analyse participants in the group to which they were randomized was substantial	'High' risk of bias for Part 1 OR 'High' risk of bias for Part 2

Table 6. Suggested mapping of signalling questions to risk-of-bias judgements for bias due to deviations from intended interventions (*effect of assignment to intervention*). This is only a suggested decision tree: all default judgements can be overridden by assessors.

Signalling question					Domain level judgement
Part 1: Questions 2.1 to 2.5					
2.1	**2.2**	**2.3**	**2.4**	**2.5**	
Participants aware?	**Personnel aware?**	**Any deviations?**	**Balanced deviations?**	**Affecting outcomes?**	**Default risk of bias for part 1**
Both 2.1 & 2.2 N/PN		NA	NA	NA	Low
Either 2.1 or 2.2 Y/PY/NI		N/PN	NA	NA	Low
Either 2.1 or 2.2 Y/PY/NI		NI	NA	NA	Some concerns
Either 2.1 or 2.2 Y/PY/NI		Y/PY	Y/PY	NA	Some concerns
Either 2.1 or 2.2 Y/PY/NI		Y/PY	N/PN/NI	N/PN	Some concerns
Either 2.1 or 2.2 Y/PY/NI		Y/PY	N/PN/NI	N/PN/NI	High
Part 2: Questions 2.6 and 2.7					
2.6 Appropriate analysis?		**2.7 Potential impact on result due to switching groups in analysis?**			**Default risk of bias for part 2**
Y/PY		NA			Low
N/PN/NI		N/PN			Some concerns
N/PN/NI		Y/PY/NI			High
Criteria for the domain					
'Low' risk of bias in Part 1 AND 'Low' risk of bias in Part 2					Low
'Some concerns' in either Part 1 OR in Part 2, AND NOT 'High' risk in either part					Some concerns
'High' risk of bias in in either Part 1 OR in Part 2					High

Y/PY = 'Yes' or 'Probably yes'; N/PN = 'No' or 'Probably no'; NI = 'No information'; NA = Not applicable

Figure 2. Suggested algorithm for reaching risk-of-bias judgements for bias due to deviations from intended interventions (*effect of assignment to intervention*).



Box 7. The RoB 2 tool (part 4): Risk of bias due to deviations from the intended interventions (*effect of adhering to intervention*)

Signalling questions	Elaboration	Response options
2.1. Were participants aware of their assigned intervention during the trial?	If participants are aware of their group assignment, it is more likely that additional health-related behaviours will differ between the intervention groups. Blinding participants, which is most commonly achieved through use of a placebo or sham intervention, may prevent such differences.	Y / PY / PN / N / NI
2.2. Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?	If those involved in caring for participants or making decisions about their health care are aware of assigned intervention, then implementation of the intended intervention, or administration of additional co-interventions, may differ between the intervention groups. Blinding carers and people delivering the interventions, which is most commonly achieved through use of a placebo, may prevent such differences.	Y / PY / PN / N / NI
2.3. <u>If Y/PY/NI to 2.1 or 2.2</u>: Were important co-interventions balanced across intervention groups?	Important co-interventions are the interventions or exposures (1) that are inconsistent with the trial protocol, (2) that trial participants might receive with or after starting their assigned intervention, (3) that may be related to the intervention received and (4) that are prognostic for the outcome. Bias will arise if there is imbalance in such co-interventions between the intervention groups.	NA / Y / PY / PN / N / NI
2.4. Could failures in implementing the intervention have affected the outcome?	Answer 'No' or 'Probably no' if implementation of the intervention was successful for most participants.	Y / PY / PN / N / NI
2.5. Did study participants adhere to the assigned intervention regimen?	Lack of adherence includes imperfect compliance with a sustained intervention, cessation of intervention, crossovers to the comparator intervention and switches to another active intervention. Consider available information on the proportion of study participants who continued with their assigned intervention throughout follow up, and answer 'No' or 'Probably no' if the proportion who did not adhere is high enough to raise concerns. Answer 'Yes' for studies of interventions that are administered once, so that imperfect adherence is not possible, and if all or most participants received the assigned intervention.	Y / PY / PN / N / NI
2.6. <u>If N/PN/NI to 2.3 or 2.5 or Y/PY/NI to 2.4</u>: Was an appropriate analysis used to estimate the effect of adhering to the intervention?	Both 'as treated' analyses (comparing trial participants according to the intervention they actually received) or naïve 'per-protocol' analyses (excluding trial participants who did not receive their allocated intervention) will usually be inappropriate for estimating the effect of adhering to intervention (the 'per-protocol' effect). However, it is possible to conduct appropriate analyses in the presence of some types of deviation from the intended intervention. Examples include: (1) instrumental variable analyses to estimate the effect of receiving the assigned intervention in trials in which a single intervention, administered at baseline, is compared with standard of care; and (2) inverse probability weighting to adjust for censoring of participants who cease adherence to their assigned intervention, in trials of sustained treatment strategies.	NA / Y / PY / PN / N / NI

	Such methods depend on strong assumptions, which should be appropriate and justified if the answer to this question is Yes or Probably Yes. If an important co-intervention was administered to all participants in one intervention group, adjustments cannot be made to overcome this. Some examples of analysis strategies that would not be appropriate to estimate the effect of adhering to intervention are (i) 'Intention to treat (ITT) analysis', (ii) 'per protocol analysis', (iii) 'as-treated analysis', (iv) 'analysis by treatment received'.	
Risk-of-bias judgement	See Table 7, Table 8 and Figure 3	Low / High / Some concerns
Optional: What is the predicted direction of bias due to deviations from intended interventions?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

Table 7. Reaching risk-of-bias judgements for bias due to deviations from intended interventions (*effect of adhering to intervention*)

Low risk of bias	(i.1) Participants, carers and people delivering the interventions were unaware of intervention groups during the trial OR (i.2.1) Participants, carers or people delivering the interventions were aware of intervention groups AND (i.2.2) The important co-interventions were balanced across intervention groups AND (ii) Failures in implementing the intervention could not have affected the outcome AND (iii) Study participants adhered to the assigned intervention regimen
Some concerns	(i.1.1) Participants, carers and people delivering the interventions were unaware of intervention groups during the trial AND (i.1.2.1) Failures in implementing the intervention <u>could</u> have affected the outcome OR (i.1.2.2) Study participants did <u>not</u> adhere to the assigned intervention regimen OR (i.2.1) Participants, carers or people delivering the interventions were <u>aware</u> of intervention groups AND (i.2.2) The important co-interventions were balanced across intervention groups AND (i.2.3.1) Failures in implementing the intervention <u>could</u> have affected the outcome OR (i.2.3.2) Study participants did <u>not</u> adhere to the assigned intervention regimen OR (i.3.1) Participants, carers or people delivering the interventions were <u>aware</u> of intervention groups AND (i.3.2) The important co-interventions were <u>not</u> balanced across intervention groups AND (ii) An appropriate analysis was used to estimate the effect of adhering to the intervention
High risk of bias	(i.1.1) Participants, carers and people delivering the interventions were unaware of intervention groups during the trial AND (i.1.2.1) Failures in implementing the intervention <u>could</u> have affected the outcome

OR

(i.1.2.2) Study participants did not adhere to the assigned intervention regimen

OR

(i.2.1) Participants, carers or people delivering the interventions were aware of intervention groups

AND

(i.2.2) The important co-interventions were balanced across intervention groups

AND

(i.2.3.1) Failures in implementing the intervention could have affected the outcome

OR

(i.2.3.2) Study participants did not adhere to the assigned intervention regimen

OR

(i.3.1) Participants, carers or people delivering the interventions were aware of intervention groups

AND

(i.3.2) The important co-interventions were not balanced across intervention groups

AND

(ii) An appropriate analysis was not used to estimate the effect of adhering to the intervention

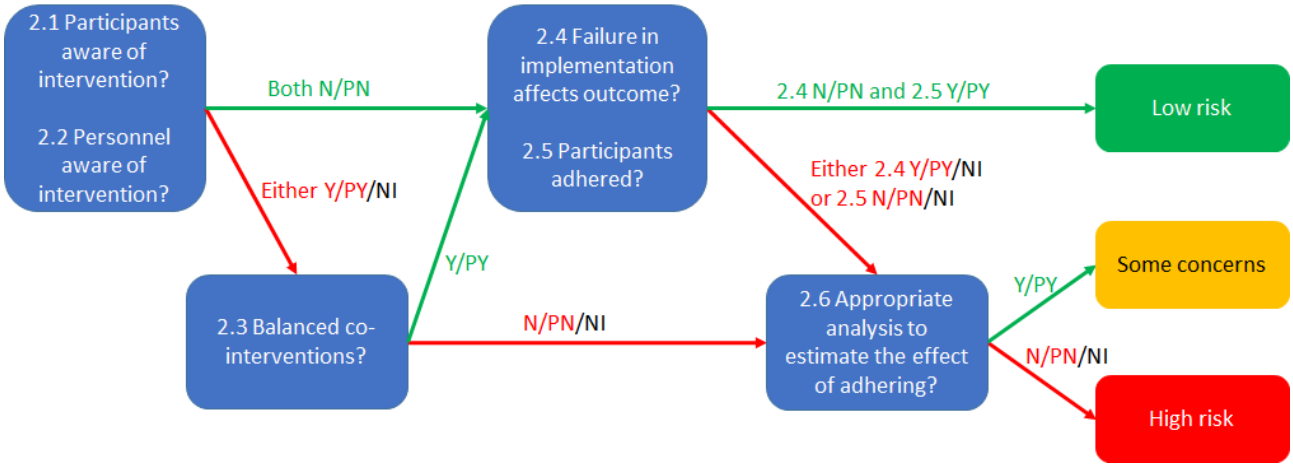
Table 8. Suggested mapping of signalling questions to risk-of-bias judgements for bias due to deviations from intended interventions (*effect adhering to intervention*). This is only a suggested decision tree: all default judgements can be overridden by assessors.

Signalling question						Domain level judgement
2.1	2.2	2.3	2.4	2.5	2.6	
Participant aware?	**Personnel aware?**	**Balanced co-ints?**	**Failure in implementation affects outcome?**	**Participants adhered?**	**Appropriate analysis?**	**Default risk of bias**
Both 2.1 & 2.2 N/PN		NA	N/PN	Y/PY	NA	Low
Either 2.1 or 2.2 Y/PY/NI		Y/PY	N/PN	Y/PY	NA	Low
Both 2.1 & 2.2 N/PN		NA	Y/PY/NI	Y/PY	Y/PY	Some concerns
Both 2.1 & 2.2 N/PN		NA	N/PN	N/PN/NI	Y/PY	Some concerns
Either 2.1 or 2.2 Y/PY/NI		Y/PY	Y/PY/NI	Y/PY	Y/PY	Some concerns
Either 2.1 or 2.2 Y/PY/NI		Y/PY	N/PN	N/PN/NI	Y/PY	Some concerns
Either 2.1 or 2.2 Y/PY/NI		N/PN/NI	Any response		Y/PY	Some concerns
Both 2.1 & 2.2 N/PN		NA	Y/PY/NI	Y/PY	N/PN/NI	High
Both 2.1 & 2.2 N/PN		NA	N/PN	N/PN/NI	N/PN/NI	High
Either 2.1 or 2.2 Y/PY/NI		Y/PY	Y/PY/NI	Y/PY	N/PN/NI	High
Either 2.1 or 2.2 Y/PY/NI		Y/PY	N/PN	N/PN/NI	N/PN/NI	High

Either 2.1 or 2.2 Y/PY/NI	N/PN/NI	Any response	N/PN/NI	High

Y/PY = 'Yes' or 'Probably yes'; N/PN = 'No' or 'Probably no'; NI = 'No information'; NA = "Not applicable"

Figure 3. Suggested algorithm for reaching risk-of-bias judgements for bias due to deviations from intended interventions (*effect adhering to intervention*). This is only a suggested decision tree: all default judgements can be overridden by assessors.



6 Detailed guidance: bias due to missing outcome data

6.1 Background

Randomization provides a fair comparison between two or more intervention groups by balancing, on average, the distribution of known and unknown prognostic factors at baseline between the intervention groups. Missing measurements of the outcome, for example due to dropout during the study, may lead to bias in the intervention effect estimate.

Possible reasons for missing outcome data include (76):

- participants withdraw from the study or cannot be located ('loss to follow-up' or 'dropout');
- participants do not attend a study visit at which outcomes should have been measured;
- participants attend a study visit but do not provide relevant data;
- data or records are lost or are unavailable for other reasons; and
- participants can no longer experience the outcome, for example because they have died.

This domain addresses risk of bias due to missing outcome data, including biases introduced by procedures used to impute, or otherwise account for, the missing outcome data.

Some participants may be excluded from an analysis for reasons other than missing outcome data. In particular, a naïve 'per protocol' analysis is restricted to participants who received the intended intervention (see section 1.3.1). Potential bias introduced by such analyses, or by other exclusions of eligible participants for whom outcome data are available, is addressed in the domain 'Bias due to deviations from intended interventions' (see section 5), in which the final signalling questions examine whether the analysis approach was appropriate. **This is a notable change from the previous Cochrane RoB tool for randomized trials**, in which the domain addressing bias due to incomplete outcome data addressed both genuinely missing data and data deliberately excluded by the trial investigators.

6.1.1 Missing outcome data when estimating the effect of assignment to intervention

The effect of assignment to intervention should be estimated using an intention-to-treat (ITT) analysis (5). As noted in section 1.3.1, the principles underlying ITT analyses are (6, 7):

(1) analyse participants in the intervention groups to which they were randomized, regardless of the intervention they actually received;
(2) include all randomized participants in the analysis; and
(3) measure outcome data on all participants.

While the first and second principles can always be followed for participants for whom outcome data are available, measuring outcome data on all participants is frequently difficult or impossible to achieve in practice. Therefore, it can often only be followed by making assumptions about the missing values of the outcome.

Even when an analysis is described as ITT, it may exclude participants with missing outcome data, and so be at risk of bias (such analyses may be described as 'modified intention-to-treat' (mITT) analyses). Therefore, assessments of risk of bias due to missing outcome data should be based on the issues addressed in the signalling questions for this domain, and not on the way that trial authors described the analysis.

6.1.2 Missing outcome data when estimating the effect of adhering to intervention

As noted above, the potential for bias arising from exclusion of eligible trial participants who did not adhere to their assigned intervention is addressed in the domain 'Bias due to deviations from intended interventions' (see section 5). Such analyses may be additionally biased if participants are excluded due to missing outcome data. Similarly, appropriate methods to estimate the effect of adhering to intervention (for example, instrumental variable analyses to estimate the effect of a single intervention that is administered at baseline) may be biased by missing outcome data.

The circumstances in which missing outcome data lead to bias are similar regardless of the effect of interest, so there is a single set of signalling questions for this domain.

6.1.3 When do missing outcome data lead to bias?

Statistical analyses excluding individuals with missing outcome data are examples of 'complete-case' analyses (analyses restricted to individuals in whom there were no missing values of included variables). To understand when missing outcome data lead to bias in such analyses, we need to consider:

(1) The **true value of the outcome** in participants with missing outcome data. This is the value of the outcome that should have been measured but was not;
(2) The **missingness mechanism**, which is the process that led to outcome data being missing.

Whether missing outcome data lead to bias depends on the relationship between the true value of the outcome and the missingness mechanism. Equivalently, we can consider whether the measured (non-missing) outcomes are systematically different from the missing outcomes (the true values in participants with missing outcome data). For example, consider a trial of cognitive behavioural therapy (CBT) compared with usual care for depression. If participants who are more depressed are less likely to return for follow-up, then whether depression is missing depends on its true value. Equivalently, the measured depression outcomes will differ systematically from the true values of the missing depression outcomes.

Below, we summarise situations in which missing outcome data do and do not lead to bias in the estimated intervention effect of intervention in a complete-case analysis:

(1) When missingness in the outcome is unrelated to its true value, within each intervention group, missing outcome data **will not lead to bias**. In this situation, the complete cases are representative of those with missing data. For example, missing outcome data would not lead to bias if the missingness occurred by chance, because of failure of an automatic measuring device.
(2) When missingness in the outcome depends on both the intervention group and the true value of the outcome, missing outcome data **will lead to bias** (in this situation data are often described as 'missing not at random' (MNAR)). For example, the results of a complete-case analysis will be biased in a placebo-controlled trial of an antidepressant drug for symptoms of depression, in which (1) participants with continuing symptoms of depression are more likely to be lost to follow up, and (2) side effects of the drug cause participants assigned to that drug to drop out of the study.
(3) When missingness in the outcome is related to its true value and, additionally, the effect of the experimental intervention differs from that of the comparator intervention, missing outcome data **will lead to bias except in the special case described below**. For example, in the trial of an antidepressant drug described above, its estimated effect on symptoms of depression will be biased if (1) participants with continuing symptoms of depression are more likely to be lost to follow up, and (2) the drug affects symptoms of depression, compared with placebo.
    - The special case is that when the outcome variable is dichotomous, and the intervention effect estimate is an odds ratio, missing outcome data **will not lead to bias** even if missingness depends on the true value of the outcome, providing that missingness is not also related to intervention group assignment. For example, the odds ratio from a trial comparing the effect of an experimental and comparator intervention on all-cause mortality will not be biased by missing outcome data, even if outcome data are more likely to be missing in participants who died, and mortality was less likely in the experimental than comparator intervention group. This is because the proportional reduction in the risk of death applies to both the intervention and control groups, so cancels out in the calculation of the odds ratio. This exception does not apply if the outcome is a dichotomous variable derived from a numerical variable (for example, a classification of hypertension based on measured blood pressure): odds ratios will still be biased in this situation.

      Because many dichotomous outcomes are derived from numerical variables, and because it is difficult to distinguish situations in which outcome data do and do not depend on intervention group assignment, the signalling questions for this domain do not distinguish odds ratios from other measures of intervention effect. Further we do not distinguish between situations in which the effect of the experimental intervention does and does not differ from that of the comparator intervention, because it is usually not possible to exclude a difference (absence of evidence is not evidence of absence).

(4) There are further exceptions, which are not listed because they are hard to identify in practice, and likely to be rare.

If the intervention effect differs between participant subgroups ('effect modification'), then the considerations above apply separately within the subgroups. If missingness in the outcome differs between the subgroups, this will change the overall intervention effect estimate, because the subgroup proportions analysed will differ from those originally randomised. This issue is not considered further, because differences in the distribution of effect modifiers between trials are usually considered to be a source of heterogeneity rather than bias.

6.1.3.1 Analyses that adjust for participant characteristics at baseline

Some trial analyses adjust for characteristics of participants at baseline, and so exclude participants in whom data on baseline characteristics are missing. Exclusions because of missing data on baseline characteristics only lead to bias if, additionally, missingness in the outcome depends on its true value (see above). Therefore, they are not considered separately. An exception is if baseline characteristics are collected retrospectively, and the outcome causes missingness in the baseline characteristics: this has the same implications for bias as when missingness in the outcome depends on its the true value, discussed above. For example, in a trial of a new treatment for acute stroke, where the outcome was 24-hour mortality, baseline characteristics such as duration of symptoms could not be collected in those who died before recovering consciousness.

Conversely, it might be possible to reduce or remove bias due to missing outcome data by accounting for participant characteristics at baseline. For example, suppose that the outcome variable is blood pressure, and that both older people and those in the experimental intervention group were more likely to drop out of the trial. This would lead to bias in the estimated effect of intervention on blood pressure, because older people tend to have higher blood pressure. The bias would be removed in an analysis adjusting for age at baseline, if this fully accounts for the relation between missingness in the outcome and its true value.

6.1.3.2 When is the amount of missing outcome data small enough to exclude bias?

Unfortunately, there is no sensible threshold for 'small enough' in relation to the proportion of missing outcome data.

In situations where missing outcome data lead to bias, the extent of bias will increase as the amount of missing outcome data increases. There is a tradition of regarding a proportion of less than 5% missing outcome data as "small" (with corresponding implications for risk of bias), and over 20% as "large". However, the potential impact of missing data on estimated intervention effects depends on the proportion of participants with missing data, the type of outcome and (for dichotomous outcome) the risk of the event. For example, consider a study of 1000 participants in the intervention group where the observed mortality is 2% for the 900 participants with outcome data (18 deaths). Even though the proportion of data missing is only 10%, if the mortality rate in the 100 missing participants is 20% (20 deaths), the overall true mortality of the intervention group would be nearly double (3.8% vs. 2%) that estimated from the observed data.

6.1.4 How can we identify evidence of bias due to missing outcome data?

It is not possible to examine directly whether the chance that the outcome is missing depends on its true value: judgements of risk of bias will depend on the circumstances of the trial. Therefore, we can only be sure that there is no bias due to missing outcome data when the outcome is measured in all participants, when the proportion of missing outcome data is sufficiently low that any bias is too small to be of importance, or when sensitivity analyses (conducted by either the trial authors or the review authors) confirm that plausible values of the missing outcome data could make no important difference to the estimated intervention effect.

However, indirect evidence that missing outcome data are likely to cause bias can come from examining: (1) differences between the proportion of missing outcome data in the experimental and comparator intervention groups and (2) reasons that outcome data are missing.

6.1.4.1 Differing proportions of missing outcome data

If the intervention has an effect, and missingness in the outcome depends on its true value, then the proportion of participants with missing data is likely to differ between the intervention groups. Therefore, differing proportions of missing outcome data in the experimental and comparator intervention groups provide evidence of potential bias.

It is possible that missing outcome data do not lead to bias, even when the proportion of missing outcome data differs between intervention groups. This will only be the case if the chance of the outcome being missing is not related to its true value: for example if engagement with the trial was greater in the experimental than

comparator intervention group but engagement, and hence missing outcome data, was not related to the outcome. We expect that it will be unusual for review authors to be confident that the chance that the outcome is missing is unrelated to its true value.

6.1.4.2 Examining reasons for missing outcome data

Trial reports may provide reasons why participants have missing data. For example, trials of haloperidol to treat dementia reported various reasons such as 'lack of efficacy', 'adverse experience', 'positive response', 'withdrawal of consent' and 'patient ran away', and 'patient sleeping' (77). It is likely that some of these (for example, 'lack of efficacy' and 'positive response') are related to the true values of the missing outcome data. Therefore, these reasons increase the risk of bias if the effects of the experimental and comparator interventions differ, or if the reasons are related to intervention group (for example, 'adverse experience').

In practice, our ability to assess risk of bias will be limited by the extent to which trial authors collected and reported reasons that outcome data were missing.

The situation most likely to lead to bias is when reasons for missing outcome data differ between the intervention groups: for example if participants who became seriously unwell withdrew from the comparator group while participants who recovered withdrew from the experimental intervention group.

6.1.5 Time-to-event data

Many trial analyses use methods for 'time-to-event' data, in which the outcome is a dichotomous variable that indicates whether the outcome event was observed in each participant. The follow up time ends either when the outcome event occurs or when observation stops for other reasons. Follow-up times for participants in whom the outcome event was not observed before observation stopped are said to be 'censored'. Intervention effects in analyses of time-to-event data are typically estimated as rate ratios or hazard ratios.

Results of time-to-event analyses will be unbiased only if censoring is 'non-informative', which means that censoring times for censored participants are unrelated to the (subsequent) times at which outcome events occur. For example, if all participants are followed until a specified date after which follow up ends, then censoring can be assumed to be non-informative. Informative censoring implies that the chance that the outcome is not observed depends on its true value, as discussed in detail above. Therefore, considerations of bias due to missing data are similar for time-to-event data. In the presence of informative censoring, a time-to-event analysis will be biased if:

- the chance that the follow up is censored also depends on the intervention group (for example, if censoring is more likely because participants in the experimental intervention group are lost to follow up because of severe side effects); and
- the effect of the experimental and comparator interventions on the outcome differs.

For similar reasons to those described above, either differences in rates of censoring or differing reasons for censoring may provide evidence that censoring was informative.

6.1.6 Statistical methods for handling missing outcome data

Trial authors may present statistical analyses (in addition to or instead of complete case analyses) that attempt to address the potential for bias caused by missing outcome data (76, 78-80). The most common approaches are:

(1) single imputation (i.e., generate a complete dataset by filling in the missing values of the outcome);
(2) multiple imputation (generate multiple complete datasets based on a predictive distribution for the outcome variable); and
(3) methods that do not require a complete data set such as likelihood-based methods, moment-based methods, and semiparametric models for survival data (76, 81).

Imputation approaches replace missing values by one or more new values. In single imputation, only one estimate is filled in. Commonly used approaches include 'last observation carried forward' (LOCF) and 'baseline observation carried forward' (BOCF). Each of these is unlikely to remove the bias that occurs when missingness in the outcome depends on its true value. Further, they generally improve precision artificially (so that the confidence interval for the intervention effect estimate is too narrow), since they do not reflect uncertainty about the missing outcomes. Therefore, intervention effect estimates based on these methods should be considered as at low risk of bias only when there is clear justification (77, 82).

In multiple imputation, multiple values of the missing outcomes are drawn at random from a predictive distribution, forming multiple distinct filled-in datasets (83, 84). These multiple datasets are analysed to produce a single summary estimate and confidence interval that reflect the uncertainty associated with missing data (unlike single imputation methods). However, multiple imputation methods will not remove or reduce the bias that occurs when missingness in the outcome depends on its true value, unless such missingness can be explained by measured variables. In particular, imputing missing outcome data based only on intervention group will give results that are near-identical to those from a complete-case analysis, and does not reduce or remove bias when outcome data are MNAR.

It may be possible to reduce bias associated with missing outcome data when values of the outcome are measured repeatedly over time and these measurements are used to predict the missing outcome data (85). Even when such an approach is used, review authors should consider carefully whether loss to follow up is plausibly related to the outcome trajectory after the last recorded measurement.

Bias because of loss to follow up can also be addressed by modelling its probability over time, and using the probability of loss to follow up to give more weight to participants who resemble those lost. The aim is to conduct a weighted analysis in which characteristics of participants after baseline are unrelated to loss to follow up: this is an example of a semiparametric approach to dealing with missing data (see approach (3) above). As with imputation-based approaches, such analyses will not remove bias if missingness in the outcome depends on its true value, even after accounting for the variables used to derive the weights.

6.1.7 Sensitivity analyses

Sensitivity analyses can be performed to assess the potential impact of missing outcome data, based on assumptions about the relationship between missingness in the outcome and its true value. They may be conducted by trial authors or by review authors. However, they are only helpful in judging risk of bias if they address the potential relationship between missingness in the outcome and its true value. The methods summarized in section 6.1.6 can be extended to conduct such sensitivity analyses (86, 87).

Several methods are available for review authors to assess the robustness of a result from a randomized trial in the presence of missing data (88). For dichotomous outcomes, Higgins and colleagues propose a strategy involving different assumptions about how the risk of the event among the missing participants differs from the risk of the event among the observed participants, taking account of uncertainty introduced by the assumptions (77). Akl and colleagues propose a suite of simple imputation methods, including a similar approach to that of Higgins and colleagues based on relative risks of the event in missing versus observed participants (89). Similar ideas can be applied to continuous outcome data (90-92). Particular care is required to avoid double counting events, since it can be unclear whether reported numbers of events in trial reports apply to the full randomized sample or only to those who did not drop out (93).

Although there is a tradition of implementing "worst case" and "best case" analyses clarifying the extreme boundaries of what is theoretically possible, such analyses may not be informative for the most plausible scenarios (77).

6.2 Empirical evidence of bias due to missing outcome data

Empirical research has investigated the adequacy with which missing data are addressed in reports of trials. A study that included 71 trial reports from four general medical journals, concluded that missing data are common and often inadequately handled in the statistical analysis (94).

Concerns over bias resulting from missing outcome data are driven mainly by theoretical considerations. Several empirical studies have looked at whether various aspects of missing data are associated with the magnitude of effect estimates (15, 31, 68, 73, 95-97). In a systematic review of meta-epidemiological studies, missing data were associated with overestimation of effect estimates in some studies, but underestimation or no difference in others(30). Many of the studies do not differentiate between missing outcome data and participants being excluded from the analysis (see also Section 5.2). In one study, having a dropout rate >20% (versus ≤20%) was not associated with different effect estimates on average (68).

6.3 Using this domain of the tool

(1) Risk of bias will be low if outcome data are available for all, or nearly all, randomized participants. The meaning of 'nearly all' in this context is that the number of participants with missing outcome data is so

small that their outcomes, whatever they were, could have made no important difference to the estimated effect of intervention. If this is the case then no further signalling questions need be answered. Absence of information about the extent of missing outcome data (for example, when no CONSORT flow diagram was provided in the trial report) will usually lead to a judgement of high risk of bias for this domain.

(2) Risk of bias will be low if sensitivity analyses, conducted by either the trial authors or the review authors (see Section 6.1.7), confirm that the finding is robust to plausible values of the missing outcome data. Such analyses are likely to be particularly useful when the amount of missing data is sufficiently large for the potential impact on the estimated effect of intervention to be substantial. If sensitivity analyses confirm that the result is robust, then it may be regarded as at low risk of bias and no further signalling questions need be answered.

(3) As explained in section 6.1.3, missing outcome data can only lead to bias if the chance that the outcome is missing depends on its true value. It may be possible to exclude this based on reported reasons for missing outcome data (for example, if outcome data are only missing because of failure of a measuring instrument) or closure of a centre in a multicentre trial. However, if it is possible that **missingness in the outcome could depend on its true value** then review authors will need to consider the proportions of and reasons for missing outcome data.

(4) A difference between the experimental and comparator intervention groups in the proportions of missing outcome data may indicate a risk of bias (see section 6.1.4). For time-to-event-data, review authors should consider whether rates of censoring (loss to follow-up) differ between the intervention groups.

(5) Either reasons for missing outcome data reported by trial authors, or the circumstances of the trial, may lead review authors to conclude that it is **likely that missingness in the outcome depended on its true value**.

(6) Differing reasons for missing outcome data in the experimental and comparator intervention groups may lead to substantial bias. For example, in a trial of an experimental intervention aimed at smoking cessation there would be serious bias if comparator intervention participants left the study due to a lack of enthusiasm at receiving nothing novel (and continued to smoke) while experimental intervention participants left the study due to successful cessation of smoking.

6.4 Signalling questions and criteria for judging risk of bias

Signalling questions for this domain are provided in Box 8. Criteria for reaching risk-of-bias judgements are given in Table 11, and an algorithm for implementing these is provided in Table 12 and Figure 4.

Box 8. The RoB 2 tool (part 5): Risk of bias due to missing outcome data

Signalling questions	Elaboration	Response options
3.1 Were data for this outcome available for all, or nearly all, participants randomized?	The appropriate study population for an analysis of the intention to treat effect is all randomized participants. Note that imputed data should be regarded as missing data, and not considered as "outcome data" in the context of this question. Only answer 'No information' if the trial report provides no information about the extent of missing outcome data. This situation will usually lead to a judgement that there is a high risk of bias due to missing outcome data. "Nearly all" should be interpreted as that the number of participants with missing outcome data is so small that their outcomes, whatever they were, could have made no important difference to the estimated effect of intervention. For continuous outcomes, availability of data from 95% (or possibly 90%) of the participants would often be sufficient. For dichotomous outcomes, the proportion required is directly linked to the risk of the event. If the observed number of events is much greater than the number of participants with missing outcome data, the bias would necessarily be small.	Y / PY / PN / N / NI
3.2 If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?	Evidence that the result was not biased by missing outcome data may come from: (1) analysis methods that correct for bias; or (2) sensitivity analyses showing that results are little changed under a range of plausible assumptions about the relationship between missingness in the outcome and its true value. However, imputing the outcome variable, either through methods such as 'last-observation-carried-forward' or via multiple imputation based only on intervention group, should not be assumed to correct for bias due to missing outcome data.	NA / Y / PY / PN / N
3.3 If N/PN to 3.2: Could missingness in the outcome depend on its true value?	If loss to follow up, or withdrawal from the study, could be related to participants' health status, then it is possible that missingness in the outcome was influenced by its true value. However, if all missing outcome data occurred for documented reasons that are unrelated to the outcome then the risk of bias due to missing outcome data will be low (for example, failure of a measuring device or interruptions to routine data collection).	NA / Y / PY / PN / N / NI
3.4 If Y/PY/NI to 3.3: Do the proportions of missing outcome data differ between intervention groups?	If there is a difference between the effects of the experimental and comparator interventions on the outcome, and the missingness in the outcome is influenced by its true value, then the proportions of missing outcome data are likely to differ between intervention groups. Therefore, such a difference may indicate a risk of bias due to missing outcome data. The answer to this question only influences risk of bias judgements via the answer to question 3.5. For time-to-event-data, the question should be interpreted as "Do rates of censoring (loss to follow-up) differ between the intervention groups?"	NA / Y / PY / PN / N / NI
3.5 If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?	This question distinguishes between situations in which (i) missingness in the outcome could depend on its true value (assessed as 'Some concerns') from those in which (ii) it is likely that missingness in the outcome depended on its true value (assessed as 'High'). Four reasons for answering 'Yes' are: 1. The most likely explanation for differences between intervention groups in the proportions of missing outcome data is that missingness in the outcome depends on its true value (see answer to 3.4 above); 2. Reported reasons for missing outcome data provide evidence that missingness in the outcome depends on its true value; 3. Reported reasons for missing outcome data differ between the intervention groups;	NA / Y / PY / PN / N / NI

	4. The circumstances of the trial make it likely that missingness in the outcome depends on its true value. For example, in trials of interventions to treat schizophrenia it is widely understood that continuing symptoms make drop out more likely.	
Risk-of-bias judgement	See Table 9, Table 10 and Figure 4.	Low / High / Some concerns
Optional: What is the predicted direction of bias due to missing outcome data?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

Table 9. Reaching risk-of-bias judgements for bias due to missing outcome data

Low risk of bias	(i) Outcome data were available for all, or nearly all, randomized participants OR (ii) There is evidence that the result was not biased by missing outcome data OR (iii) Missingness in the outcome could not depend on its true value
Some concerns	(i) Outcome data were <u>not</u> available for all, or nearly all, randomized participants AND (ii) There is <u>not</u> evidence that the result was not biased by missing outcome data AND (iii) Missingness in the outcome <u>could</u> depend on its true value AND (iv.1.1) The proportions of missing outcome data do not differ between intervention groups AND (iv.1.2)It is unlikely that missingness in the outcome depended on its true value AND (iv.1.3) Reasons for missingness <u>differ</u> between the intervention groups OR (iv.2.1) The proportions of missing outcome data do not differ between intervention groups AND (iv.2.2) It is <u>likely</u> that missingness in the outcome depended on its true value AND (iv.2.3) Reasons for missingness do not differ between the intervention groups OR (iv.3.1) The proportions of missing outcome data <u>differ</u> between intervention groups AND (iv.3.2) It is unlikely that missingness in the outcome depended on its true value AND (iv.3.3) Reasons for missingness do not differ between the intervention groups
High risk of bias	(i) Outcome data were <u>not</u> available for all, or nearly all, randomized participants

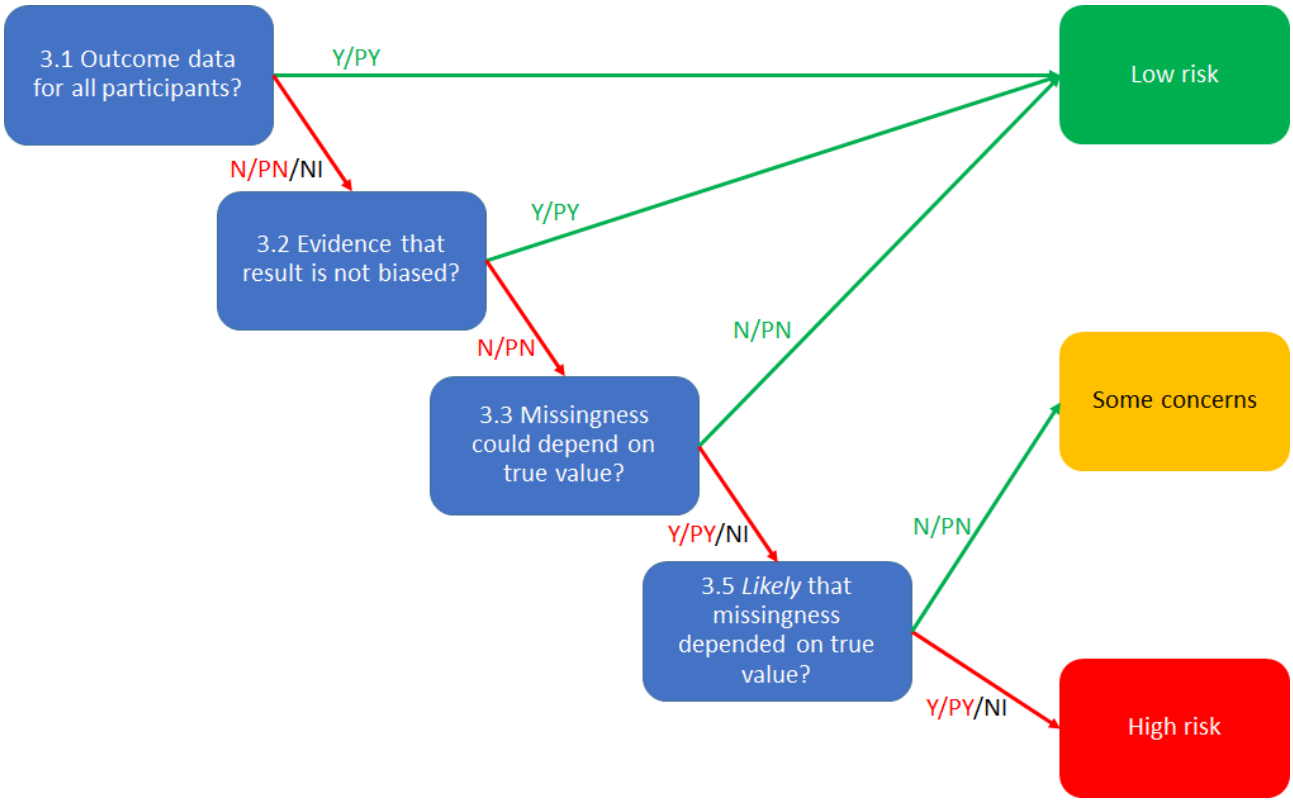
	AND (ii) There is not evidence that the result was not biased by missing outcome data AND (iii) Missingness in the outcome could depend on its true value AND (iv.1.1) The proportions of missing outcome data differ between intervention groups AND (iv.1.2) It is likely that missingness in the outcome depended on its true value OR (iv.2.1) The proportions of missing outcome data differ between intervention groups AND (iv.2.2) Reasons for missingness differ between the intervention groups OR (iv.3.1) It is likely that missingness in the outcome depended on its true value AND (iv.3.2) Reasons for missingness differ between the intervention groups

Table 10. Suggested mapping of signalling questions to risk-of-bias judgements for bias due to missing outcome data. This is only a suggested decision tree: all default judgements can be overridden by assessors.

Signalling question					Domain-level judgement
3.1	3.2	3.3	3.4	3.5	Default risk of bias
Complete data?	Evidence of no bias?	Depend on true?	Proportions differ?	Likely depend on true?	
Y/PY	NA	NA	NA	NA	Low
N/PN/NI	Y/PY	NA	NA	NA	Low
N/PN/NI	N/PN	N/PN	NA	NA	Low
N/PN/NI	N/PN	Y/PY/NI	Any response	N/PN	Some concerns
N/PN/NI	N/PN	Y/PY/NI	Any response	Y/PY/NI	High

Y/PY = 'Yes' or 'Probably yes'; N/PN = 'No' or 'Probably no'; NI = 'No information'; NA = "Not applicable"

Figure 4. Suggested algorithm for reaching risk-of-bias judgements for bias due to missing outcome data. This is only a suggested decision tree: all default judgements can be overridden by assessors.



7 Detailed guidance: bias in measurement of the outcome

7.1 Background

Errors in the measurement of outcomes can bias intervention effect estimates. These errors are often referred to as **measurement error** (for continuous outcomes), **misclassification** (for dichotomous or categorical outcomes) or **under-ascertainment/over-ascertainment** (for events). Measurement errors may be **differential** or **non-differential** in relation to intervention assignment.

- Differential measurement errors are systematically different between experimental and comparator intervention groups. Such errors are less likely when outcome assessors are blinded to intervention assignment.
- Non-differential measurement errors are unrelated to intervention assignment.

This domain relates primarily to assessment of the risk of differential errors in measuring the outcome variable. Box 9 explains why non-differential measurement errors are not addressed in detail.

Box 9. Implications of non-differential measurement error for intervention effect estimates

Errors in measurement or classification of an outcome that are unrelated to intervention assignment are known as 'non-differential' errors. They will generally not cause bias in intervention effect estimates of mean differences for continuous outcomes but are likely to attenuate (i.e. produce bias towards the null in) intervention effect estimates such as odds ratios, risk ratios or hazard ratios when the outcome is dichotomous or categorical. There are situations in which non-differential measurement error can bias effect estimates away from the null, but these are usually considered unlikely to occur in randomized trials.

The consequences of non-differential measurement error in an outcome depend on the intervention effect measure and the nature of the measurement error. For example:

- There will be no bias due to non-differential errors if the effect measure is a difference measure (such as a mean difference) and the measurement error acts additively on the outcome. For instance, if a blood pressure measuring device used in both intervention groups systematically produces a measurement that is 10 mmHg too high (an additive error), then this error will apply equally to mean values in the experimental and comparator intervention groups, and on computation of the difference between the means, the errors will 'cancel out'.
- There will be bias due to non-differential errors if the effect measure is a difference measure and both (1) the measurement error acts multiplicatively on the outcome and (2) the intervention effect is non-zero. For instance, if the blood pressure measuring device systematically produces a measurement that is 120% of the truth (a multiplicative error), then:
    - in the absence of an intervention effect the mean difference will be unbiased (taking the correct value of zero); but
    - in the presence of an intervention effect, the difference in means will be over-estimated by 20% (e.g. true means of 100 and 110 mmHg (difference 10 mmHg) would be measured as 120 and 132 mmHg (difference 12 mmHg));
    - if blood pressure measurements are used to classify patients as hypertensive or normotensive then estimates of risk ratios or odds ratios will be biased in the presence of an intervention effect.

  Note that errors in continuous measurements are typically assumed to be additive rather than multiplicative.
- The effects of non-differential measurement error are different for different effect measures (i.e. odds ratios, risk ratios and risk differences). As for continuous outcomes, the key consideration is whether the error acts in a manner that is 'congruent' with the effect measure. A thorough discussion is provided by Rothman and colleagues (98).

The first signalling question in this domain asks whether the method of measuring the outcome was inappropriate. Addressing non-differential measurement error in more detail would add substantial complexity. Furthermore, a meaningful assessment would require primary reports of trials to distinguish explicitly between the outcome and the way that it is measured, and to consider the relationship between the

true value and the measured value so that judgements can be made about whether errors are 'congruent' with the chosen intervention effect measures. For example, hypertension is usually diagnosed based on one or more blood pressure measurements, each of which is subject to measurement error. However, a trial result based on whether participants were diagnosed with hypertension at the end of follow is not usually considered inherently biased due to non-differential measurement error.

Consideration of risk of bias in this domain depends on:

(1) whether the method of measuring the outcome is appropriate;
(2) whether measurement or ascertainment of the outcome differs, or could differ, between intervention groups;
(3) who is the outcome assessor;
(4) whether the outcome assessor is blinded to intervention assignment; and
(5) whether the assessment of outcome is likely to be influenced by knowledge of intervention received.

(1) Outcomes in randomized trials should be assessed using appropriate outcome measures. For example, portable blood glucose machines used by trial participants may not reliably measure below 3.1 mmol, leading to an inability to detect differences in rates of severe hypoglycaemia between an insulin intervention and placebo, and under-representation of the true incidence of this adverse effect. Such a measurement would be inappropriate for this outcome.

(2) Outcomes should be measured or ascertained using a method that is comparable across intervention groups. This is usually the case for pre-specified outcomes. However, problems may arise with passive collection of outcome data, as is often the case for unexpected adverse effects. For example:

- In a placebo-controlled trial, severe headaches occur more frequently in participants assigned to a new drug than those assigned to placebo. These lead to more MRI scans being done in the experimental intervention group, and therefore to more diagnoses of symptomless brain tumours, even though the drug does not increase the incidence of brain tumours.
- Clemens et al identified the potential for what they called "diagnostic testing bias" in trials of the protective effect of BCG vaccine against tuberculosis (99). BCG vaccination usually leads to an easily identified scar, so that there is a potential for bias if tuberculosis is identified only using passive follow up, and either participants are less likely to seek care or assessors are less likely to order a radiograph if a scar is present. A systematic review found evidence that estimated protection was lower in trials assessed as at higher risk of such bias (100).

Such bias was described by Sackett: "an innocent exposure may become suspect if, rather than causing a disease, it causes a sign or symptom which precipitates a search for the disease" (101).

Even for a pre-specified outcome measure, the nature of the intervention may lead to methods of measuring the outcome that are not comparable across intervention groups. For example, an intervention involving additional visits to a healthcare provider may lead to additional opportunities for outcome events to be identified, compared with the comparator intervention.

(3) The outcome assessor can be:

- the **participant** when the outcome is a participant-reported outcome such as pain, quality of life, or self-completed questionnaire evaluating depression, anxiety or function;
- the **intervention provider** when the outcome is the result of a clinical examination, the occurrence of a clinical event or a therapeutic decision such as decision to offer a surgical intervention or decision to discharge the patient; or
- an **outcome assessor** who is an observer not directly involved in the intervention provided to the participant, such as an adjudication committee, a biologist performing an automatized complementary test, or a health professional recording outcomes for inclusion in health records or disease registries.

(4) Blinding of outcome assessors is often possible (and often done) even when blinding of participants and personnel during the trial is not feasible. However, it is particularly difficult for participant-reported outcomes: for example, in a trial comparing surgery with usual care on pain at 3 months, it is impossible to blind the outcome assessor (i.e. the participant). Impossibility of blinding outcome assessors does not mean that the resulting potential for bias can be ignored: review authors must always assess the risk of bias due to error in measuring the outcome.

(5) For studies in which outcome assessors were not blinded, whether the assessment of outcome is likely to be influenced by knowledge of the intervention received will depend on whether the assessment of outcome involves judgement. In turn, this depends on the type of outcome. We distinguish five different type of outcomes as follows.

1 Participant-reported outcomes

Participant-reported outcomes are any reports coming directly from participants about how they function or feel in relation to a health condition and its therapy, without interpretation of the participant's responses by a clinician, or anyone else. Participant-reported outcomes include any outcome evaluation obtained directly from participants through interviews, self-completed questionnaires, diaries or other data collection tools such as hand-held devices and web-based forms (102). Examples include pain, nausea and health-related quality of life.

The outcome assessor here is **the participant,** even if a blinded interviewer is questioning the participant and completing a questionnaire on their behalf. The interviewer is not considered to be the outcome assessor in a strict sense but rather a facilitator of the measurement.

For participant-reported outcomes, the assessment of outcome is **potentially influenced** by knowledge of intervention received, leading to a judgement of at least 'Some concerns'. Review authors will need to judge whether it is likely that participants' reporting of the outcome was influenced by knowledge of intervention received, in which case risk of bias is considered to be high. For example, a severe or unexpected adverse effect recorded some time after the start of the intervention may be considered unlikely to be influenced by knowledge of the intervention received. On the other hand, level of pain reported at the end of a course of acupuncture, in a study comparing acupuncture with no treatment, is likely to be affected by knowledge of the intervention received.

2 Observer-reported outcomes not involving judgement

These are outcomes reported by an external observer (e.g. an intervention provider, independent researcher, or physician not involved in the care provided to participants such as a radiologist) that do not involve any judgement from the observer. Examples include all-cause mortality or the result of an automated test.

The outcome assessor here is **the observer**. For observer-reported outcomes not involving judgement the assessment of outcome is usually **not likely to be influenced** by knowledge of intervention received.

3 Observer-reported outcomes involving some judgment

These are outcomes reported by an external observer (e.g. an intervention provider) that involve some judgement, such as is involved in a clinical examination. Examples include tests involving assessment of a radiograph, clinical examination and clinical events other than death (e.g. myocardial infarction) that require judgements based on medical records.

The outcome assessor here is **the observer**. If the observer is aware of the intervention received then assessment of the outcome is **potentially influenced** by this knowledge, leading to a judgement of at least 'Some concerns'. Review authors will need to judge whether it is likely that assessment of the outcome was influenced by knowledge of intervention received, in which case risk of bias is considered to be high.

4 Outcomes that reflect decisions made by the intervention provider

These are outcomes that reflect a decision made by the intervention provider. The recording of this decision does not involve any judgement. However, the decision itself can be influenced by knowledge of intervention received. For example, in a trial comparing the impact of laparoscopic versus small-incision cholecystectomy on hospital stay, it was essential to keep the carers blinded to the intervention received to make sure their decision to discharge participants was influenced only by the clinical evaluation of the participants. In general, examples of intervention provider decision outcomes include hospitalization, stopping treatment, referral to a different ward, performing a caesarean section, stopping ventilation and discharge of the participant.

The outcome assessor here is **the care provider making the decision**. The assessment of outcome is usually **likely to be influenced** by knowledge of intervention received, if the care provider is aware of

this. This is particularly important when preferences, expectations or hunches regarding the effect of the experimental intervention are strong.

5 Composite outcomes

A composite outcome combines multiple end points into a single outcome. Typically, participants who have experienced any of a specified set of endpoints are considered to have experienced the composite outcome. Examples include major adverse cardiac and cerebrovascular events (MACCE). Composite endpoints can also be constructed from continuous outcome measures.

Assessment of risk of bias for composite outcomes should take into account the frequency or contribution of each component of the composite outcome and take into account the risk of bias due to the most influential components.

7.2 Empirical evidence of bias in measurement of the outcome

Empirical evidence of bias in measurement of the outcome largely comes from studies exploring whether reporting of "double blinding" is associated with intervention effects. These studies are summarized in section 5.2. In addition, lack of blinding of outcome assessors in randomized trials has been shown to be associated with more exaggerated estimated intervention effects, by 34% on average, measured as odds ratio (103).

7.3 Using this domain of the tool

The first question in this domain is a screening question to identify rare instances when the method of measuring the outcome was inappropriate. This is unlikely to be the case for pre-specified outcomes in randomized trials, but may be the case for some adverse effects, particularly if there are no systems in place to define the adverse outcome or to identify its occurrence. This question **does not aim to assess whether the choice of outcome being evaluated was sensible** (e.g. because it is an appropriate surrogate or proxy for the main outcome of interest).

The second question addresses whether measurement or ascertainment of the outcome measurement could have differed between the intervention groups. Methods are likely to be comparable in most randomized trials. The question aims to identify unusual situations in which there were systematic differences between the groups, for example because the experimental intervention involved more encounters with healthcare professionals and so more opportunities to identify the outcome than in the control group, or because of 'diagnostic detection bias' in the context of passively collected outcomes (see section 7.1).

The subsequent questions address blinding and its potential implications for assessments of outcomes. It is often clear whether outcome assessments were made blinded to intervention assignment. If blinding was successfully implemented, then the risk of bias due to differential measurement error is low.

The importance of lack of blinding of the outcome assessor will depend on the extent to which the assessment can be influenced by knowledge of the intervention assignment. For objectively assessed outcomes such as all-cause mortality, assessment of the outcome is unlikely to have been influenced but for highly subjective outcomes such as 'clinical impression of improvement', knowledge of the intervention received is much more important.

When the outcome assessor is not blinded and the outcome **could have** been influenced by knowledge of intervention received, review authors should assess whether **it is likely that** such influence occurred. In doing so, they should account for both conscious and subconscious influences on the outcome assessor (both 'interests' and 'expectations') These could vary according to:

- the comparator (higher risk of bias if the comparator is no treatment or usual care than when the comparator is another active intervention),
- involvement of the outcome assessor in participants' care (lower risk of bias if the outcome assessor is an independent researcher).
- influence of other actors. For example, for participant-reported outcomes recorded through interview, risk of bias might be lower if the person interviewing the participant is an independent researcher compared with when the interviewer is the care provider and involved in the administration of the intervention (such as surgeons interviewing patients).

Timing of outcome assessment could also influence the likelihood that assessment of the outcome was influenced by knowledge of intervention received. For example, in a trial comparing supervised exercise with shockwave treatment (104), if the assessment is at 12 months, knowledge of intervention received – which ceased at 12 weeks post-randomization – might not matter very much. This may be the case if many other treatments have been used by participants in the interim, so that participants have largely forgotten what they received many months ago.

7.4 Signalling questions and criteria for judging risk of bias

Signalling questions for this domain are provided in Box 10. Criteria for reaching risk-of-bias judgements are given in Table 11 and an algorithm for implementing these is provided in Table 12 and Figure 5.

Box 10. The RoB 2 tool (part 6): Risk of bias in measurement of the outcome

Signalling questions	Elaboration	Response options
4.1 Was the method of measuring the outcome inappropriate?	This question aims to identify methods of outcome measurement (data collection) that are unsuitable for the outcome they are intended to evaluate. The question *does not* aim to assess whether the choice of outcome being evaluated was sensible (e.g. because it is a surrogate or proxy for the main outcome of interest). In most circumstances, for pre-specified outcomes, the answer to this question will be 'No' or 'Probably no'. Answer 'Yes' or 'Probably yes' if the method of measuring the outcome is inappropriate, for example because: (1) it is unlikely to be sensitive to plausible intervention effects (e.g. important ranges of outcome values fall outside levels that are detectable using the measurement method); or (2) the measurement instrument has been demonstrated to have poor validity.	Y / PY / PN / N / NI
4.2 Could measurement or ascertainment of the outcome have differed between intervention groups?	Comparable methods of outcome measurement (data collection) involve the same measurement methods and thresholds, used at comparable time points. Differences between intervention groups may arise because of 'diagnostic detection bias' in the context of passive collection of outcome data, or if an intervention involves additional visits to a healthcare provider, leading to additional opportunities for outcome events to be identified.	Y / PY / PN / N / NI
4.3 If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?	Answer 'No' if outcome assessors were blinded to intervention status. For participant-reported outcomes, the outcome assessor is the study participant.	Y / PY / PN / N / NI
4.4 If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?	Knowledge of the assigned intervention could influence participant-reported outcomes (such as level of pain), observer-reported outcomes involving some judgement, and intervention provider decision outcomes. They are unlikely to influence observer-reported outcomes that do not involve judgement, for example all-cause mortality.	NA / Y / PY / PN / N / NI
4.5 If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?	This question distinguishes between situations in which (i) knowledge of intervention status could have influenced outcome assessment but there is no reason to believe that it did (assessed as 'Some concerns') from those in which (ii) knowledge of intervention status was likely to influence outcome assessment (assessed as 'High'). When there are strong levels of belief in either beneficial or harmful effects of the intervention, it is more likely that the outcome was influenced by knowledge of the intervention received. Examples may include patient-reported symptoms in trials of homeopathy, or assessments of recovery of function by a physiotherapist.	NA / Y / PY / PN / N / NI
Risk-of-bias judgement	See Table 11, Table 12 and Figure 5.	Low / High / Some concerns

Optional: What is the predicted direction of bias in measurement of the outcome?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

Table 11. Reaching risk-of-bias judgements for bias in measurement of the outcome

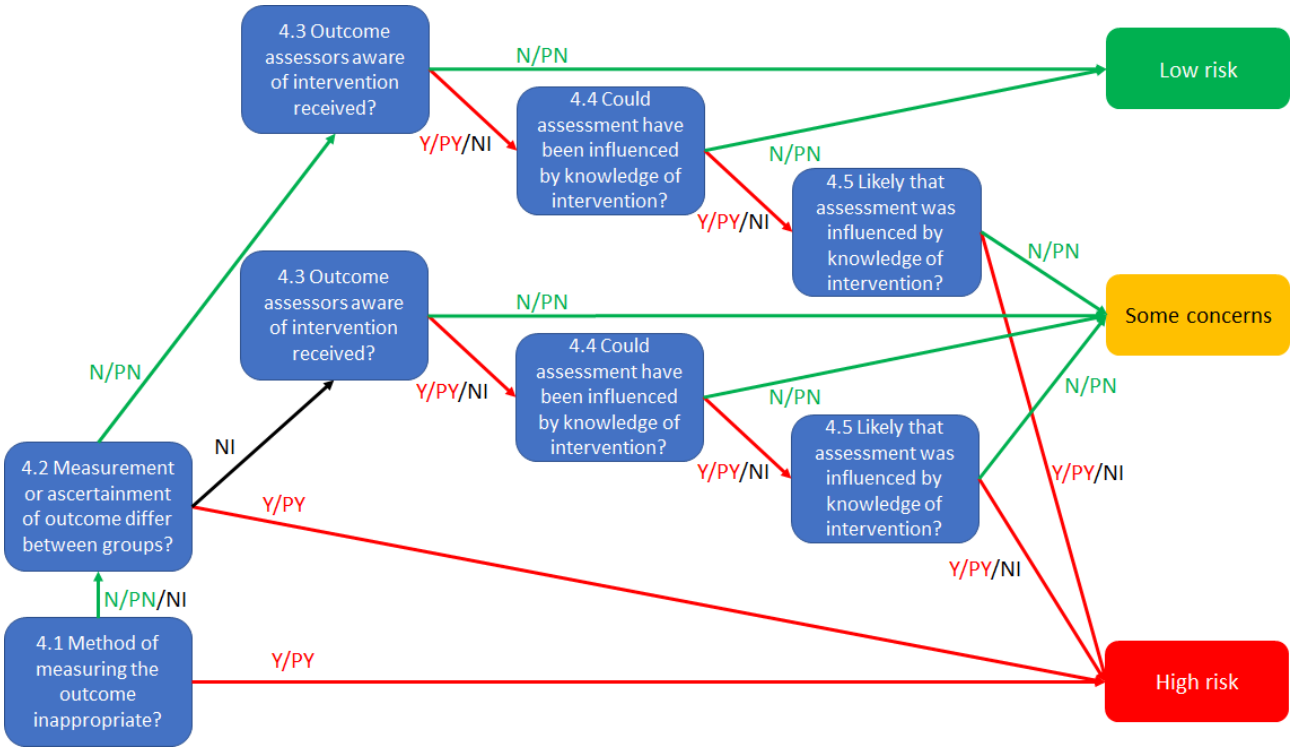
Low risk of bias	(i) The method of measuring the outcome was not inappropriate AND (ii) The measurement or ascertainment of the outcome did not differ between intervention groups AND (iii.1) The outcome assessors were unaware of the intervention received by study participants OR (iii.2) The assessment of the outcome could not have been influenced by knowledge of the intervention received
Some concerns	(i.1) The method of measuring the outcome was not inappropriate AND (i.2) The measurement or ascertainment of the outcome did not differ between intervention groups AND (i.3) The assessment of the outcome could have been influenced by knowledge of the intervention received AND (i.4) It is unlikely that assessment of the outcome was influenced by knowledge of intervention received OR (ii.1) The method of measuring the outcome was not inappropriate AND (ii.2) There is no information on whether the measurement or ascertainment of the outcome could have differed between intervention groups AND (ii.3.1) The outcome assessors were unaware of the intervention received by study participants OR (ii.3.2) The assessment of the outcome could not have been influenced by knowledge of the intervention received
High risk of bias	(i) The method of measuring the outcome was inappropriate OR (ii) The measurement or ascertainment of the outcome could have differed between intervention groups OR (iii) It is likely that assessment of the outcome was influenced by knowledge of the intervention received

Table 12. Suggested mapping of signalling questions to risk-of-bias judgements for bias in measurement of the outcome. This is only a suggested decision tree: all default judgements can be overridden by assessors.

Signalling question					Domain level judgement
4.1	4.2	4.3	4.4	4.5	Default risk of bias
Inappropriate?	Differed between groups?	Aware?	Could be influenced?	Likely to be influenced?	
N/PN/NI	N/PN	N/PN	NA	NA	Low
N/PN/NI	N/PN	Y/PY/NI	N/PN	NA	Low
N/PN/NI	N/PN	Y/PY/NI	Y/PY/NI	N/PN	Some concerns
N/PN/NI	N/PN	Y/PY/NI	Y/PY/NI	Y/PY/NI	High risk
N/PN/NI	NI	N/PN	NA	NA	Some concerns
N/PN/NI	NI	Y/PY/NI	N/PN	NA	Some concerns
N/PN/NI	NI	Y/PY/NI	Y/PY/NI	N/PN	Some concerns
N/PN/NI	NI	Y/PY/NI	Y/PY/NI	Y/PY/NI	High risk
Y/PY	Any response	Any response	Any response	Any response	High risk
Any response	Y/PY	Any response	Any response	Any response	High risk

Y/PY = 'Yes' or 'Probably yes'; N/PN = 'No' or 'Probably no'; NI = 'No information'; NA = "Not applicable"

Figure 5. Suggested algorithm for reaching risk-of-bias judgements for bias in measurement of the outcome. This is only a suggested decision tree: all default judgements can be overridden by assessors.



8 Detailed guidance: bias in selection of the reported result

8.1 Background

This domain addresses bias that arises because the reported result is selected (based on its direction, magnitude or statistical significance) from among multiple intervention effect estimates that were calculated by the trial authors. We call this **bias in selection of the reported result**. Consideration of risk of bias requires distinction between:

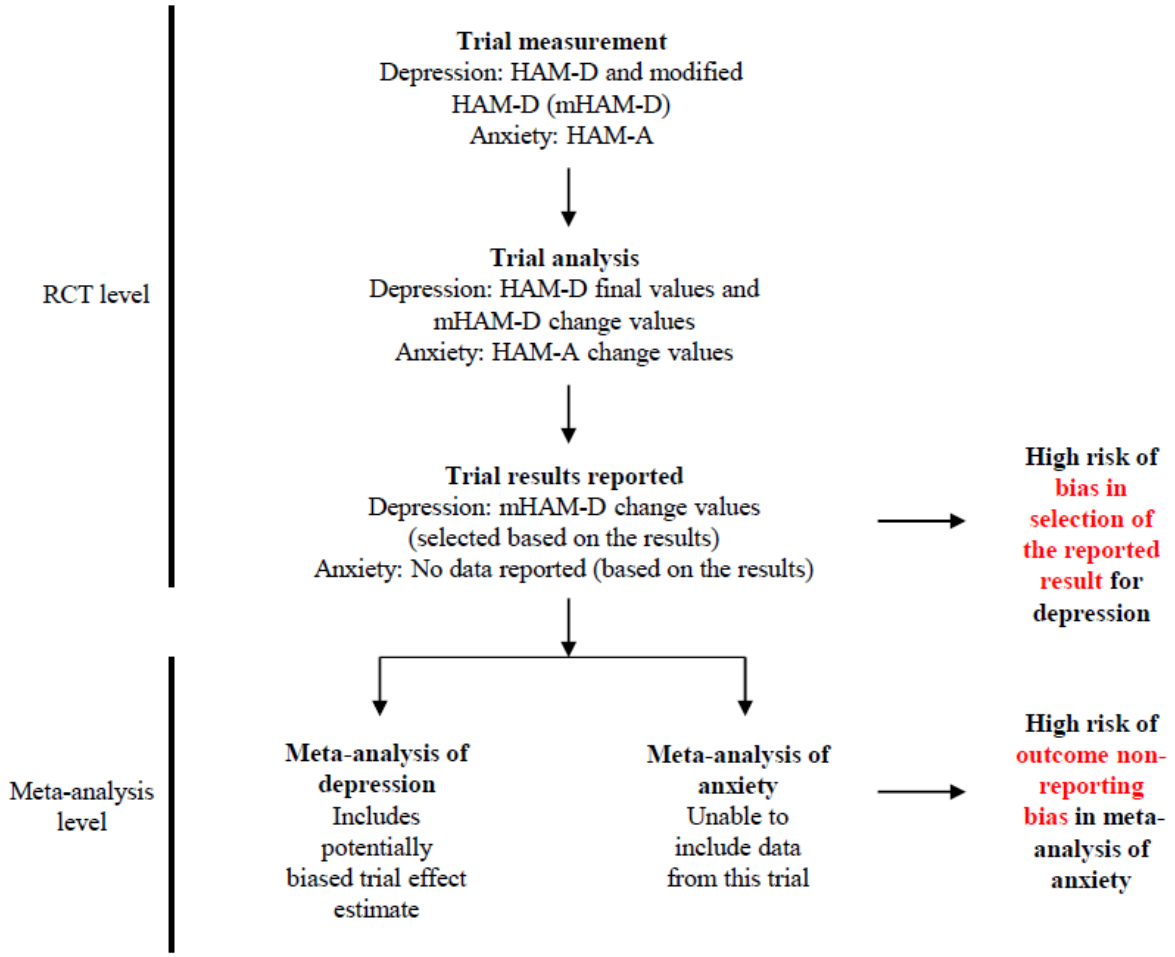
- An **outcome domain.** This is a state or endpoint of interest, irrespective of how it is measured (e.g. presence or severity of depression);
- A specific **outcome measurement** (e.g. measurement of depression using the Hamilton rating scale 6 weeks after starting intervention); and
- An **outcome analysis**. This is a specific result obtained by analysing one or more outcome measurements (e.g. the difference in mean change in Hamilton rating scale scores from baseline to 6 weeks between experimental and comparator groups).

This domain does not address bias due to selective non-reporting (or incomplete reporting) of outcome domains that were measured and analysed by the trial authors (105). For example, deaths of trial participants may be recorded by the trialists, but the reports of the trial might contain no data for deaths, or state only that the effect estimate for mortality was not statistically significant. Such bias puts the result of a synthesis at risk because results are omitted based on their direction, magnitude or statistical significance. It should therefore be addressed at the review level, as part of an integrated assessment of the risk of reporting bias (106).

An example of bias in selection of the reported result, and how this differs from outcome non-reporting bias, is shown in Figure 6. In this example, depression is measured using two scales (HAM-D and a modification of the HAM-D) and analysed using both final values and change from baseline values. The effect estimate for depression (modified HAM-D change values) was selected by the trialists because it was most favourable to the experimental intervention. The available result from the trial for depression is therefore at risk of bias due to selective reporting, and this bias would be addressed by the RoB 2 tool. In contrast, anxiety was measured and analysed by the trialists, but no data were reported because the effect estimate was not favourable to the experimental intervention. Such non-reporting will cause bias in a synthesis of results for anxiety but does not, by itself, put fully reported effect estimates from the same trial at risk of bias.

The separation of selection in reporting a result for an outcome from selective non-reporting of an outcome domain **is a notable change from version 1 of the Cochrane RoB tool for randomized trials**.

Figure 6. Examples of bias in selection of the reported result and outcome non-reporting bias



8.1.1 The role of core outcome sets

Recognition of the serious implications of selective non-reporting of outcomes by trial authors, and substantial variability between trials in the same clinical area in the choice of outcomes measured, led to initiatives to develop **core outcome sets** (107). These are recommended lists of a small number of essential outcome domains that should be measured in all trials in a specified clinical setting. They are usually derived through a formal consensus process by eliciting the perceived importance of various outcomes from clinicians and patients (108).

Widespread adoption of core outcome sets should reduce the occurrence of selective non-reporting of outcomes. Their impact on selection of the reported result is, however, likely to be more limited, since they have not consistently included recommendations on how the core outcomes should be defined and measured. Such work is ongoing (109). Development of fully defined and validated sets of core **outcome measures** will provide an important step in reducing bias in the selection of the reported result.

8.1.2 Selective reporting of a result contributing to the synthesis

This domain considers both (i) **selective reporting of a particular outcome measurement** from multiple measurements assessed within an outcome domain; and (ii) **selective reporting of a particular analysis** from multiple analyses of a specific outcome measurement. Either type of selective reporting will lead to bias if selection is based on the direction, magnitude or statistical significance of the effect estimate.

Selective reporting of a particular outcome measurement occurs when the reported effect estimate was selected (based on the results) from among effect estimates for multiple outcome measurements for an outcome domain. Examples include:

- reporting only one or a subset of time points at which the outcome was measured;

- use of multiple measurement instruments (e.g. pain scales) and only reporting data for the instrument with the most favourable result;
- having multiple assessors measure an outcome domain (e.g. clinician-rated and patient-rated depression scales) and only reporting data for the measure with the most favourable result; and
- reporting only the most favourable subscale (or a subset of subscales) for an instrument when measurements for other subscales were available.

Selective reporting of a particular analysis occurs when results are selected (based on the results) from intervention effects estimated in multiple ways. For example:

- carrying out analyses of both change scores and post-intervention scores adjusted for baseline;
- multiple analyses of a particular outcome measurement with and without adjustment for potential confounders (or with adjustment for different sets of potential confounders);
- a continuously scaled outcome converted to categorical data on the basis of multiple cut-points; and
- effect estimates generated for multiple composite outcomes with full reporting of just one or a subset.

Bias in selection of the reported result typically arises from a desire for findings to be sufficiently noteworthy to merit publication, and this could be the case if previous evidence (or a prior hypothesis) is either supported or contradicted. Bias of this kind can arise for both harms and benefits, although the motivations (and direction of bias) underlying selection of effect estimates for harms and benefits may differ. For example, in trials comparing an experimental intervention with placebo, trialists who have a preconception or vested interest in showing that the experimental intervention is beneficial and safe may be inclined to be selective in reporting efficacy estimates that are statistically significant and favourable to the experimental intervention, along with harm estimates that are not significantly different between groups. In contrast, other trialists may selectively report harm estimates that are statistically significant and unfavourable to the experimental intervention if they believe that publicizing the existence of a harm will increase their chances of publishing in a high impact journal. Such motivations are not always easy to decipher; for example, in trials comparing active interventions, different trialists may have different preconceptions about the efficacy and safety of the different interventions.

8.2 Empirical evidence of bias in selection of the reported result

In a systematic review comparing different sections within trial reports, and comparing trial reports with their protocols, discrepancies were often found in definitions of composite outcomes, handling of missing data, unadjusted versus adjusted analyses, and subgroup analyses (110). Such discrepancies are suggestive of bias in selection of the reporting result.

8.3 Using this domain of the tool

8.3.1 Importance of seeking the analysis intentions of a trial

We strongly encourage review authors to attempt to retrieve the pre-specified analysis intentions for each trial. Doing so allows for the identification of any outcome measures or analyses that have been omitted from, or added to, the results report, post hoc.

Analysis intentions may be documented in a variety of sources, including the trial registry entry (e.g. ClinicalTrials.gov record), trial protocol or design paper (which may be published in a journal or available via the funder's website). The statistical analysis plan (SAP) often provides the most details, but may not be published. If the researchers' pre-specified intentions are available in sufficient detail, then planned outcome measurements and analyses can be compared with those presented in the published report(s).

When comparing analysis intentions with the publication(s), the dates of such documents must be considered carefully. There should be a "date-stamp" confirming that the analysis intentions were finalized before unblinded outcome data were available to the trial authors (other than staff who provided confidential data to a monitoring committee). Amendments or updates to analysis intentions should also be retrieved and compared with the original intentions. These are usually date-stamped in trial registries or journal publications, so review authors can determine when changes were made.

Review authors should ideally ask the study authors to supply the study protocol and full statistical analysis plan if these are not publicly available. In addition, if outcome measures and analyses mentioned in an article or protocol are not reported, study authors could be asked to clarify whether those outcome measures were in fact analysed and, if so, to supply the data.

Trial protocols should describe how unexpected adverse outcomes (that potentially reflect unanticipated harms) will be collected and analysed. However, results based on spontaneously-reported adverse outcomes may lead to concerns that these were selected based on the finding being noteworthy.

8.3.2 Inferring selective reporting when analysis intentions are unavailable

For some trials, the analysis intentions will not be readily available. It is still possible to assess the risk of bias in selection of the reported result. For example, outcome measures and analyses listed in the methods section of an article can be compared with those that are reported. Furthermore, outcome measures and analyses can be compared across different papers describing trial. In addition, the following questions may help review authors to infer selective reporting:

(1) Are subscales aggregated in an unusual manner?

(2) Is there a discrepancy between different reports in the designation of the primary and secondary outcomes or specific outcomes?

(3) Is there any suggestion that multiple adjusted analyses were carried out but only one (or a subset) was reported? Were one or more adjusted analyses performed but none reported?

(4) Have the researchers categorized continuous outcome measures in an unusual way? Are different cut-points for creating categories reported across multiple publications relating to the same study?

(5) Is there is a discrepancy between different reports in the sample of participants analysed?

(6) Has an unusual composite outcome been reported? Could different definitions of a composite outcome have been considered, for example by grouping different combinations of unanticipated adverse events under a category of "major adverse event" or "minor adverse event"?

It is important to recognize that some differences between analysis intentions and publication may be due to legitimate changes to the protocol. For example, planned subgroup analyses or planned cut points for continuous outcome measures may need to be modified because the distribution of data differed from what was anticipated, resulting in subgroups with no data or very uneven spread. Although such changes should be reported in publications, few trialists do so (110). Further, trialists may amend their analysis intentions before conducting any analyses, yet not update the publicly available trial registry record or protocol. Therefore, contact with authors to seek clarification for any discrepancies identified will be necessary.

Insufficient detail in some documents may preclude a full assessment of the risk of bias in selection of the reported result (e.g. trialists only state in the trial registry record that they will measure "pain", without specifying the measurement scale, time point or metric that will be used). Review authors should indicate insufficient information alongside their responses to signalling questions.

8.4 Signalling questions and criteria for judging risk of bias

Signalling questions for this domain are provided in Box 11. Criteria for reaching risk-of-bias judgements are given in Table 13, and an algorithm for implementing these is provided in Table 14 and

Figure 7.

Box 11. The RoB 2 tool (part 7): Risk of bias in selection of the reported result

Signalling questions	Elaboration	Response options
5.1 Was the trial analysed in accordance with a pre-specified plan that was finalized before unblinded outcome data were available for analysis?	If the researchers' pre-specified intentions are available in sufficient detail, then planned outcome measurements and analyses can be compared with those presented in the published report(s). To avoid the possibility of selection of the reported result, finalization of the analysis intentions must precede availability of unblinded outcome data to the trial authors. Changes to analysis plans that were made before unblinded outcome data were available, or that were clearly unrelated to the results (e.g. due to a broken machine making data collection impossible) do not raise concerns about bias in selection of the reported result.	Y / PY / PN / N / NI
Is the numerical result being assessed likely to have been selected, on the basis of the results, from...		
5.2. ... multiple outcome measurements (e.g. scales, definitions, time points) within the outcome domain?	A particular outcome domain (i.e. a true state or endpoint of interest) may be **measured** in multiple ways. For example, the domain pain may be measured using multiple scales (e.g. a visual analogue scale and the McGill Pain Questionnaire), each at multiple time points (e.g. 3, 6 and 12 weeks post-treatment). If multiple measurements were made, but only one or a subset is reported on the basis of the results (e.g. statistical significance), there is a high risk of bias in the fully reported result. Answer 'Yes' or 'Probably yes' if: There is clear evidence (usually through examination of a trial protocol or statistical analysis plan) that a domain was measured in multiple ways, but data for only one or a subset of measures is fully reported (without justification), and the fully reported result is likely to have been selected on the basis of the results. Selection on the basis of the results can arise from a desire for findings to be newsworthy, sufficiently noteworthy to merit publication, or to confirm a prior hypothesis. For example, trialists who have a preconception, or vested interest in showing, that an experimental intervention is beneficial may be inclined to selectively report results that are favourable to the experimental intervention. Answer 'No' or 'Probably no' if: There is clear evidence (usually through examination of a trial protocol or statistical analysis plan) that all reported results for the outcome domain correspond to all intended outcome measurements. or There is only one possible way in which the outcome domain can be measured (hence there is no opportunity to select from multiple measures). or	Y / PY / PN / N / NI

	Outcome measurements are inconsistent across different reports on the same trial, but the trialists have provided the reason for the inconsistency and it is not related to the nature of the results. Answer 'No information' if: Analysis intentions are not available, or the analysis intentions are not reported in sufficient detail to enable an assessment, and there is more than one way in which the outcome domain could have been measured.	
5.3 ... multiple analyses of the data?	A particular outcome domain may be **analysed** in multiple ways. Examples include: unadjusted and adjusted models; final value *vs* change from baseline *vs* analysis of covariance; transformations of variables; different definitions of composite outcomes (e.g. 'major adverse event'); conversion of continuously scaled outcome to categorical data with different cut-points; different sets of covariates for adjustment; and different strategies for dealing with missing data. Application of multiple methods generates multiple effect estimates for a specific outcome domain. If multiple estimates are generated but only one or a subset is reported on the basis of the results (e.g. statistical significance), there is a high risk of bias in the fully reported result. Answer 'Yes' or 'Probably yes' if: There is clear evidence (usually through examination of a trial protocol or statistical analysis plan) that a domain was analysed in multiple ways, but data for only one or a subset of analyses is fully reported (without justification), and the fully reported result is likely to have been selected on the basis of the results. Selection on the basis of the results arises from a desire for findings to be newsworthy, sufficiently noteworthy to merit publication, or to confirm a prior hypothesis. For example, trialists who have a preconception or vested interest in showing that an experimental intervention is beneficial may be inclined to selectively report analyses that are favourable to the experimental intervention. Answer 'No' or 'Probably no' if: There is clear evidence (usually through examination of a trial protocol or statistical analysis plan) that all reported results for the outcome domain correspond to all intended analyses. or There is only one possible way in which the outcome domain can be analysed (hence there is no opportunity to select from multiple analyses). or Analyses are inconsistent across different reports on the same trial, but the trialists have provided the reason for the inconsistency and it is not related to the nature of the results. Answer 'No information' if:	Y / PY / PN / N / NI

	Analysis intentions are not available, or the analysis intentions are not reported in sufficient detail to enable an assessment, and there is more than one way in which the outcome domain could have been analysed.	
Risk-of-bias judgement	See Table 13, Table 14 and Figure 7.	Low / High / Some concerns
Optional: What is the predicted direction of bias due to selection of the reported result?	If the likely direction of bias can be predicted, it is helpful to state this. The direction might be characterized either as being towards (or away from) the null, or as being in favour of one of the interventions.	Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

Table 13. Reaching risk-of-bias judgements for bias in selection of the reported result

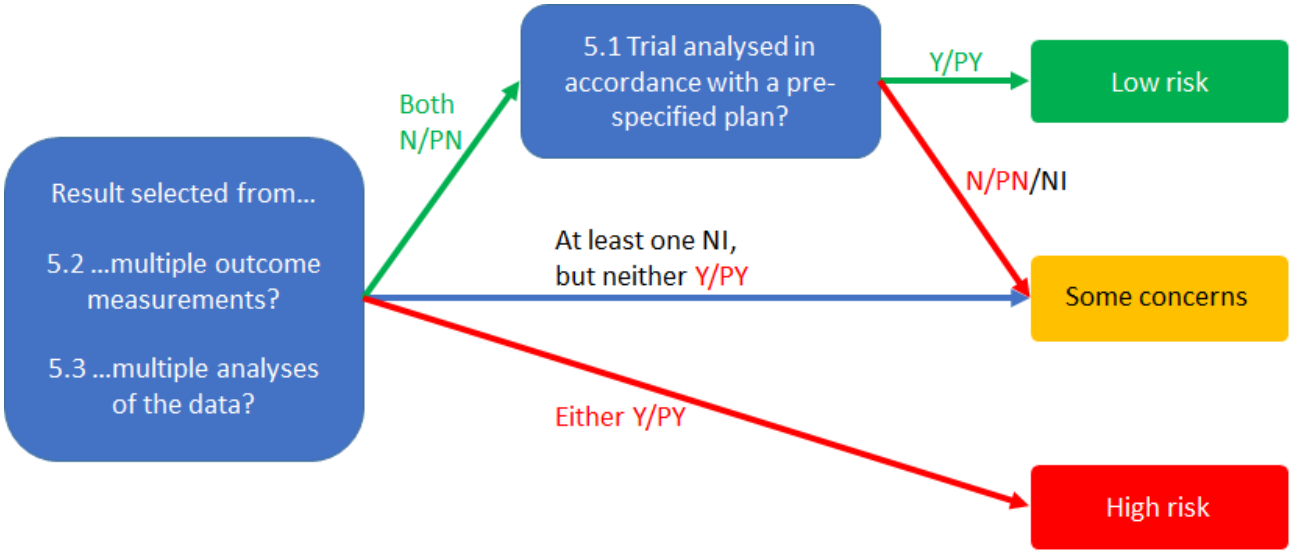
Low risk of bias	(i) The trial was analysed in accordance with a pre-specified plan that was published before outcome data were available for analysis AND (ii) Reported outcome data are unlikely to have been selected, on the basis of the results, from multiple outcome measurements (e.g. scales, definitions, time points) within the outcome domain AND (iii) Reported outcome data are unlikely to have been selected, on the basis of the results, from multiple analyses of the data
Some concerns	(i.1) The trial was not analysed in accordance with a pre-specified plan that was published before outcome data were available for analysis AND (i.2) Reported outcome data are unlikely to have been selected, on the basis of the results, from multiple outcome measurements (e.g. scales, definitions, time points) within the outcome domain AND (i.3) Reported outcome data are unlikely to have been selected, on the basis of the results, from multiple analyses of the data OR (ii)There is no information on whether reported outcome data are likely to have been selected, on the basis of the results, from multiple outcome measurements (e.g. scales, definitions, time points) within the outcome domain AND from multiple analyses of the data
High risk of bias	(i) Reported outcome data are likely to have been selected, on the basis of the results, from multiple outcome measurements (e.g. scales, definitions, time points) within the outcome domain OR (ii) Reported outcome data are likely to have been selected, on the basis of the results, from multiple analyses of the data

Table 14. Suggested mapping of signalling questions to risk-of-bias judgements for bias in selection of the reported result.

Signalling question			Domain level judgement
5.1	**5.2**	**5.3**	**Default risk of bias**
In accordance with plan?	**Selected from multiple outcomes?**	**Selected from multiple analyses?**	
Y/PY	N/PN	N/PN	Low
N/PN/NI	N/PN	N/PN	Some concerns
Any answer	N/PN	NI	Some concerns
Any answer	NI	N/PN	Some concerns
Any answer	NI	NI	Some concerns
Any answer	Either 5.2 or 5.3 Y/PY		High

Y/PY = 'Yes' or 'Probably yes'; N/PN = 'No' or 'Probably no'; NI = 'No information'

Figure 7. Suggested algorithm for reaching risk-of-bias judgements for bias in selection of the reported result. This is only a suggested decision tree: all default judgements can be overridden by assessors.



9 Acknowledgements

The development of the RoB 2 tool was supported by the MRC Network of Hubs for Trials Methodology Research (MR/L004933/2- N61), with the support of the host MRC ConDuCT-II Hub (Collaboration and innovation for Difficult and Complex randomised controlled Trials In Invasive procedures - MR/K025643/1), by MRC research grant MR/M025209/1, and by a grant from The Cochrane Collaboration.

10 Contributors

Core group: Julian Higgins, Jonathan Sterne, Jelena Savović, Matthew Page, Roy Elbers

Contributors to bias domain development: Natalie Blencowe, Isabelle Boutron, Christopher Cates, Rachel Churchill, Mark Corbett, Nicky Cullum, Jonathan Emberson, Sally Hopewell, Asbjørn Hróbjartsson, Sharea Ijaz, Peter Jüni, Jamie Kirkham, Toby Lasserson, Tianjing Li, Barney Reeves, Sasha Shepperd, Ian Shrier, Lesley Stewart, Kate Tilling, Ian White, Penny Whiting

Other contributors: Henning Keinke Andersen, Mike Clarke, Jon Deeks, Miguel Hernán, Daniela Junqueira, Yoon Loke, Geraldine MacDonald, Richard Morris, Mona Nasser, Nishith Patel, Jani Ruotsalainen, Holger Schünemann, Jayne Tierney, Sunita Vohra, Liliane Zorzela

11 References

1. Mansournia MA, Higgins JPT, Sterne JAC, Hernán MA. Biases in randomized trials: a conversation between trialists and epidemiologists. Epidemiology. 2016 (Published online 29 September).
2. Sterne JAC, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in `meta-epidemiological' research. Stat Med. 2002;21(11):1513-24.
3. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273:408-12.
4. Savović J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012;157(6):426-38.
5. Fergusson D, Aaron SD, Guyatt G, Hebert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ. 2002;325(7365):652-4.
6. Menerit CL. Clinical Trials – Design, Conduct, and Analysis. Second Edition: Oxford University Press; 2012.
7. Piantadosi S. Clinical Trials: A Methodologic perspective. Second Edition: Wiley; 2005.
8. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319(7211):670-4.
9. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? Clin Trials. 2007;4(4):350-6.
10. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. BMC Med Res Methodol. 2014;14:118.
11. Hernan MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. Clin Trials. 2012;9(1):48-55.
12. Shrier I, Steele RJ, Verhagen E, Herbert R, Riddell CA, Kaufman JS. Beyond intention to treat: what is the right question? Clin Trials. 2014;11(1):28-37.
13. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. N Engl J Med. 2017;377(14):1391-8.
14. Savovic J, Turner RM, Mawdsley D, Jones HE, Beynon R, Higgins JPT, et al. Association between risk-of-bias assessments and results of randomized trials in Cochrane Reviews: the ROBES meta-epidemiologic study. Am J Epidemiol. 2017.
15. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408-12.
16. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 2001;323(7303):42-6.
17. Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. Ann Intern Med. 2002;136(3):254-9.
18. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet. 2002;359(9305):515-9.

19. Schulz KF, Grimes DA. The Lancet Handbook of Essential Concepts in Clinical Research. Edinburgh, UK: Elsevier; 2006.
20. Brown S, Thorpe H, Hawkins K, Brown J. Minimization--reducing predictability for multi-centre trials whilst retaining balance within centre. Stat Med. 2005;24(24):3715-27.
21. Clark L, Fairhurst C, Torgerson DJ. Allocation concealment in randomised controlled trials: are we getting better? BMJ. 2016;355:i5663.
22. Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. Lancet. 2002;359(9310):966-70.
23. Berger VW, Ivanova A, Knoll MD. Minimizing predictability while retaining balance through the use of less restrictive randomization procedures. Stat Med. 2003;22(19):3017-28.
24. Berger VW. Quantifying the magnitude of baseline covariate imbalances resulting from selection bias in randomized clinical trials. Biometrical journal Biometrische Zeitschrift. 2005;47(2):119-27; discussion 28-39.
25. Schulz KF. Subverting randomization in controlled trials. JAMA. 1995;274(18):1456-8.
26. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 2001;323(7303):42-6.
27. Turner L, Shamseer L, Altman DG, Weeks L, Peters J, Kober T, et al. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. The Cochrane database of systematic reviews. 2012;11:MR000030.
28. Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol. 2007;36(4):847-57.
29. Corbett MS, Higgins JPT, Woolacott NF. Assessing baseline imbalance in randomised trials: implications for the Cochrane risk of bias tool. Research synthesis methods. 2014;5(1):79-85.
30. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hrobjartsson A, Savovic J. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLoS One. 2016;11(7):e0159267.
31. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA. 2002;287(22):2973-82.
32. Hartling L, Hamm MP, Fernandes RM, Dryden DM, Vandermeer B. Quantifying bias in randomized controlled trials in child health: a meta-epidemiological study. PLoS One. 2014;9(2):e88008.
33. Unverzagt S, Prondzinsky R, Peinemann F. Single-center trials tend to provide larger treatment effects than multicenter trials: a systematic review. J Clin Epidemiol. 2013;66(11):1271-80.
34. Chapin RB, Williams DC, Adair RF. Diabetes control improved when inner-city patients received graphic feedback about glycosylated hemoglobin levels. J Gen Intern Med. 2003;18(2):120-4.
35. Saugstad OD, Rootwelt T, Aalen O. Resuscitation of asphyxiated newborn infants with room air or oxygen: an international controlled trial: the Resair 2 study. Pediatrics. 1998;102(1):e1.
36. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359(9306):614-8.
37. Bellomo R, Chapman M, Finfer S, Hickling K, Myburgh J. Low-dose dopamine in patients with early renal dysfunction: a placebo-controlled randomised trial. Australian and New Zealand Intensive Care Society (ANZICS) Clinical Trials Group. Lancet. 2000;356(9248):2139-43.
38. Smilde TJ, van Wissen S, Wollersheim H, Trip MD, Kastelein JJ, Stalenhoef AF. Effect of aggressive versus conventional lipid lowering on atherosclerosis progression in familial hypercholesterolaemia (ASAP): a prospective, randomised, double-blind trial. Lancet. 2001;357(9256):577-81.
39. de Gaetano G. Low-dose aspirin and vitamin E in people at cardiovascular risk: a randomised trial in general practice. Collaborative Group of the Primary Prevention Project. Lancet. 2001;357(9250):89-95.
40. Brightling CE, Monteiro W, Ward R, Parker D, Morgan MD, Wardlaw AJ, et al. Sputum eosinophilia and short-term response to prednisolone in chronic obstructive pulmonary disease: a randomised controlled trial. Lancet. 2000;356(9240):1480-5.
41. Fu R, Vandermeer BW, Shamliyan TA, O'Neil ME, Yazdi F, Fox SH, et al. AHRQ Methods for Effective Health Care: Handling Continuous Outcomes in Quantitative Synthesis. Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008.
42. Schulz KF. Subverting randomization in controlled trials. JAMA. 1995;274(18):1456-8.
43. Altman DG, Bland JM. How to randomize. BMJ. 1999;319:703-4.
44. Senn S. Seven myths of randomisation in clinical trials. Stat Med. 2013;32(9):1439-50.
45. Wright IS, Marple CD, Beck DF. Report of the Committee for the Evaluation of Anticoagulants in the Treatment of Coronary Thrombosis with Myocardial Infarction; a progress report on the statistical analysis of the first 800 cases studied by this committee. Am Heart J. 1948;36(6):801-15.

46. Trowman R, Dumville JC, Torgerson DJ, Cranny G. The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. J Clin Epidemiol. 2007;60(12):1229-33.
47. Carlisle JB. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. Anaesthesia. 2017;72(8):944-52.
48. Badawy MSH, Hassanin IMA. The value of magnesium sulfate nebulization in treatment of acute bronchial asthma during pregnancy. Egyptian Journal of Chest Diseases and Tuberculosis. 2014;63(2):285-9.
49. Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. BMJ. 2016;355:i4919.
50. Hernan MA, Scharfstein D. Cautions as Regulators Move to End Exclusive Reliance on Intention to Treat. Ann Intern Med. 2018;168(7):515-6.
51. Madsen MV, Gotzsche PC, Hrobjartsson A. Acupuncture treatment for pain: systematic review of randomised clinical trials with acupuncture, placebo acupuncture, and no acupuncture groups. BMJ. 2009;338:a3115.
52. Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hrobjartsson A, et al. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. PLoS Med. 2006;3(10):e425.
53. Jensen JS, Bielefeldt AO, Hrobjartsson A. Active placebo control groups of pharmacological interventions were rarely used but merited serious consideration: a methodological overview. J Clin Epidemiol. 2017;87:35-46.
54. Bello S, Moustgaard H, Hrobjartsson A. Unreported formal assessment of unblinding occurred in 4 of 10 randomized clinical trials, unreported loss of blinding in 1 of 10 trials. J Clin Epidemiol. 2017;81:42-50.
55. Fergusson D, Glass KC, Waring D, Shapiro S. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. BMJ. 2004;328(7437):432.
56. Rees JR, Wade TJ, Levy DA, Colford JM, Jr., Hilton JF. Changes in beliefs identify unblinding in randomized controlled trials: a method to meet CONSORT guidelines. Contemp Clin Trials. 2005;26(1):25-37.
57. Hrobjartsson A, Forfang E, Haahr MT, Als-Nielsen B, Brorson S. Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding. Int J Epidemiol. 2007;36(3):654-63.
58. Sackett DL. Commentary: Measuring the success of blinding in RCTs: don't, must, can't or needn't? Int J Epidemiol. 2007;36(3):664-5.
59. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA. 2001;285(15):2000-3.
60. Boutron I, Estellat C, Ravaud P. A review of blinding in randomized controlled trials found results inconsistent and questionable. J Clin Epidemiol. 2005;58(12):1220-6.
61. Haahr MT, Hrobjartsson A. Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors. Clin Trials. 2006;3(4):360-5.
62. Montori VM, Bhandari M, Devereaux PJ, Manns BJ, Ghali WA, Guyatt GH. In the dark: the reporting of blinding status in randomized controlled trials. J Clin Epidemiol. 2002;55(8):787-90.
63. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Bmj. 2010;340:c332.
64. Weinstein JN, Tosteson TD, Lurie JD, Tosteson AN, Hanscom B, Skinner JS, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. JAMA. 2006;296(20):2441-50.
65. De Bruyne B, Fearon WF, Pijls NH, Barbato E, Tonino P, Piroth Z, et al. Fractional flow reserve-guided PCI for stable coronary artery disease. N Engl J Med. 2014;371(13):1208-17.
66. Biere SS, van Berge Henegouwen MI, Maas KW, Bonavina L, Rosman C, Garcia JR, et al. Minimally invasive versus open oesophagectomy for patients with oesophageal cancer: a multicentre, open-label, randomised controlled trial. Lancet. 2012;379(9829):1887-92.
67. St Peter SD, Adibe OO, Juang D, Sharp SW, Garey CL, Laituri CA, et al. Single incision versus standard 3-port laparoscopic appendectomy: a prospective randomized trial. Ann Surg. 2011;254(4):586-90.
68. Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012;157(6):429-38.
69. Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 2014;43(4):1272-83.

70. Nuesch E, Reichenbach S, Trelle S, Rutjes AW, Liewald K, Sterchi R, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. Arthritis Rheum. 2009;61(12):1633-41.
71. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine--selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ. 2003;326(7400):1171-3.
72. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. J Clin Epidemiol. 2007;60(7):663-9.
73. Abraha I, Cherubini A, Cozzolino F, De Florio R, Luchetta ML, Rimland JM, et al. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ. 2015;350:h2445.
74. Tierney JF, Stewart LA. Investigating patient exclusion bias in meta-analysis. Int J Epidemiol. 2005;34(1):79-87.
75. Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: systematic review. BMJ. 2010;340:c2697.
76. National Research Council. The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press; 2010.
77. Higgins JPT, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. Clin Trials. 2008;5(3):225-39.
78. Fleming TR. Addressing missing data in clinical trials. Ann Intern Med. 2011;154(2):113-7.
79. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Second Edition. Hoboken, NJ: Wiley; 2002.
80. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367(14):1355-60.
81. Robins J, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. Journal of the American Statistical Association. 1995;90:106-21.
82. May GS, DeMets DL, Friedman LM, Furberg C, Passamani E. The randomized clinical trial: bias in analysis. Circulation. 1981;64(4):669-73.
83. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: John Wiley & Sons; 1987.
84. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
85. Sullivan TR, White IR, Salter AB, Ryan P, Lee KJ. Should multiple imputation be the method of choice for handling missing data in randomized trials? Stat Methods Med Res. 2016:962280216683570.
86. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. Clin Trials. 2012;9(4):396-407.
87. Leurent B, Gomes M, Faria R, Morris S, Grieve R, Carpenter JR. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. Pharmacoeconomics. 2018.
88. Akl EA, Kahale LA, Agoritsas T, Brignardello-Petersen R, Busse JW, Carrasco-Labra A, et al. Handling trial participants with missing outcome data when conducting a meta-analysis: a systematic survey of proposed approaches. Syst Rev. 2015;4:98.
89. Akl EA, Johnston BC, Alonso-Coello P, Neumann I, Ebrahim S, Briel M, et al. Addressing dichotomous data for participants excluded from trial analysis: a guide for systematic reviewers. PLoS One. 2013;8(2):e57132.
90. Ebrahim S, Akl EA, Mustafa RA, Sun X, Walter SD, Heels-Ansdell D, et al. Addressing continuous data for participants excluded from trial analysis: a guide for systematic reviewers. J Clin Epidemiol. 2013;66(9):1014-21 e1.
91. Ebrahim S, Johnston BC, Akl EA, Mustafa RA, Sun X, Walter SD, et al. Addressing continuous data measured with different instruments for participants excluded from trial analysis: a guide for systematic reviewers. J Clin Epidemiol. 2014;67(5):560-70.
92. Mavridis D, White IR, Higgins JP, Cipriani A, Salanti G. Allowing for uncertainty due to missing continuous outcome data in pairwise and network meta-analysis. Stat Med. 2015;34(5):721-41.
93. Akl EA, Kahale LA, Ebrahim S, Alonso-Coello P, Schunemann HJ, Guyatt GH. Three challenges described for identifying participants with missing data in trials reports, and potential solutions suggested to systematic reviewers. J Clin Epidemiol. 2016;76:147-54.
94. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials. 2004;1(4):368-76.
95. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med. 2001;135(11):982-9.

96. Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. Stat Med. 2007;26(14):2745-58.
97. Nuesch E, Trelle S, Reichenbach S, Rutjes AW, Burgi E, Scherer M, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ. 2009;339:b3244.
98. Rothman KJ, Greenland S, Lash TL. Validity in epidemiological studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology. 3rd ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2008. p. 128-47.
99. Clemens JD, Chuong JJ, Feinstein AR. The BCG controversy. A methodological and statistical reappraisal. JAMA. 1983;249(17):2362-9.
100. Mangtani P, Abubakar I, Ariti C, Beynon R, Pimpin L, Fine PE, et al. Protection by BCG vaccine against tuberculosis: a systematic review of randomized controlled trials. Clin Infect Dis. 2014;58(4):470-80.
101. Sackett DL. Bias in analytic research. J Chronic Dis. 1979;32(1-2):51-63.
102. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health and quality of life outcomes. 2006;4:79.
103. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012;344:e1119.
104. Engebretsen K, Grotle M, Bautz-Holter E, Ekeberg OM, Juel NG, Brox JI. Supervised exercises compared with radial extracorporeal shock-wave therapy for subacromial shoulder pain: 1-year results of a single-blind randomized controlled trial. Phys Ther. 2011;91(1):37-47.
105. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ. 2010;340:c365.
106. Page MJ, Higgins JPT. Rethinking the assessment of risk of bias due to selective reporting: a cross-sectional study. Systematic reviews. 2016;5(1):108.
107. Williamson P, Altman D, Blazeby J, Clarke M, Gargon E. Driving up the quality and relevance of research through the use of agreed core outcomes. J Health Serv Res Policy. 2012;17(1):1-2.
108. Kirkham JJ, Davis K, Altman DG, Blazeby JM, Clarke M, Tunis S, et al. Core Outcome Set-STAndards for Development: The COS-STAD recommendations. PLoS Med. 2017;14(11):e1002447.
109. Cochrane Skin Group Core Outcome Set Initiative [Available from: https://www.uniklinikum-dresden.de/COUSIN.
110. Dwan K, Altman DG, Clarke M, Gamble C, Higgins JPT, Sterne JAC, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. PLoS Med. 2014;11(6):e1001666.