Revised Cochrane risk of bias tool for randomized trials (RoB 2.0)

Edited by Julian PT Higgins, Jelena Savović, Matthew J Page, Jonathan AC Sterne
on behalf of the development group for RoB 2.0

20th October 2016

Contents

1 Introduction

The RoB 2.0 tool provides a framework for considering the risk of bias in the findings of any type of randomized trial. The assessment is structured into a series of domains through which bias might be introduced into a trial. All domains are mandatory, and no further domains should be added. We offer several templates for addressing these domains, tailored to the following study designs.

(1) Randomized parallel group trials.
(2) Cluster-randomized trials (including those in which multiple body parts are randomized to the same intervention).
(3) Randomized cross-over trials and other matched designs.

The RoB 2.0 assessment for individually randomized trials (including cross-over trials) has five domains, as follows.

(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.

We have avoided terms used in the previous version of the tool (e.g. selection bias, performance bias, attrition bias, detection bias) because we have found that they cause a lot of confusion (1). In particular, the same terms are sometimes used to refer to different types of bias, and different types of bias are often described by a host of different terms.

For cluster-randomized trials, an additional domain is included ((1b) Bias arising from the timing of identification and recruitment of individual participants)

Assessment of risk of bias is specific to a particular result for a particular outcome (and time point) in the study. However, some domains will apply generally to the whole study (such as biases arising from the randomization process); some will apply mainly to the outcome being measured (such as bias due to deviations from intended intervention); some will apply mainly to the measurement method used for the outcome (such as bias in measurement of outcomes); and some will apply to the specific result (such as bias in selection of the reported result).

Trials will frequently (if not 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. At this point, we have not 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.

In future work we plan to develop recommendations for how to extract information so that items are collected without unnecessary duplication and in the order in which they are likely to be encountered in reports of a trial.

This document describes the main features of the RoB 2.0 tool, and provides guidance for its application to individually randomized, parallel group trials. Supplementary documents provide (i) additional considerations for cluster-randomized trials and cross-over trials; (ii) full presentations of the RoB 2.0 tool for each of the three trial designs.

1.1 Specifying the nature of the target comparison (effect of interest)

We offer two variants of the tool, representing an important distinction between two types of intervention effect that review authors might be interested in quantifying. These are:

(1) the effect of **assignment** to the interventions at baseline (regardless of whether the interventions are received during follow-up); or
(2) the effect of **starting and adhering to** the interventions as specified in the trial protocol.

For example, to inform a health policy question about whether to recommend an intervention in a particular health system we would probably estimate the effect of *assignment to intervention*, whereas to inform a care decision by an individual patient we would wish to estimate the effect of receiving the intervention according to a specified protocol (that is, the effect of starting and adhering to the intervention). Review authors need to define the intervention effect in which they are interested in the review (or in each meta-analysis), and apply the risk of bias tool appropriately to this effect. Differences in the risk of bias assessment mainly relate to biases due to deviations from the intended intervention, and issues related to the two types of effect of interest are discussed in more detail in section 2.3.2.

Note that in the context of RoB 2.0, specification of the "effect of interest" does not relate to choice of treatment effect metric (odds ratio, risk difference etc.).

1.2 Signalling questions

A key feature of the tool is the inclusion of signalling questions within each domain of bias. The signalling questions aim to be reasonably factual in nature. Their primary purpose is to facilitate 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 their simplicity and clarity, the signalling questions are phrased such that a response of "Yes" may be indicative of either a low or high risk of bias, depending on the question.

Responses of "Yes" and "Probably yes" have similar implications, as do responses of "No" and "Probably no". The definitive versions ("Yes" and "No") would 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 measures of agreement are applied to answers to the signalling questions, we recommend grouping the pairs of responses.

The "No information" response should be used only when insufficient data are reported to permit a reasonable judgement to be made.

Some signalling questions are answered only if the response to a previous question is "Yes" or "Probably yes" (or "No" or "Probably no"). For such signalling questions an additional response option "Not applicable" is available.

Signalling questions should be answered independently, i.e. the answer to one question should not affect answers to other questions in the same domain or other domains.

1.2.1 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 the 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 when possible.

1.3 Risk of bias judgements

1.3.1 Domain-level judgements about risk of bias

RoB 2.0 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 addressed.

Use of the word "judgement" is important for the risk of bias assessment. Review authors must use the answers to the signalling questions as well as any other additional understanding of the study and its context to determine whether the result of the study is at risk of bias. "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.

The **response options for each risk of bias judgement** are:

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

The key to applying the tool is to make domain-level judgements about risk of bias that mean the same across the five domains with respect to concern about the impact of bias on the trustworthiness of the result. If domain-level judgements are made consistently, then judging the overall risk of bias for a particular outcome is relatively straightforward (see 1.3.4). Review authors need to consider both the severity of the bias in a particular domain and the relative consequences of bias in different domains.

1.3.2 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 essential that the reasons are provided for any judgements of "Some concerns" or "High" risk of bias.

1.3.3 Direction of bias

The tool includes an optional component to judge 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 participant from their assigned intervention to the other intervention would (in the context interest in the effect of adhering to intervention) lead to bias towards the null. However, for other domains, the bias is likely to favour one of the interventions being compared. Such favouring will lead to an increase or decrease in the effect estimate, affecting the magnitude or direction of the effect size. Many biases are likely to fall into this category, although the direction of the bias may be less easy to predict; examples include manipulation of the randomization process, awareness of interventions received 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 attempt to guess it.**

1.3.4 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 be used to map risk of bias judgements within domains to a single risk of bias 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 be at **some concerns** in at least one domain for this result.
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.

Declaring a study to be at a particular level of risk of bias for an individual domain will mean that the study as a whole has a risk of bias at least this severe (for the result being assessed). Therefore a judgement of "High" risk of bias within any domain should have similar implications for the study as a whole, irrespective of which domain is being assessed.

The typical mapping of domain-level judgements to overall judgements is described in Table 1. However, in some cases several domain-level judgements of "Some concerns" for the same outcome might be considered to be additive, so that "Some concerns" in multiple domains can lead to an overall judgement of "High" risk of bias overall for that outcome or group of outcomes.

2 Detailed guidance

2.1 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 *starting and adhering to intervention*. It is also helpful to document the sources that are used to complete the assessment. The RoB 2.0 standard template includes questions to capture these details (Box 1).

Box 1. The RoB 2.0 tool (part 1): Preliminary considerations

Study design

☐ Randomized parallel group trial

☐ Cluster-randomized trial

☐ Randomized cross-over or other matched design

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 your aim for this study…?

☐ to assess the effect of *assignment to intervention*

☐ to assess the effect of *starting and adhering to intervention*

Which of the following sources have you obtained to help inform your risk of bias judgements (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

2.2 Bias arising from the randomization process

2.2.1 Introduction

The unique strength of randomization is that, if successfully accomplished, it prevents bias in allocating interventions to participants. The success of randomization in this respect depends on fulfilling several interrelated processes. A rule for allocating interventions to participants must be specified, based on some chance (random) process. We call this **allocation sequence generation**. Furthermore, steps must be taken to secure strict implementation of that schedule of random assignments by preventing foreknowledge of the forthcoming allocations. This process is often termed **allocation sequence concealment**. Thus, one suitable method for assigning interventions would be to use a simple random (and therefore unpredictable) sequence, and to conceal the upcoming allocations from those involved in enrolment into the trial.

The starting point for an unbiased intervention study is the use of a mechanism that ensures that the same sorts of participants receive each intervention. An allocation sequence must be generated that, if perfectly implemented, would balance prognostic factors, on average, evenly across intervention groups. Prognostic factors are factors that predict the outcome (e.g. severity of illness, presence of comorbidities), and if they are not balanced across the intervention arms they result in confounding. Confounding is common in observational studies of the effects of interventions, because treatment decisions in routine care are often influenced by prognostic factors. The key purpose of randomization is to eliminate confounding, so that the intervention groups are balanced with respect to known and unknown prognostic factors. It can be argued that other assignment rules, such as alternation (alternating between two interventions) or rotation (cycling through more than two interventions), can achieve the same thing (2). However, a theoretically unbiased rule is insufficient to prevent bias in practice. If future assignments can be anticipated, either by predicting them or by knowing them, then bias can arise due to the selective enrolment and non-enrolment of participants into a study in the light of the upcoming intervention assignment.

Future assignments may be anticipated for several reasons. These include (i) knowledge of a deterministic assignment rule, such as by alternation, date of birth or day of admission; (ii) knowledge of the sequence of assignments, whether randomized or not (e.g. if a sequence of random assignments is posted on the wall); (iii) ability to predict assignments successfully, based on previous assignments (which may sometimes be possible when randomization methods are used that attempt to ensure an exact ratio of allocations to different interventions). Complex interrelationships between theoretical and practical aspects of allocation in intervention studies make the assessment of bias in the randomization process challenging. Perhaps the most important among the practical aspects is concealment of the allocation sequence, that is, the use of mechanisms to prevent foreknowledge of the next assignment.

Knowledge of the next assignment – for example, from a table of random numbers openly posted on a bulletin board – can cause 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 deliberately directed to the "appropriate" intervention, which can often be accomplished by delaying a participant's entry into the trial until the next appropriate allocation appears. In epidemiological terms, this may introduce **confounding**.

A randomized sequence is not always completely unpredictable, even if mechanisms to try and maintain allocation concealment are in place. For example, unsealed allocation envelopes may be opened, while translucent envelopes may be held against a bright light to reveal the contents (Schulz 1995a, Schulz 1995b, Jüni 2001). Personal accounts suggest that many allocation schemes have been deciphered by investigators because the methods of concealment were inadequate (Schulz 1995a). Another example is the use of blocked randomization in an unblinded trial, or in a blinded trial where the blinding is broken, for example because of characteristic side effects. When blocked randomization is used, and when the assignments are revealed to the recruiter after each person is enrolled into the trial, then it is sometimes possible to predict future assignments. This is particularly the case when blocks are of a fixed size and are not divided across multiple recruitment centres (3).

Unfortunately, information on methods for generation of sequence generation and allocation concealment is often not fully reported in publications of randomized trials. 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 (4). Lack of description of methods of randomization and allocation concealment does not necessarily mean that these methods were inappropriate (5). This can sometimes be due to poor reporting or limited word count in journals.

The success (or not) of randomization in producing comparable groups is often claimed on the basis of examining baseline values of important prognostic factors. 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 (4). Thus Corbett et al have argued that risk of bias assessments should consider whether participant characteristics are balanced between intervention groups (6).

2.2.2 Empirical evidence of bias

A recent meta-analysis (7) of seven meta-epidemiological studies (8-14) found that inadequate or unclear (versus adequate) method of sequence generation was associated with a small (7%) exaggeration of intervention effect estimates. The bias was greater in trials of subjective outcomes, while trials assessing all-cause mortality and other objective outcomes appeared unbiased, on average. Similarly, a modest (10%) exaggeration of intervention effect estimates was observed in a meta-analysis (7) of seven empirical studies (8-13, 15) for trials with inadequate/unclear (versus adequate) concealment of allocation. The average bias associated with inadequate allocation concealment was greatest in trials of subjective outcomes and in trials of complementary and alternative medicine, with no evidence of bias in trials of mortality or other objective outcomes. Three empirical studies that assessed the effect of baseline imbalances on intervention effect estimates found no evidence that imbalances inflate intervention effect estimates (12, 16, 17), but all estimates were imprecise. Also, there was little evidence that intervention effect estimates were exaggerated in trials without adjustment for confounders (16) or in unblinded trials with block randomization (12). However, each characteristic was only examined in a single small study.

2.2.3 Assessing random sequence generation

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

Randomization with no constraints to generate an allocation sequence is called **simple randomization** or **unrestricted randomization**. In principle, this could be achieved by allocating interventions using methods such as repeated coin-tossing, throwing dice or dealing previously shuffled cards (18, 19). More usually a list of random assignments is generated by a computer, or randomization may be achieved by referring to a published list of random numbers.

> *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."

Sometimes **restricted randomization** is used to generate a sequence to ensure particular allocation ratios to the intervention groups (e.g. 1:1). Blocked randomization is a common form of restricted randomization (18, 19). Blocking ensures that the numbers of participants to be assigned to each of the comparison groups will be balanced within blocks of, for example, five in one group and five in the other for every 10 consecutively entered participants. The block size may be randomly varied to reduce the likelihood of foreknowledge of intervention assignment (random permuted blocks).

> *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."

Also common is **stratified randomization**, in which restricted randomization is performed separately within strata. This generates separate randomization schedules for subsets of participants defined by potentially important prognostic factors, such as disease severity and study centres. If simple (rather than restricted) randomization was used in each stratum, then stratification would have no effect but the randomization would still be valid. Risk of bias may be judged in the same way whether or not a trial claims to have stratified its randomization.

Another approach that incorporates both the general concepts of stratification and restricted randomization is **minimization**, which can be used to make small groups closely similar with respect to several characteristics. The use of minimization should not automatically be considered to put a study at risk of bias. However, some methodologists remain cautious about the acceptability of minimization, particularly when it is used without any random component, while others consider it to be very attractive (20).

Other adequate types of randomization that are sometimes used are biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization (18, 21, 22). If these or other approaches are encountered, consultation with a statistician may be necessary.

2.2.3.1 Inadequate methods of sequence generation

Systematic methods, such as alternation, assignment based on date of birth, case record number and date of presentation, are sometimes referred to as "quasi-random". 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: "We allocated patients to the intervention group based on the week of the month."

Example of non-random sequence generation: "Patients born on even days were assigned to Treatment A and patients born on odd days were assigned to Treatment B."

2.2.3.2 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. It is, unfortunately, common for authors to use the term "randomized" even when it is not justified: many trials with declared systematic allocation are described by the authors as randomized. Sometimes trial authors provide some information, but they incompletely define their approach and do not confirm some random component in the process. For example, authors may state that blocked randomization was used, but the process of selecting the blocks, such as a random number table or a computer random number generator, was not specified. Other information available (i.e. answering "Probably yes; or "Probably no" to the signalling question) or a response of "No information" may be provided.

Sometimes no information at all about sequence generation is reported. Some assessors may decide that if the allocation sequence was adequately concealed this means that the sequence likely to be random or unpredictable. Although this is often a reasonable assumption, it is possible to produce a biased (unbalanced) allocation sequence which has been successfully concealed from participants and enrolling investigators. This could be done deliberately or could be unintentional. For example, an error in the computer randomization code might place sicker people preferentially in one group when stratified randomization is used. However, this could be the case even if the random sequence was simply described as "computer generated".

2.2.4 Assessing concealment of allocation sequence

Some review authors confuse allocation concealment with blinding of assigned interventions. Allocation concealment seeks to prevent bias in intervention assignment by protecting the allocation sequence before and until assignment, and can always be successfully implemented regardless of the study topic (23, 24). In contrast, blinding seeks to prevent bias by protecting the sequence after assignment (24, 25), 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.

Among the different methods used to conceal allocation, central randomization by a third party is perhaps the most desirable. Methods using envelopes are more susceptible to manipulation than other approaches (23). 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.

Adequate methods of allocation sequence concealment
Table 2 provides minimal criteria for a judgement of adequate concealment of allocation sequence (left) and extended criteria, which provide additional assurance that concealment of the allocation sequence was indeed adequate (right).

Examples of allocation sequence concealment (as compiled by Schulz and Grimes (26)):

> " . . . 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." (27).
>
> ". . . concealed in sequentially numbered, sealed, opaque envelopes, and kept by the hospital pharmacist of the two centres." (28).
>
> "Treatments were centrally assigned on telephone verification of the correctness of inclusion criteria . . ." (29).
>
> "Glenfield Hospital Pharmacy Department did the randomization, distributed the study agents, and held the trial codes, which were disclosed after the study." (30).

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 (IVRS)), fax, email or an interactive web response system (IWRS), 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.

2.2.5 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 (31). The RoB 2.0 tool includes consideration of situations in which baseline characteristics indicate that something may have gone wrong with the randomization process. This does not include chance imbalances, which we discuss in 2.2.5.1.

Severe baseline imbalances may arise as a result of deliberate actions of the trialists attempting to subvert the randomization process (32). They may also occur because of unintentional actions or errors that occurred due to insufficient safeguards. An example of the latter would be an error in writing a minimization programme such as writing a "plus" instead of a "minus", leading to maximizing instead of minimizing differences between groups. Such errors would lead to a failure of the randomization process despite not being a deliberate attempt to subvert it.

Indicators from baseline imbalance that randomization was not performed adequately include the following, which we address in turn.

(1) unusually large differences between intervention group sizes;
(2) a substantial excess in statistically significant differences in baseline characteristics than would be expected by chance alone;
(3) imbalance in key prognostic factors (or baseline measures of outcome variables) that are unlikely to be due to chance;
(4) excessive similarity in baseline characteristics that is not compatible with chance;
(5) surprising absence of one or more key characteristics that would be expected to be reported.

The first indicator of baseline imbalance is a substantial difference in the numbers of participants randomized to each intervention group compared with the intended allocation ratio. One example is a 1948 trial comparing anticoagulation medication to conventional treatment for myocardial infarction (33). 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 very 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 so that they would receive the new anti-coagulant (33).

The second indicator of baseline imbalance is a substantial excess of baseline differences between groups beyond what would be expected by chance. It is widely understood that statistical tests of differences in baseline characteristics should not be used in truly randomized trials, because it is known that the null hypothesis (i.e. that the participants were randomized) is true. However, they do provide a valid test of the different null hypothesis that randomization was truly used and implemented successfully. Statistical tests (e.g. P values) for comparisons between groups are often presented in tables of baseline characteristics, but they must be interpreted appropriately. Under randomization, one in 20 variables would be expected to be statistically significant at a 5% significance level, purely by chance. However, if a substantial excess of 1 in 20 baseline characteristics differ between groups, or if P values are extremely small, then this is less likely to have occurred by chance and may suggest problems with the randomization process.

The third indicator of baseline imbalance is an important difference in prognostic factors or baseline measures of outcome. This is because these are the factors that might influence those recruiting participants into the study. They therefore have more potential to be manipulated by investigators who want to influence the result of the trial. Knowledge of key prognostic factors for the topic area is therefore helpful when assessing baseline comparability. The review team should, where possible, identify in advance the key prognostic factors that may influence the outcome of interest. Key prognostic factors are likely to be identified both through the knowledge of subject matter experts who are members of the review group, and through initial (scoping) reviews of the literature. Discussions with health professionals who make intervention decisions for the target patient or population groups may also be helpful. The magnitude and importance of differences in prognostic factors should not be assessed using P values. Instead, minimum clinically important differences for continuous outcomes might be considered. For example, if a 2 point difference in VAS pain (0-10 scale) is an accepted important difference, review authors may decide that a 1 point difference is an important difference between groups at baseline. A convenient way to compare baseline differences is to compute standardized mean differences between groups, and this is becoming increasingly common.

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 (34).

A fourth cause for concern when considering baseline balance is when baseline characteristics are excessively *similar* across intervention groups, such that they are unlikely to be compatible with the randomization methods. Note that restricted randomization methods (see section 2.2.3) tend to give rise the groups that are more similar at baseline than simple randomization methods.

A fifth cause for concern regarding the comparability of groups arises if baseline data are not available for a potentially relevant measure (e.g. key prognostic factors, baseline measures). 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 (35). 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. The fact this was not reported may lead us to suspect that disease severity differed between groups at baseline.

To be able to assess baseline imbalance, baseline data should be presented 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 missing data needs to be considered. This practice of reporting baseline characteristics of analysed participants only is not common in trials of clinical medicine, but it may be very common in other areas (e.g. social care).

2.2.5.1 Chance imbalances at baseline

In trials using large samples (usually meaning at least 100 in each randomized group (18, 19, 21)), simple randomization generates comparison groups of relatively similar sizes. In trials using small samples, simple randomization will sometimes produce an allocation sequence leading to groups that differ, by chance, quite substantially in size or in the occurrence of prognostic factors (i.e. "case-mix" variation) (36).

Chance imbalances are not a source of bias, and the RoB 2.0 tool does not aim to identify imbalances in baseline variables that have arisen due to chance. It can be difficult, however, to distinguish chance imbalances from imbalances due to problems with the randomization process.

If a meta-analysis included a reasonable number of studies, chance imbalances would be expected to act in opposite directions and would not be a concern across the body of evidence. Across several trials, however, a pattern might emerge such that imbalances tend to favour one intervention over the other. This provides a suggestion that the imbalances are due to bias rather than chance.

2.2.6 Using the "Bias arising from the randomization process" domain of the RoB 2.0 tool

Bias in the process of randomization is the only domain in the tool that is assessed at the study level.

Signalling questions for this domain are provided in Box 2. Note that the answer to one signalling question should not affect answers to other questions. Therefore, 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. Any concerns about the observed imbalance should be raised 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. A "High" risk of bias should be given only if there are concerns sufficient to suspect that the study result itself is at high risk of bias. 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 (32). Judgements should be supported with their rationale in the free text box provided.

Box 2. The RoB 2.0 tool (part 2): Risk of bias arising from the randomization process

Signalling questions	Elaboration	Response options
1.1 Was the allocation sequence random?	"Yes" if a random component was used in the sequence generation process such as using a computer generated random numbers, referring to a random number table, minimization, coin tossing; shuffling cards or envelopes; throwing dice; or drawing of lots. Minimization may be implemented without a random element, and this is considered to be equivalent to being random. "No" if the sequence is non-random, such that it is either likely to introduce confounding, or is predictable or difficult to conceal, e.g. alternation, methods based on dates (of birth or admission) or patient record numbers, allocation decision made by clinicians or participants, based on the availability of the intervention, or any other systematic or haphazard method. If the only information about randomization methods is to state that the study is randomized, then this signalling question should generally be answered as "No information". There may be situations in which a judgement is 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, then 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. Similarly, if participants and personnel are all unaware of intervention assignments throughout/during the trial (blinding or masking), this may be an indicator that the allocation process was also concealed, but this will not necessarily always be the case. If the allocation sequence was clearly concealed but there is no information about how the sequence was generated, it will often be reasonable to assume that the sequence was random (although this will not necessarily always be the case).	Y / PY / PN / N / NI
1.2 Was the allocation sequence concealed until participants were recruited and assigned to interventions?	"Yes" if any form of remote or centrally administered randomization, where the process of allocation is controlled by an outsourced unit or organization, independent of the enrolment personnel (e.g. independent central pharmacy, telephone or internet-based randomization service providers). "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 (e.g. "Probably yes" or "Probably no"). "No" if there is reason to suspect the enrolling investigator or the participant had knowledge of the forthcoming allocation.	Y / PY / PN / N / NI
1.3 Were there baseline imbalances that suggest a problem with the randomization process?	*NB Imbalances that are small and compatible with chance should not be highlighted using the RoB 2.0 tool; chance imbalances are not 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) unusually large differences between intervention group sizes; or (2) a substantial excess in statistically significant differences in baseline characteristics than would be expected by chance alone; or (3) imbalance in key prognostic factors (or baseline measures of outcome variables) that are unlikely to be due to chance.	Y / PY / PN / N / NI

	An answer of "Yes/Probably yes" may exceptionally be given if the groups are surprisingly balanced in a way that appears incompatible with chance and the randomization methods, thus raising suspicion about the methods used. In some circumstances, it may be reasonable to answer "Yes/Probably yes" (rather than "No information") when there is a surprising lack of information on baseline characteristics when such information could reasonably be expected to be available/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 be used to 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).	
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

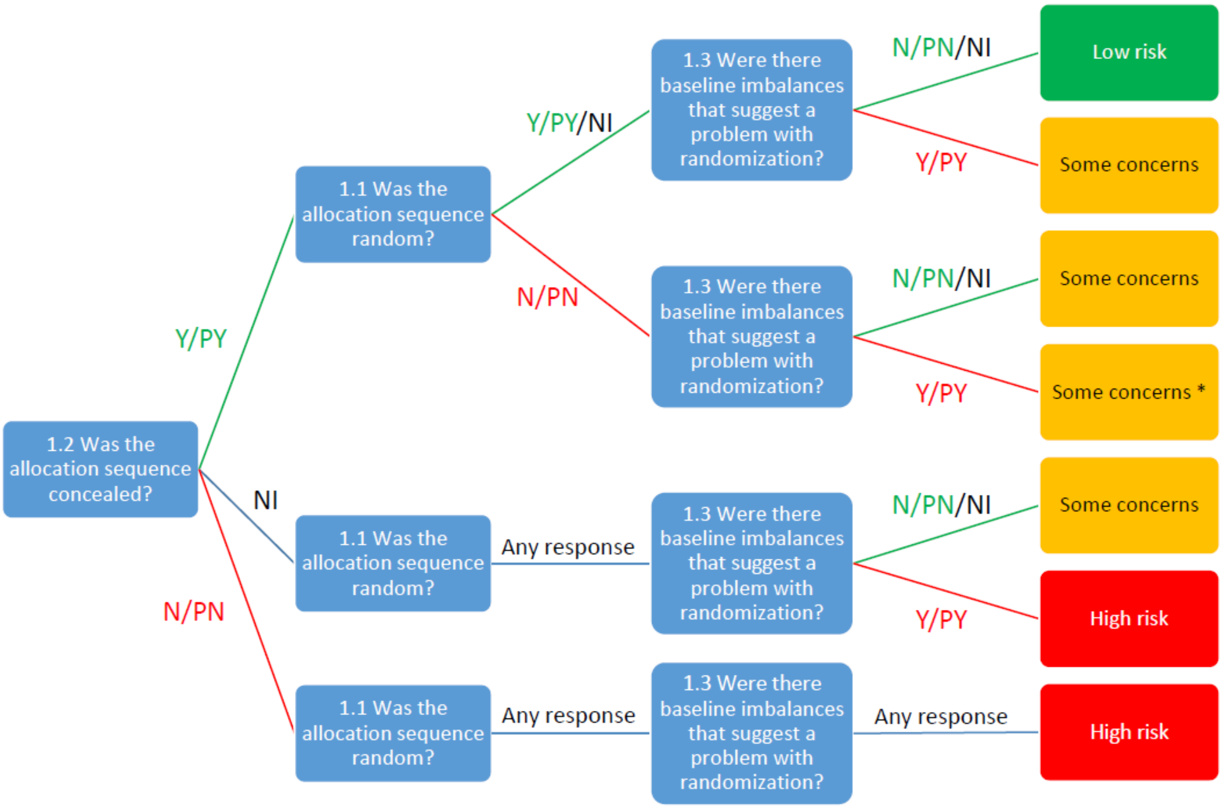
Low risk of bias	Allocation was adequately concealed AND There are no baseline imbalances across intervention groups at baseline appear to be compatible with chance AND An adequate (random or otherwise unpredictable) method was used to generate allocation sequence OR There is no information about the method used to generate the allocation sequence.
Some concerns	Allocation was adequately concealed AND There is a problem with the method of sequence generation OR Baseline imbalances suggest a problem with the randomization process OR No information is provided about concealment of allocation AND Baseline imbalances across intervention groups appear to be compatible with chance OR No information to answer any of the signalling questions.
High risk of bias	Allocation sequence was not concealed OR No information is provided about concealment of allocation sequence AND Baseline imbalances suggest a problem 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	**1.2**	**1.3**	**Default risk of bias**	**Remarks**
Y/PY	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 base might be reached.
Y/PY	N/PN	Any response	High	
Y/PY	NI	NI/N/PN	Some concerns	
Y/PY	NI	Y/PY	High	
N/PN	Y/PY	Any response	Some concerns	
N/PN	N/PN	Any response	High	
N/PN	NI	NI/N/PN	Some concerns	
N/PN	NI	Y/PY	High	
NI	Y/PY	NI/N/PN	Low	
NI	Y/PY	Y/PY	Some concerns	Substantial baseline imbalance may lead to a judgement of High risk of bias
NI	N/PN	Any response	High	
NI	NI	NI/N/PN	Some concerns	
NI	NI	Y/PY	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. (*In some cases a judgement of "High risk" would be appropriate.). This is only a suggested decision tree: all default judgements can be overridden by assessors.


1.2 Was the allocation sequence concealed?
Y/PY
NI
N/PN
1.1 Was the allocation sequence random?
Y/PY/NI
N/PN
1.1 Was the allocation sequence random?
Any response
1.1 Was the allocation sequence random?
Any response
1.3 Were there baseline imbalances that suggest a problem with randomization?
N/PN/NI
Y/PY
Low risk
Some concerns
1.3 Were there baseline imbalances that suggest a problem with randomization?
N/PN/NI
Y/PY
Some concerns
Some concerns *
1.3 Were there baseline imbalances that suggest a problem with randomization?
N/PN/NI
Y/PY
Some concerns
High risk
1.3 Were there baseline imbalances that suggest a problem with randomization?
Any response
High risk

2.3 Bias due to deviations from intended interventions

2.3.1 Introduction

This domain relates to biases that arise when there are systematic differences between the care provided to intervention and comparator groups, which represent a deviation from the intended interventions. Such differences (see Box 2) could reflect either additional aspects of care, 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. To be able to assess risk of bias for this domain, it is essential to understand what the intended interventions were.

In randomized trials, bias due to deviations from intended interventions can sometimes be reduced or avoided by masking (blinding) participants, carers, healthcare providers and trial personnel to the interventions received. Such masking, if successful, should prevent knowledge of the intervention assignment from influencing the "co-interventions" (interventions other than the intended interventions), contamination (application of one of the interventions in participants intended to receive the other), switches from the intended interventions to other interventions, non-adherence to interventions, or failure to implement some or all of the intervention as intended).

As described in Section 1.1, the RoB 2.0 tool offers two types of assessment for this domain depending on what type of question is of interest to the review authors, and specifically whether this is

(1) the effect of **assignment** to the interventions at baseline (regardless of whether the interventions are received or adhered to during follow-up); or
(2) the effect of **starting and adhering to** the interventions as specified in the trial protocol.

In the following section we discuss these two effects of interest in more detail.

2.3.2 The role of the nature of the target comparison (effect of interest)

Typically, the effect of interest will be either the effect of **assignment** to the interventions at baseline, regardless of whether the interventions are received and adhered to during the follow-up, or the effect of **starting and adhering to** the interventions as specified in the protocol.

(1) In randomized trials, the effect of **assignment** to intervention is estimated by an **intention-to-treat (ITT) analysis** that includes all randomized participants. In an ITT analysis, participants are analysed in the intervention groups to which they were randomized. This ensures that the benefits of randomization – that the two intervention groups should be similar with respect to measured and unmeasured prognostic factors – are maintained.

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 do not reflect usual practice, are not balanced between the intervention groups and also influence the outcome. For example, more closely monitoring patients randomized to a novel intervention than those randomized to standard care would increase the risk of bias, unless such care was 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 as equivalent to bias and include such non-routine deviations as an important consideration in the RoB 2.0 tool.

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. Although ITT effects may be regarded as 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 (37, 38), and are problematic for non-inferiority or equivalence studies (37), or for estimating harms.

(2) Patients and other stakeholders are often interested in the effect of **starting and adhering to the intervention** as described in the trial protocol (the **per protocol effect**). To examine this we need to look at issues such as how well the intervention was implemented, how well participants adhered to it (without discontinuing or switching to another intervention), and whether unintended co-interventions were received alongside the intended intervention and were unbalanced across intervention groups. Masking of on

participants, carers and trial personnel (see below) should ensure that unintended co-interventions were balanced across intervention groups.

It is possible to use data from randomized trials to estimate the effect of starting and adhering to intervention. However, the commonly used approaches are generally problematic. In particular, unadjusted analyses based on the intervention actually received, or naïve "per protocol" analyses restricted to individuals in each intervention group who started and adhered to the interventions can be biased, if prognostic factors influenced intervention received. Advanced statistical methods that adjust for the effects of these prognostic factors permit appropriate adjustment for such bias (37, 38), although applications of such methods are relatively rare. Methods that use randomization status as an instrumental variable bypass the need to adjust for prognostic factors, but they are not always applicable (37, 38).

2.3.3 The role of masking (or blinding)

Several types of people can be masked (or blinded) in a clinical trial. Here we focus on participants, carers and trial personnel (including healthcare providers). Lack of masking of participants or healthcare providers could bias the results by affecting the *actual* outcomes of the participants in the trial. This may be due to a lack of expectation of efficacy in a control group, or behaviours that differ between intervention groups (for example, differential drop-out, differential cross-over to an alternative intervention, or differential administration of co-interventions).

Masking of outcome assessors is considered in the "Bias in Measurement of Outcomes" domain.

Masking can be impossible in some contexts, for example in a trial comparing a surgical with a non-surgical intervention. 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 of differential behaviours by patients and healthcare providers. When interest is in the effect of assignment to intervention, absence of masking need not lead to bias, providing that all deviations from the intended intervention reflect the care that would routinely be received outside of the context of the trial.

An attempt to mask participants, carers and personnel to intervention group does not ensure successful blinding in practice. Blinding can be compromised for most interventions. For many blinded drug trials, the side effects of the drugs allow the possible detection of which intervention is being received for some participants, unless the study compares two rather similar interventions, e.g. drugs with similar side effects, or uses an active placebo (39). 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 (40, 41), and some reviews of such reports have been published (40, 42). Evidence of correct guesses exceeding 50% would seem to suggest that blinding may have been broken, but in fact can simply reflect the 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 (43). 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. As a consequence, review authors should consider carefully whether to take any notice of the findings of such an exercise.

Study reports often describe masking in broad terms, such as "double blind". This term makes it impossible to know who was blinded (25). Such terms are also used very inconsistently (44-46), and the frequency of explicit reporting of the masking status of study participants and personnel remains low even in trials published in top journals (47), despite recommendations in the CONSORT Statement to be explicit (48). A review of methods used for masking highlights the variety of methods used in practice (39).

Empirical evidence of bias due to deviations from the intended interventions largely comes from studies exploring whether 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 subjective outcomes was associated with a 23% exaggeration of odds ratio (8). By comparison, there was little evidence of such bias in trials of mortality or other objective outcomes, in a meta-analysis of meta-epidemiological studies (7). 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) (49, 50).

2.3.4 Co-interventions

Relevant co-interventions are the interventions or exposures that individuals might receive with or after starting the intervention of interest, which are related to the intervention received and which are prognostic for the outcome of interest. Efforts to specify likely co-interventions should ideally be made in advance (at protocol writing stage). They are likely to be identified through the expert knowledge of members of the review group, via initial (scoping) reviews of the literature, and after discussions with health professionals.

Box 3. 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). An ITT analysis found no difference between intervention groups in the primary outcome (Short Form-36 bodily pain and physical function scales) at two years. However, two years after randomization 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. The authors performed an "as treated" analysis which, in contrast to ITT, showed advantages for surgery. However, this introduced bias because participants crossing over to the surgical group had higher levels of baseline disability and pain than other trial participants (51).

Example 2: substantial numbers of patients not treated as randomized

To determine the efficacy of percutaneous coronary interventions, the FAME 2 trial (Fractional Flow Reserve versus Angiography for Multivessel Evaluation 2 trial) randomised patients with stable, functionally significant coronary artery disease to a percutaneous coronary intervention with implantation of drug-eluting stents and optimal medical therapy or to medical therapy alone. The trial was prematurely stopped because of overwhelming evidence of benefit on the primary composite outcome of death from any cause, nonfatal myocardial infarction, or urgent revascularization in the ITT analysis. There was a pronounced, statistically significant 77% relative risk reduction in urgent revascularization, which drove the difference between groups in the primary outcome. In contrast, there was only a 21% relative risk reduction in the composite of death or myocardial infarction. However, since 41% of patients allocated to optimal medical therapy had crossed over to percutaneous coronary intervention it was likely that the ITT analysis of the composite of death or myocardial infarction was biased towards the null (52).

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. 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. Therefore the study result was at risk of bias, because the co-interventions were not balanced between intervention groups and were likely to impact on the outcome (53).

2.3.5 Using the "Bias due to deviations from intended interventions" domain of the RoB 2.0 tool

Risk of bias assessments for bias due to deviations from the intended interventions are different depending on whether interest focusses on the effect of assignment to intervention or the effect of starting and adhering to intervention.

2.3.5.1 *The effect of assignment to intervention)*

When considering the risk of bias from deviations from intended intervention it is important to consider specifically:

1. who was and was not blinded; and
2. the impact of any deviations from intended intervention that do not reflect usual practice.

Risk of bias may be higher for some outcomes than for others, even if the same people were aware of intervention assignments during the trial. For example, knowledge of the assigned intervention may impact on behavioural outcomes (such as number of clinic visits), while not impacting on physiological outcomes or mortality.

Signalling questions for this domain are provided in Box 4. 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.

2.3.5.2 *Risk of bias judgement (effect of starting and adhering to intervention)*

When considering the risk of bias from deviations from intended intervention it is important to consider specifically:

1. who was and was not blinded; and
2. the impact of any deviations from the protocol intervention in relation to implementation of the intervention, adherence and co-interventions.

Signalling questions for this domain are provided in Box 5. 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 4. The RoB 2.0 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, so risk of bias will be higher. Masking 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 trial personnel 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. Masking carers and trial personnel, which is most commonly achieved through use of a placebo, may prevent such differences.	Y / PY / PN / N / NI
2.3. If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention beyond what would be expected in usual practice?	When interest focusses on the effect of assignment to intervention, it is important to distinguish between: (a) deviations that happen in usual practice following the intervention and so are part of the intended intervention (for example, cessation of a drug intervention because of acute toxicity); and (b) deviations from intended intervention that arise due to expectations of a difference between intervention and comparator (for example because participants feel "unlucky" to have been assigned to the comparator group and therefore seek the active intervention, or components of it, or other interventions). We use the term "usual practice" to refer to the usual course of events in a non-trial context. Because deviations that arise due to expectations of a difference between intervention and comparator are not part of usual practice, they may lead to biased effect estimates that do not reflect what would happen to participants assigned to the interventions in practice. Trialists do not always report (and do not necessarily know) whether deviations that are not part of usual practice actually occurred. Therefore the answer "No information" may be appropriate. However, if such deviations *probably* occurred you should answer "Probably yes".	NA / Y / PY / PN / N / NI
2.4. If Y/PY to 2.3: Were these deviations from intended intervention unbalanced between groups *and* likely to have affected the outcome?	Deviations from intended interventions that do not reflect usual practice will be important if they affect the outcome, but not otherwise. Furthermore, bias will arise only if there is imbalance in the deviations across the two groups.	NA / Y / PY / PN / N / NI
2.5 Were any participants analysed in a group different from the one to which they were assigned?	This question addresses one of the fundamental aspects of an "intention-to-treat" approach to the trial analysis: that participants are analysed in the groups to which they were assigned through randomization. If some participants did not receive their assigned intervention, and such participants were analysed according to intervention received, then the balance between intervention groups created by randomization is lost.	Y / PY / PN / N / NI

2.6 If Y/PY/NI to 2.5: Was there potential for a substantial impact (on the estimated effect of intervention) of analysing participants in the wrong group?	Risk of bias will be high in a randomized trial in which sufficiently many participants were analysed in the wrong intervention group that there could have been a substantial impact on the results. There is potential for a substantial impact if more than 5% of participants were analysed in the wrong group, but for rare events there could be an impact for a smaller proportion.	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*)

Low risk of bias:	Participants, carers and personnel were unaware of intervention groups during the trial. OR Participants, carers or personnel were aware of intervention groups during the trial but any deviations from intended intervention reflected usual practice. OR Participants, carers or personnel were aware of intervention groups during the trial but any deviations from intended intervention were unlikely to impact on the outcome. AND No participants were analysed in the wrong intervention groups (that is, on the basis of intervention actually received rather than of randomized allocation)
Some concerns:	Participants, carers or personnel were aware of intervention groups and there is no information on whether there were deviations from usual practice that were likely to impact on the outcome and were imbalanced between intervention groups. OR Some participants were analysed in the wrong intervention groups (on the basis of intervention actually received rather than of randomized allocation) but there was little potential for a substantial impact on the estimated effect of intervention.
High risk of bias:	Participants, carers or personnel were aware of intervention groups and there were deviations from intended interventions that were unbalanced between the intervention groups and likely to have affected the outcome. OR Some participants were analysed in the wrong intervention groups (on the basis of intervention actually received rather than of randomized allocation), and there was potential for a substantial impact on the estimated effect of intervention.

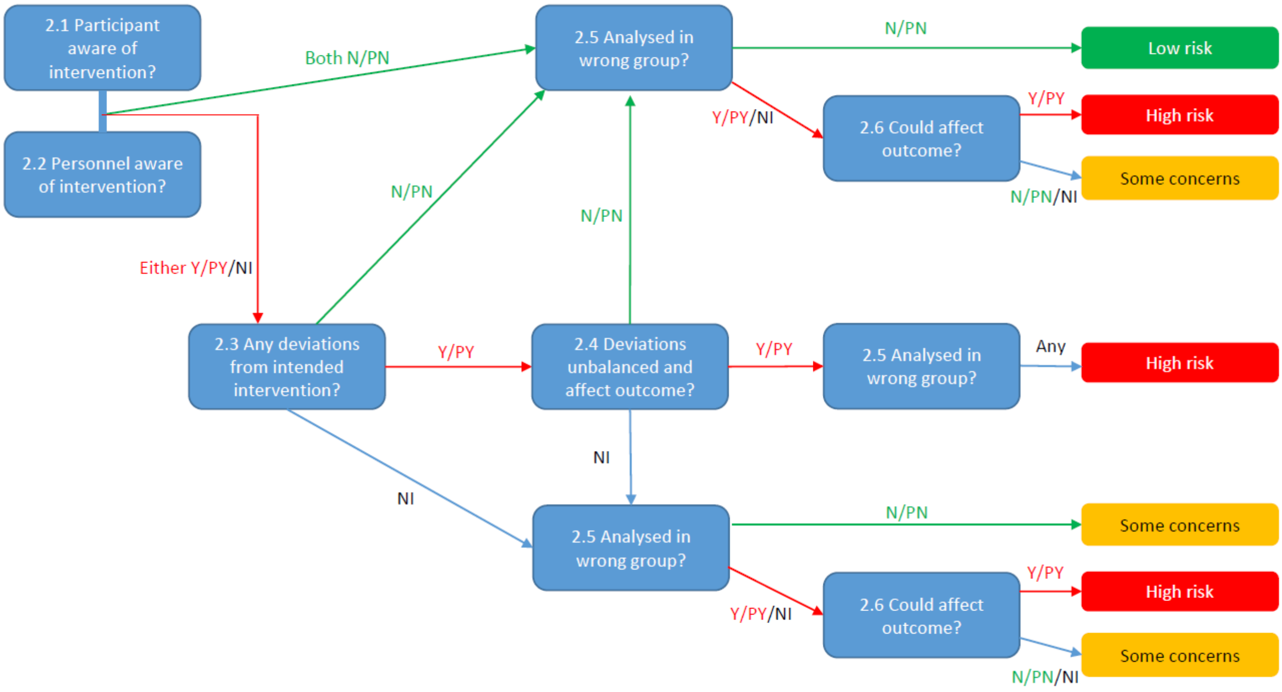
Table 6. Suggested mapping of signalling questions to risk of bias judgements for bias due to deviations from intended interventions (*effect of assignment to intervention*). This is only a suggested decision tree: all default judgements can be overridden by assessors.

Signalling question						Domain level judgement
2.1 **Patient aware?**	**2.2** **Personnel aware?**	**2.3** **Any deviations?**	**2.4** **Unbalanced deviations?**	**2.5** **Wrong group?**	**2.6** **Affect outcome?**	**Default risk of bias**
Both 2.1 & 2.2 N/PN		NA	NA	N/PN	NA	Low
Both 2.1 & 2.2 N/PN		NA	NA	Y/PY/NI	Y/PY	High
Both 2.1 & 2.2 N/PN		NA	NA	Y/PY/NI	N/PN/NI	Some concerns

Either 2.1 or 2.2 Y/PY/NI	Y/PY	N/PN	N/PN	NA	Low
Either 2.1 or 2.2 Y/PY/NI	Y/PY	N/PN	Y/PY/NI	Y/PY	High
Either 2.1 or 2.2 Y/PY/NI	Y/PY	N/PN	Y/PY/NI	N/PN/NI	Some concerns
Either 2.1 or 2.2 Y/PY/NI	Y/PY	Y/PY	Any response	Any response	High
Either 2.1 or 2.2 Y/PY/NI	Y/PY	NI	N/PN	NA	Some concerns
Either 2.1 or 2.2 Y/PY/NI	Y/PY	NI	Y/PY/NI	Y/PY	High
Either 2.1 or 2.2 Y/PY/NI	Y/PY	NI	Y/PY/NI	N/PN/NI	Some concerns
Either 2.1 or 2.2 Y/PY/NI	N/PN	NA	N/PN	NA	Low
Either 2.1 or 2.2 Y/PY/NI	N/PN	NA	Y/PY/NI	Y/PY	High
Either 2.1 or 2.2 Y/PY/NI	N/PN	NA	Y/PY/NI	N/PN/NI	Some concerns
Either 2.1 or 2.2 Y/PY/NI	NI	NA	N/PN	NA	Some concerns
Either 2.1 or 2.2 Y/PY/NI	NI	NA	Y/PY/NI	Y/PY	High
Either 2.1 or 2.2 Y/PY/NI	NI	NA	Y/PY/NI	N/PN/NI	Some concerns

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*). This is only a suggested decision tree: all default judgements can be overridden by assessors.


2.1 Participant aware of intervention?
2.2 Personnel aware of intervention?
Both N/PN
2.5 Analysed in wrong group?
N/PN
Low risk
Y/PY/NI
2.6 Could affect outcome?
Y/PY
High risk
N/PN/NI
Some concerns
N/PN
N/PN
Either Y/PY/NI
2.3 Any deviations from intended intervention?
Y/PY
2.4 Deviations unbalanced and affect outcome?
Y/PY
2.5 Analysed in wrong group?
Any
High risk
NI
NI
2.5 Analysed in wrong group?
N/PN
Some concerns
Y/PY/NI
2.6 Could affect outcome?
Y/PY
High risk
N/PN/NI
Some concerns

Box 5. The RoB 2.0 tool (part 4): Risk of bias due to deviations from the intended interventions (*effect of starting and 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, so risk of bias will be higher. Masking participants, which is most commonly achieved through use of a placebo, may prevent such differences.	Y / PY / PN / N / NI
2.2. Were carers and trial personnel aware of participants' assigned intervention during the trial?	If those involved in caring for participants and those otherwise involved in the trial are aware of group assignment, then it is more likely that implementation of the intended intervention, or the administration of additional co-interventions, will differ between the intervention groups. Masking carers and trial personnel, which is most commonly achieved through use of a placebo, may prevent such differences.	Y / PY / PN / N / NI
2.3. If Y/PY/NI to 2.1 or 2.2: Were important co-interventions balanced across intervention groups?	Risk of bias will be higher if unplanned co-interventions were implemented in a way that would bias the estimated effect of intervention. Co-interventions will be important if they affect the outcome, but not otherwise. Bias will arise only if there is imbalance in such co-interventions between the intervention groups. Consider the co-interventions, including any pre-specified co-interventions, that are likely to affect the outcome and to have been administered in this study. Consider whether these co-interventions are balanced between intervention groups.	NA / Y / PY / PN / N / NI
2.4. Was the intervention implemented successfully?	Risk of bias will be higher if the intervention was not implemented as intended by, for example, the health care professionals delivering care during the trial. Consider whether 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?	Risk of bias will be higher if participants did not adhere to the intervention as intended. Lack of adherence includes imperfect compliance, 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 this proportion is high enough to raise concerns. Answer "Yes" for studies of interventions that are administered once, so that imperfect adherence is not possible.	Y / PY / PN / N / NI
2.6. If N/PN/NI to 2.3, 2.4 or 2.5: Was an appropriate analysis used to estimate the effect of starting and adhering to the intervention?	It is possible to conduct an analysis that corrects for some types of deviation from the intended intervention. Examples of appropriate analysis strategies include inverse probability weighting or instrumental variable estimation. It is possible that a paper reports such an analysis without reporting information on the deviations from intended intervention, but it would be hard to judge such an analysis to be appropriate in the absence of such information. If everyone in one group received a co-intervention, adjustments cannot be made to overcome this. Some examples of analysis strategies that would not be appropriate to estimate the effect of intended intervention are (i) "ITT analysis", (ii) "per protocol analysis", and (iii) "analysis by treatment received".	NA / Y / PY / PN / N / NI
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 starting and adhering to intervention*)

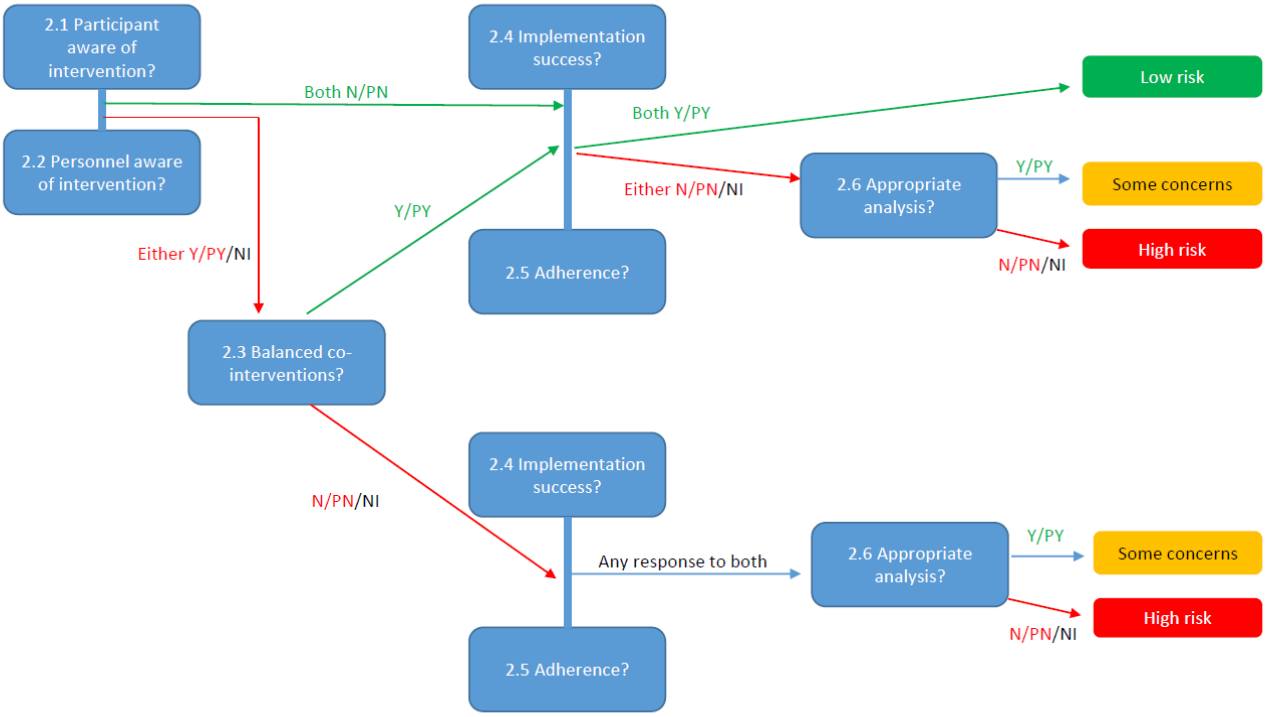
Low risk of bias:	Participants, carers and personnel were **unaware** of intervention groups during the trial and there were no deviations from the intended interventions (in terms of **implementation or adherence**) that were likely to impact on the outcome; OR Participants, carers or personnel were **aware** of intervention groups but the important **co-interventions** were balanced across intervention groups and there were no deviations from the intended interventions (in terms of **implementation or adherence**) that were likely to impact on the outcome.
Some concerns:	There were deviations from the intended interventions (in terms of **implementation and/or adherence**) that were likely to impact on the outcome; OR No information on deviations from intervention was reported for one or both of **implementation or adherence;** OR Participants, carers or personnel were **aware** of intervention groups and the important **co-interventions** were not balanced across intervention groups; OR Participants, carers or personnel were **aware** of intervention groups and no information on **co-interventions** was reported; AND The analysis was **appropriate** to estimate the effect of starting and adhering to intervention, allowing for deviations (in terms of implementation, adherence and co-intervention) that were likely to impact on the outcome.
High risk of bias:	There were deviations from the intended interventions (in terms of **implementation and/or adherence**) that were likely to impact on the outcome; OR No information on deviations from intervention was reported for one or both of **implementation or adherence**. OR Participants, carers or personnel were **aware** of intervention groups and the important **co-interventions** were not balanced between intervention groups; OR Participants, carers or personnel were **aware** of intervention groups and no information on **co-interventions** was reported; AND The analysis was **not appropriate** to estimate the effect of initating and adhering to intervention.

Table 8. Suggested mapping of signalling questions to risk of bias judgements for bias due to deviations from intended interventions (*effect starting and 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	
Patient aware?	Personnel aware?	Balanced co-interventions?	Implement success?	Adherence?	Appropriate analysis?	Default risk of bias
Both 2.1 & 2.2 N/PN		NA	Both 2.4 & 2.5 Y/PY		NA	Low
Both 2.1 & 2.2 N/PN		NA	Either 2.4 or 2.5 N/PN/NI		Y/PY	Some concerns
Both 2.1 & 2.2 N/PN		NA	Either 2.4 or 2.5 N/PN/NI		N/PN/NI	High
Either 2.1 or 2.2 Y/PY/NI		Y/PY	Both 2.4 & 2.5 Y/PY		NA	Low
Either 2.1 or 2.2 Y/PY/NI		Y/PY	Either 2.4 or 2.5 N/PN/NI		Y/PY	Some concerns
Either 2.1 or 2.2 Y/PY/NI		Y/PY	Either 2.4 or 2.5 N/PN/NI		N/PN/NI	High
Either 2.1 or 2.2 Y/PY/NI		N/PN/NI	Any response		Y/PY	Some concerns
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 starting and adhering to intervention*). This is only a suggested decision tree: all default judgements can be overridden by assessors.


2.1 Participant aware of intervention?
2.2 Personnel aware of intervention?
Both N/PN
Either Y/PY/NI
2.4 Implementation success?
2.5 Adherence?
Both Y/PY
Y/PY
Either N/PN/NI
2.6 Appropriate analysis?
Y/PY
N/PN/NI
Low risk
Some concerns
High risk
2.3 Balanced co-interventions?
N/PN/NI
2.4 Implementation success?
2.5 Adherence?
Any response to both
2.6 Appropriate analysis?
Y/PY
N/PN/NI
Some concerns
High risk

2.4 Bias due to missing outcome data

2.4.1 *Rationale for concern about bias*

Randomization provides a fair comparison between two or more intervention groups by balancing out, on average, the distribution of known and unknown prognostic factors among the participants. When measurements of the outcome are missing, for example due to dropout during the study or exclusions from the analysis, the benefit provided by randomization is jeopardized, raising the possibility that the observed effect estimate is biased.

Missing outcome data may occur for the following reasons (54):

- Participants withdraw from the study or cannot be located (lost to follow-up).
- 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.

In addition, some participants may be excluded from analysis for the following reasons:

- Some enrolled participants were later found to be ineligible.
- A "per protocol" analysis is performed (in which participants are included only if they received the intended intervention in accordance with the protocol).
- The study analysis excluded some participants for other reasons.

2.4.1.1 *Missing outcome data and types of analysis*

An intention-to-treat (ITT) analysis is often recommended as the least biased way to estimate intervention effects in randomized trials (55). More precisely, the intention-to-treat analysis provides an unbiased estimate of the effect of assigning participants to an intervention (i.e. the intention-to-treat effect), which is not the same as the effect of participants starting and adhering to intervention (section 1.1). The principles of intention-to-treat analyses are (56, 57):

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

The first principle can always be applied. However, the second is often impossible due to attrition beyond the control of the trialists. Consequently, the third principle of conducting an analysis that includes all participants can only be followed by making assumptions about the missing values. Thus very few trials can perform a true intention-to-treat analysis, especially when there is extended follow-up. In practice, study authors may describe an analysis as intention-to-treat even when some outcome data are missing. The term "intention-to-treat" does not have a clear and consistent definition, and it is used inconsistently in study reports (58-60). Review authors should use the term only to imply all three of the principles above, and should interpret with care any studies that use the term without clarification.

Review authors may also encounter analyses described as "modified intention-to-treat", which usually means that participants were excluded if they did not receive a specified minimum amount of the intended intervention. This term is also used in a variety of ways so review authors should always seek information about precisely who was included (61, 62).

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"), if the reasons for exclusions are considered inappropriate and the data are available to the review author. Review authors are encouraged to do this when possible.

2.4.1.2 *Empirical evidence of bias due to missing outcome data*

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 (8, 16, 23, 63-66) (10-14). In a systematic review of meta-epidemiological studies (7), missing data was associated with overestimation of effect estimates in some studies, but underestimation or no difference in others. For example, reporting the use of a "modified" intention-to-treat analysis (versus intention-to-treat) was associated with exaggerated effects (66), while having a dropout rate >20% (versus ≤20%) was not (8). There are notable examples of biased "per-protocol" analyses, however (67), and a review has found more exaggerated effect estimates from "per-protocol" analyses compared with "intention-to-treat" analyses of the same trials (68). 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 (69).

Per-protocol analyses will generally produce larger effect estimates than intention-to-treat analyses, and are biased if interest is in the effect of assignment to intervention. If interest focusses on the effect of starting and adhering to intervention, then both the intention-to-treat estimate and naïve per protocol estimate may under or overestimate the true effect, depending on the reasons for the missing data (37, 38, 69).

Interpretation of empirical studies is difficult because exclusions are often poorly reported, particularly before 1996 in the pre-CONSORT era. 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 (23). 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.

Empirical research also has investigated the adequacy with which incomplete outcome data are addressed in reports of trials. One 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 (70).

2.4.2 Assessing risk of bias from missing outcome data

The risk of bias arising from missing outcome data depends on several factors, including the amount and distribution across intervention groups, the reasons for outcomes being missing, the likely difference in outcome between participant with and without data, what study investigators have done to address the problem in their analyses, and the context. Therefore it is not possible to formulate a simple rule for judging a study to be at low or high risk of bias.

The following considerations should help review authors assess whether missing outcome data may put an outcome at risk for bias.

2.4.3 Amount of missing outcome data

A threshold for what constitutes "high" or "modest" degree of missing data is to a considerable extent arbitrary, and is clearly inappropriate if interpreted as implying a strict boundary. There is a tradition for regarding missing less than 5% of the outcome data as "low", and missing over 15-20% as "high". The potential impact of missing data on estimated intervention effects depends not only on the proportion of participants with missing data, but also on the type of outcome, reasons for missing data, and the difference in prognosis between participants with and without missing data. When determining the proportion of and the reasons for missing data, some random variation between groups is expected in an unbiased scenario.

For continuous outcomes, it is unlikely that notable bias will result from missing less than 5% of outcome data.

For dichotomous outcomes, the proportion required is directly linked to 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 there is only 10% missing data, if the mortality rate in the 100 missing participants is 20% (20 participants), the overall true mortality of the intervention group would be nearly double (3.8% vs. 2%) of that estimated from the observed data.

For time-to-event outcomes, missing data occur in participants who prematurely discontinue follow-up prior to the occurrence of the event or at the end of the planned follow-up period. For both types of missing data, participants who did not experience the event of interest are considered to be "censored" on the date of their last follow-up. The important consideration is whether the intervention effect in individuals who were censored is the same as the effect in other individuals. Often censoring is assumed to be a random event, which is likely to be the case when censoring is caused by an administrative decision (such as closing down the trial) but is less likely when caused by patient drop out during the trial.

In case there is "No information" available, this is a cause of some concern because missing data prevention is a core aspect of trial methodology and some degree of reporting on missing data is to be expected.

2.4.4 Reasons for missing outcome data

An understanding of the reasons for missing outcome data is critical for judging the risk of bias. A difference in the proportions of incomplete outcome data across groups is of concern if the availability of outcome data is determined by the participants' true outcomes. For example, if participants with poorer clinical outcomes are more likely to drop out due to adverse effects, and this happens mainly in the experimental group, then the effect estimate will be biased in favour of the experimental intervention. Exclusion of participants due to "inefficacy" or "failure to improve" will introduce bias if the numbers excluded are not balanced across intervention groups. Note that a non-significant result of a statistical test for differential missingness does not confirm the absence of bias, especially in small studies that have low power.

Even if incomplete outcome data are balanced in numbers across groups, bias can be introduced if the reasons for missing outcomes differ. For example, in a trial of an experimental intervention aimed at smoking cessation it is feasible that a proportion of the control intervention participants could leave the study due to a lack of enthusiasm at receiving nothing novel (and continue to smoke), and that a similar proportion of the experimental intervention group could leave the study due to successful cessation of smoking.

2.4.5 Statistical methods for handling missing outcome data

Statistical models for handling missing data involve analyses conducted by trial authors (for example, multiple imputation) or review authors (for example in a trial with binary outcomes or if reviewers have access to individual patient data). There are three common approaches for the analysis of missing data (54, 71-73): (1) complete cases analysis (i.e., ignore and discard incomplete observations with missing data); (2) imputation (i.e., fill in the missing values and then analyse the filled-in data); and (3) analysing the incomplete data by a method that does not require a complete data set (i.e. likelihood-based methods, moment-based methods, and semiparametric models for survival data) (54, 74).

Complete case analysis is problematic because the estimates produced could be biased, depending on the missing data mechanism.

For imputation, the underlying idea is that the missing value is replaced by one or more new values and the filled-in dataset is then analysed. In single imputation, only one estimate is filled in. This is often done using regression techniques to draw from a predictive distribution, last observation carried forward (LOCF) or baseline observation carried forward (BOCF). Note that LOCF and BOCF may be problematic depending on the trend of changing outcomes among participants (trend of improvement or trend of deterioration) and generally improve precision artificially. We recommend that reviewers interpret LOCF and BOCF analyses with considerable caution (75, 76).

In multiple imputation, multiple estimates are drawn from a predictive distribution of this outcome variable, forming multiple distinct filled-in datasets (77, 78). These multiple datasets are then analysed for summary measures and variances to reflect the uncertainty associated with missing data that are not reflected with simple regression imputation methods (see Box 3 on statistical methods).

2.4.6 Sensitivity analysis

Sensitivity analyses can be performed to assess the assumption made about the missing data mechanism. There is a tradition for "worst case" and "best case" analyses clarifying the extreme boundaries of what is theoretically possible, but such analyses may not be informative for what is the most probable scenario (75). We recommend that any sensitivity analyses should focus on realistic assumptions, and if results differ little between such analyses, the result may be considered "robust".

In one method, reviewers might present results based on the assumptions at the "boundary level", i.e. where the treatment no longer becomes beneficial. For example, treatment might not be considered beneficial if 25% of the treated group died and 10% of the control group died. If so, there is no need to present the results of the traditional worst-case scenario where 100% of the treated group dies and 0% of the control group dies. Once the boundary results are presented, readers would then know that assumptions that were more extreme would lead to even more extreme results.

2.4.7 Using the "Bias due to missing outcome data" domain of the RoB 2.0 tool

Signalling questions for this domain are provided in Box 6. 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 6. The RoB 2.0 tool (part 5): Risk of bias due to missing outcome data

Signalling questions	Elaboration	Response options
3.1 Were outcome data available for all, or nearly all, participants randomized?	The appropriate study population for an analysis of the intention to treat effect is all randomized patients. Note that imputed data should be regarded as missing data, and not considered as "outcome data" in the context of this question. "Nearly all" (equivalently, a low or modest amount of missing data) should be interpreted as "enough to be confident of the findings", and a suitable proportion depends on the context. 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: Are the proportions of missing outcome data and reasons for missing outcome data similar across intervention groups?	"Similar" (with regard to proportion and reasons for missing outcome data) includes some minor degree of discrepancy across intervention groups as expected by chance. Assessment of comparability of reasons for missingness requires the reasons to be reported.	NA / Y / PY / PN / N / NI
3.3 If N/PN/NI to 3.1: Is there evidence that results were robust to the presence of missing outcome data?	Evidence for robustness may come from how missing data were handled in the analysis and whether sensitivity analyses were performed by the trial investigators, or from additional analyses performed by the systematic reviewers.	NA / Y / PY / PN / N / NI
Risk of bias judgement	See Table 11, Table 12 and Figure 4.**Error! Reference source not found.**	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

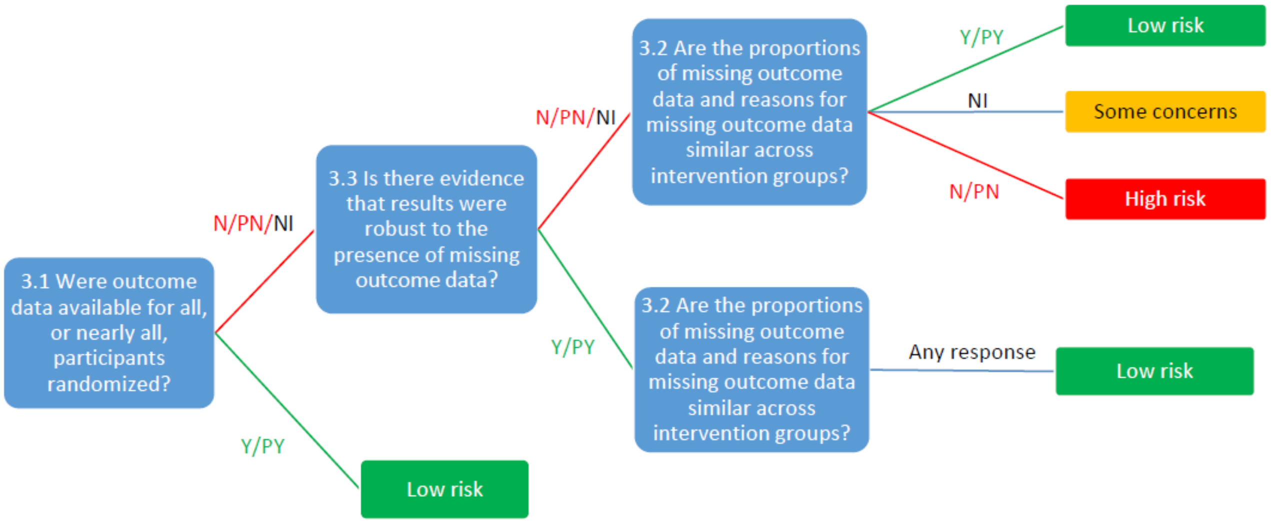
Low risk of bias:	No missing data OR Non-differential missing data (similar proportion of and similar reasons for missing data in compared groups) OR Evidence of robustness of effect estimate to missing data (based on adequate statistical methods for handling missing data and sensitivity analysis).
Some concerns:	An unclear degree of missing data or unclear information on proportion and reasons for missingness in compared groups AND There is not evidence that the effect estimate is robust to missing data.
High risk of bias:	A high degree of missing data AND Differential missing data (different proportion of or different reasons for missing data in compared groups) AND There is not evidence that the effect estimate is robust to missing data.

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**	**Default risk of bias**
Y/PY	NA	NA	Low
N/PN/NI	Any response	Y/PY	Low
N/PN/NI	Y/PY	N/PN/NI	Low
N/PN/NI	N/PN	N/PN/NI	High
N/PN/NI	NI	N/PN/NI	Some concerns

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.


3.1 Were outcome data available for all, or nearly all, participants randomized?
N/PN/NI
Y/PY
Low risk
3.3 Is there evidence that results were robust to the presence of missing outcome data?
N/PN/NI
Y/PY
3.2 Are the proportions of missing outcome data and reasons for missing outcome data similar across intervention groups?
Y/PY
Low risk
NI
Some concerns
N/PN
High risk
3.2 Are the proportions of missing outcome data and reasons for missing outcome data similar across intervention groups?
Any response
Low risk

2.5 Bias in measurement of the outcome

2.5.1 Background

Several types of people can be blinded in a clinical trial. The second of two domains in the tool that specifically address blinding focuses on blinding of outcome assessors. If people who assess outcomes are aware of intervention assignments, bias could be introduced into these assessments.

In empirical studies, 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 (79). The estimated effect has been observed to be more biased, on average, in trials with subjective outcomes (7).

Before assessing this domain, it is important to determine:

a) **Who is the outcome assessor and whether the outcome assessor is blinded**. The outcome assessor can be
   - The participant when the outcome is a participant reported outcome (PRO) such pain, quality of life, self-completed questionnaire evaluating depression, anxiety, function etc.
   - The intervention provider when the outcome is the result of a clinical exam, the occurrence of a clinical event or a therapeutic decision such as decision to perform a surgical intervention or decision to discharge the patient.
   - The outcome assessor can be an observer not directly involved in the intervention provided to the participant such as for example an adjudication committee, a biologist performing an automatized complementary test, a physician evaluating complementary test.

   Blinding of outcome assessors can sometimes be impossible. For example, in a trial comparing surgery to usual care on pain level at 3 months, it is impossible to blind the outcome assessor (i.e. the participant). However, this does not mean that potential biases can be ignored, and review authors should still assess the risk of bias due to lack of blinding of outcome assessment for all studies in their review.

b) **Whether the assessment of outcome is likely to be influenced by knowledge of intervention received**. This will depend on whether the assessment of outcome will involve some judgement or not which depends on the type of outcome. We can have 5 different type of outcomes:
   - **<u>Participant-reported outcomes (PRO)</u>**
     - <u>Definition:</u> 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.
       PROs 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 (80).
     - <u>Examples:</u> pain, nausea, health related quality of life etc.
     - <u>The outcome assessor</u> is **the participant** even if a blinded interviewer is questioning the participant and is completing a questionnaire. The interviewer is <u>not considered</u> the outcome assessor in a strict sense but rather a "facilitator".
     - <u>The assessment of outcome</u> is usually **likely to be influenced** by knowledge of intervention received.

   - **<u>Observer-reported outcomes not involving judgement</u>**
     - <u>Definition:</u> The outcome is reported by an external observer (e.g., intervention provider, independent researcher, physician not involved in the care provided to participants (radiologist) but it does not involve any judgement from the observer.
     - <u>Examples:</u> all-cause mortality; automatized complementary test (e.g., CRP and other automated non-repeatable laboratory tests).
     - <u>The outcome assessor</u> is **the observer**.
     - <u>The assessment of outcome</u> is usually **not likely to be influenced** by knowledge of intervention received.

- **Observer-reported outcomes involving some judgment**
    - Definition: The outcome is reported by an external observer (e.g., intervention provider) and involves some judgement (e.g., clinical exam)
    - Examples: complementary test involving some judgement (e.g., assessment of radiographs); clinical examination, clinical event other than death (e.g., myocardial infarction).
    - The outcome assessor is **the observer**.
    - The assessment of outcome is usually **likely to be influenced** by knowledge of intervention received.

- **Intervention provider decision outcomes**
    - Definition: The outcome is a decision made by the intervention provider. The recording of this decision does not involve any judgement. However, the decision itself can be highly 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 evolution of the participants not by the knowledge of intervention received.
    - Examples: hospitalization, stop treatment, referral to a different ward, performing a caesarean section, stop ventilation, discharge the participant.
    - The outcome assessor is **the care provider making the decision**.
    - The assessment of outcome is usually **likely to be influenced** by knowledge of intervention received. This is particularly important when the preference, expectations, hunches regarding the beneficial effect of the experimental intervention is very high.

- **Composite outcomes**
    - Definition: A composite outcome consists of combining multiple end points into a single outcome. Participants who have experienced any of the endpoint specified are considered to have experienced the composite outcome.
    - Example: Major adverse cardiac and cerebrovascular events (MACCE)
    - The assessment of outcome: Assessment of composite outcomes should take into account the frequency of each component of the composite outcome and take into account the **risk of bias of the most frequent components**.

c) What factors could influence the risk of bias when the outcome assessor is not blinded and the outcome is likely to be influenced by knowledge of intervention received

To assess the risk of bias, the reviewer should also take into account **the degree of expectation and vested interest of the outcome assessor** regarding the beneficial effect of the experimental intervention. This expectation and vested interest could vary according to:

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

2.5.2 Using the "Bias in measurement of the outcome" domain of the RoB 2.0 tool

Signalling questions for this domain are provided in Box 7. 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 7. The RoB 2.0 tool (part 6): Risk of bias in measurement of the outcome

Signalling questions	Elaboration	Response options
4.1 Were outcome assessors aware of the intervention received by study participants?	"No" if outcome assessors were blinded to intervention status. In studies where participants report their outcomes themselves (i.e., participant-reported outcome), the outcome assessor is the study participant.	Y / PY / PN / N / NI
4.2 If Y/PY/NI to 4.1: Was the assessment of the outcome likely to be influenced by knowledge of intervention received?	Knowledge of the assigned intervention may impact on participant-reported outcomes (such as level of pain), observer-reported outcomes involving some judgement, and intervention provider decision outcomes, while not impacting on other outcomes such as observer reported outcomes not involving judgement such as all-cause mortality. In many circumstances the assessment of *observer reported outcomes not involving judgement* such as all-cause mortality might be considered to be unbiased, even if outcome assessors were aware of intervention assignments.	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 due to 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

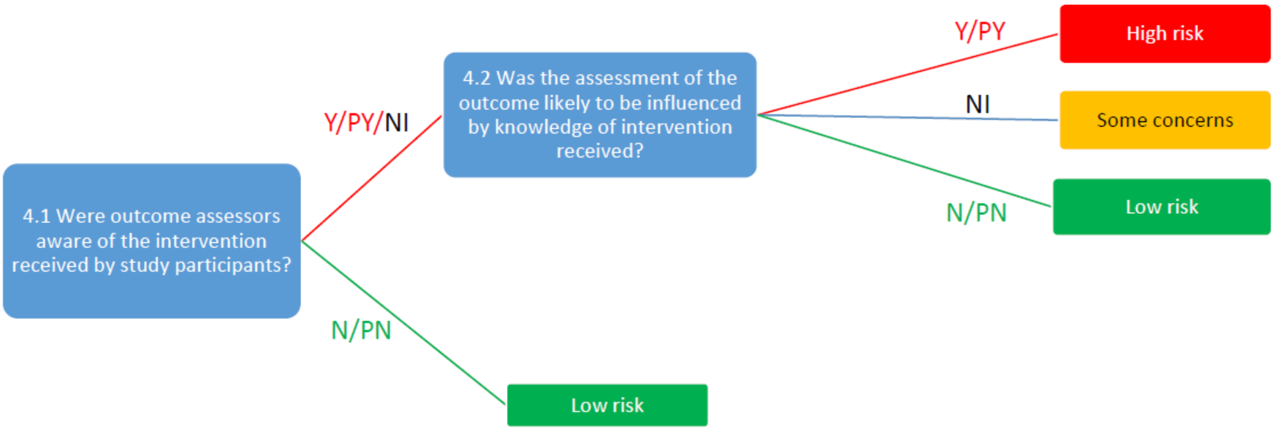
Low risk of bias	The outcome assessors were unaware of the intervention received by study participants. or The outcome assessors were aware of the intervention received by study participants, but the assessment of the outcome was unlikely to be influenced by knowledge of the intervention received.
Some concerns	There is no information available to determine whether the assessment of the outcome is likely to be influenced by knowledge of the intervention received.
High risk of bias	The assessment of the outcome was likely to be influenced by knowledge of the intervention received by study participants.

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**	**Default risk of bias**
N/PN	NA	Low
Y/PY/NI	Y/PY	High risk
Y/PY/NI	N/PN	Low risk
Y/PY/NI	NI	Some concerns

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.


4.1 Were outcome assessors aware of the intervention received by study participants?
Y/PY/NI
N/PN
4.2 Was the assessment of the outcome likely to be influenced by knowledge of intervention received?
Low risk
Y/PY
NI
N/PN
High risk
Some concerns
Low risk

2.6 Bias in selection of the reported result

2.6.1 Background

In this document we define: an **outcome domain** as a true state or endpoint of interest, irrespective of how it is measured (e.g. presence or severity of depression), an **outcome measurement** as a specific measurement made on the study participants (e.g. measurement of depression using the Hamilton rating scale 6 weeks after initiation of intervention) and an **outcome analysis** as 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 intervention and control groups).

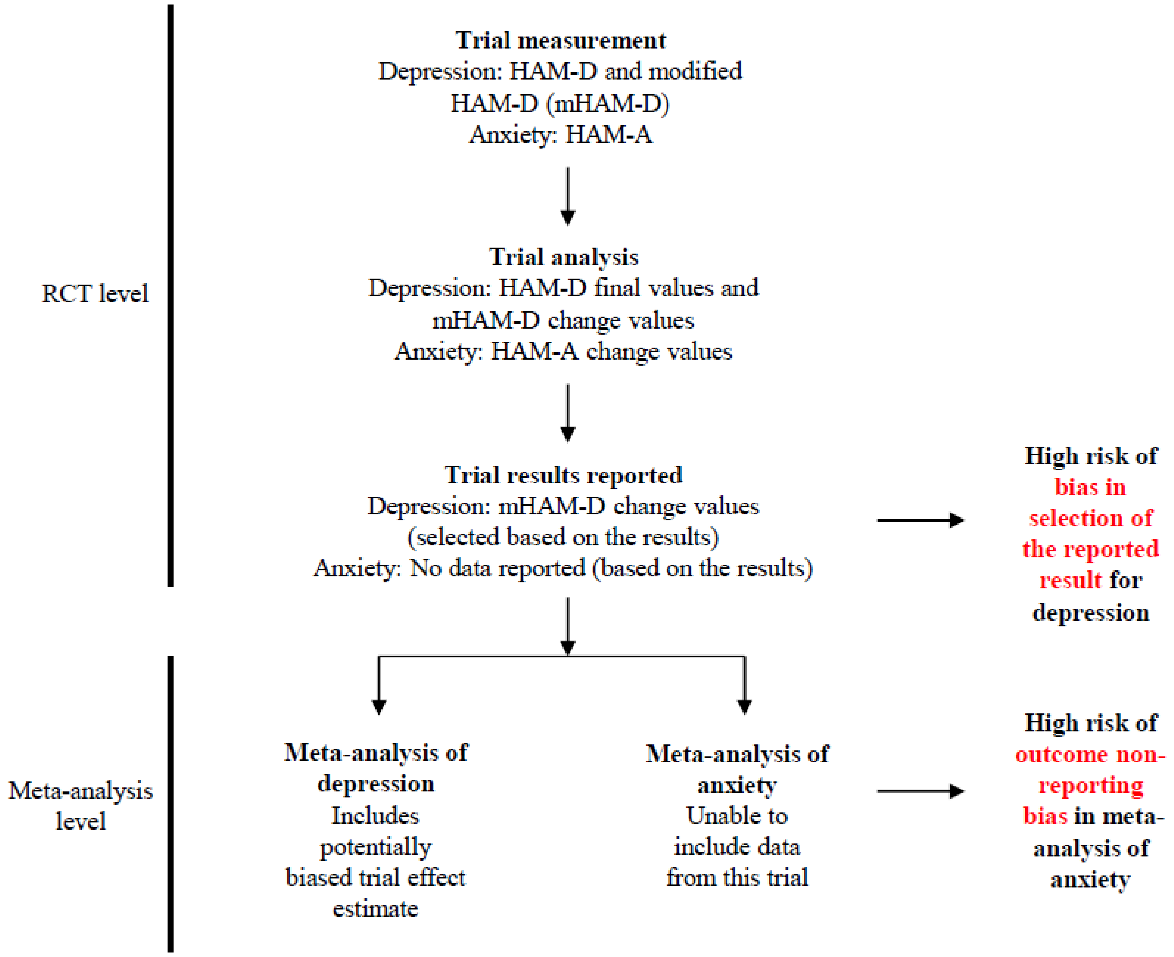
Selective reporting within clinical trials has to date mainly been described with respect to the failure to report, or partial reporting of, outcome domains that were measured and analysed (81). **Outcome non-reporting bias** arises when the outcome domain is not reported or is partially reported based on the direction, magnitude or statistical significance of the estimated intervention effect for the intervention versus the comparator ("effect estimate" or "results"). For example, the outcome mortality is recorded but trialists report no data, or state only that the effect estimate for mortality was not statistically significant. Outcome non-reporting bias in one or more of the studies included in a systematic review can put the intervention effect estimate reported by the systematic review at risk of bias (usually in the direction of exaggeration of the magnitude of effect).

Another type of selective reporting arises when the effect estimate reported for a particular outcome domain has been selected (based on the direction, magnitude or statistical significance) from multiple effect estimates that have been calculated for an outcome domain. We call this **bias in selection of the reported results**. This type of selective reporting puts effect estimates from individual primary studies at risk of bias in the same way as other bias domains considered in the RoB 2.0 tool.

An example of bias in selection of the reported results, and how this differs from outcome non-reporting bias, is shown in Figure 1. 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 reported effect estimate for depression (modified HAM-D change values) was selected by the trialists because it had the most favourable result. This means there is a high risk of bias in the fully reported effect estimate for the depression outcome domain (to be synthesized in the review) because of the primary authors' selection of the reported result (82).

In contrast, anxiety was measured and analysed, but no data were reported because the effect was unfavourable to the intervention. Not reporting anxiety does not, by itself, put other fully reported effect estimates from the trial at risk of bias. However, it is likely to bias the result of a meta-analysis of the effect of the intervention on the anxiety outcome domain (which cannot include the unreported data from this trial), if the reason for not reporting the anxiety effect estimate is a lack of significance. This means there may be a high risk of outcome non-reporting bias in a meta-analysis of anxiety.

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


RCT level
Trial measurement
Depression: HAM-D and modified HAM-D (mHAM-D)
Anxiety: HAM-A
Trial analysis
Depression: HAM-D final values and mHAM-D change values
Anxiety: HAM-A change values
Trial results reported
Depression: mHAM-D change values (selected based on the results)
Anxiety: No data reported (based on the results)
High risk of bias in selection of the reported result for depression
Meta-analysis level
Meta-analysis of depression
Includes potentially biased trial effect estimate
Meta-analysis of anxiety
Unable to include data from this trial
High risk of outcome non-reporting bias in meta-analysis of anxiety

The RoB 2.0 tool considers outcome non-reporting bias as analogous to publication bias (where an entire study is not published depending on the nature and direction of the results). Therefore, this kind of selective outcome reporting should be appraised using a different mechanism, not as part of the RoB 2.0 tool. **This is a notable departure from the previous Cochrane RoB tool for randomized trials**. Therefore, in this document we do not include signalling questions for selective non-reporting or selective partial reporting of effect estimates for outcome domains. These issues will be addressed in a new tool to assess the risk of non-reporting bias in reviews (i.e. bias due to unpublished studies or unpublished/unusable data within studies), which is under development.

Empirical evidence indicates that discrepant reporting of analyses is common in clinical trials. In a systematic review of studies comparing different sections within a trial report, or a trial report with its corresponding protocol, discrepancies were often found in definition of composite outcomes, handling of missing data, unadjusted versus adjusted analyses, handling of continuous data, and subgroup analyses (83). Such discrepancies are suggestive of bias in selection of the reporting result, as discussed in the following sections.

2.6.2 Selective reporting of a result contributing to the synthesis

We consider here the selective reporting of a **fully reported result**, that is, a result that is sufficiently reported to allow an effect estimate (or data for intervention and comparator groups) to be included in a meta-analysis (or other synthesis). This domain combines (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. Such 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 for 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 measurement with and without adjustment for potential confounders (or with adjustment for different sets of potential confounders); multiple analyses of a particular measurement with and without, or with different, methods to take account of missing data; 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 newsworthy, or 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 selectively report 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 publicising the existence of a harm will increase their chances of publishing in a high impact factor journal. Such motivations are not always easy to decipher; for example, in head-to-head trials, trialists may have different preconceptions about the efficacy and safety of the different active interventions.

2.6.3 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 paper, post-hoc.

Analysis intentions may be available 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, yet is currently rarely published. If the researchers' pre-specified intentions are available in sufficient details, then outcome measurements and analyses listed in these documents can be compared with those presented in the published report.

When comparing analysis intentions with the publication, the dates of such documents must be considered carefully. There should be a "date-stamp" confirming that the analysis intentions were finalised before conducting any analyses. If the date of a document such as a trial registry record precedes that of the final publication by, for example, only a couple of months, it is unlikely that the analysis intentions specified in the registry record were truly pre-specified. Amendments or updates to analysis intentions should also be retrieved and compared with the original intentions. Fortunately, these are usually date-stamped in trial registries or journal publications, so review authors can determine when changes were made.

Review authors should 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.

2.6.4 Inferring selective reporting when analysis intentions are unavailable

For many trials, the analysis intentions of the researchers will not be readily available. Where these remain unknown to the review authors, 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 the results that are reported. In addition, outcome measures and analyses can be compared across different results papers about a trial that may be available. In addition, the following questions may help review authors to infer selective reporting:

a. Are subscales aggregated in an unusual manner?

b. Is there a discrepancy between different reports in the designation of the primary and secondary outcomes or specific outcomes?

c. 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?

d. Have subgroups been defined in unusual ways (e.g. dose or dose frequency)?

e. Were analyses with imputation carried out but not reported without justification? Were one or more imputation methods performed but only the results of one or a subset reported?

f. 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?

g. Is there is a discrepancy between different reports in the sample of participants analysed.

h. Has an unusual composite outcome been reported?

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 cutpoints for continuous outcome measures may need to be modified because the distribution of data differed to 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 (83). 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.**

The insufficient detail in some documents may preclude a proper 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 such inadequate pre-specification of the analysis intentions in their responses to signalling questions.

When making a risk of bias judgement, review authors should consider the magnitude, direction and statistical significance of fully reported effect estimates. If there is clear evidence in a placebo-controlled trial that some measure or analysis for a particular efficacy outcome was not reported, but the reported result for that outcome is not statistically significant or is close to the null, then it is less likely that the reported effect estimate was selected based on its results.

2.6.5 Using the "Bias in selection of the reported result" domain of the RoB 2.0 tool

Signalling questions for this domain are provided in Box 8. 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 8. The RoB 2.0 tool (part 7): Risk of bias in selection of the reported result

Signalling questions		Elaboration	Response options
Are the reported outcome data likely to have been selected, on the basis of the results, from...	**5.1. ... 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. **A response of "Yes/Probably yes" is reasonable 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 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 outcome measurements that are favourable to the experimental intervention. **A response of "No/Probably no" is reasonable 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 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. **A response of "No information" is reasonable 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.	Y / PY / PN / N / NI
	5.2 ... 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; conversion of continuously scaled outcome to categorical data with different cut-points; different sets of covariates for adjustment; 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.	Y / PY / PN / N / NI

		A response of “Yes/Probably yes” is reasonable 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. **A response of “No/Probably no” is reasonable 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. **A response of “No information” is reasonable 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 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 repoted result

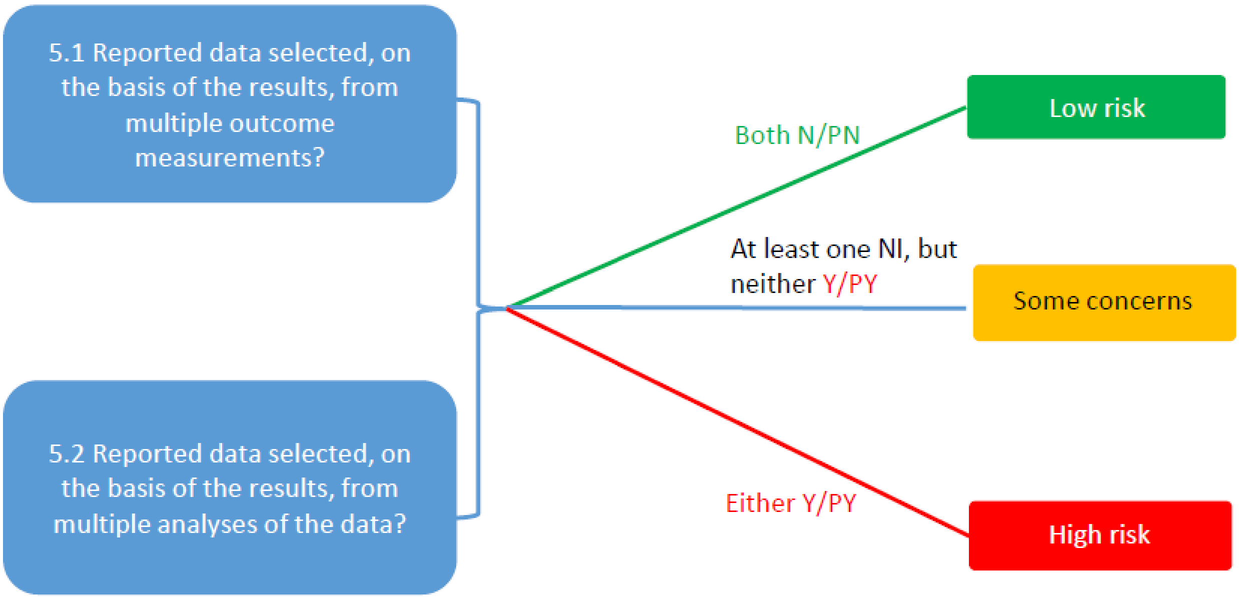
Low risk of bias	Reported outcome data are **unlikely** to have been selected, on the basis of the results, from multiple outcome measurements (e.g. scales, definitions, time points) within the outcome domain, and Reported outcome data are **unlikely** to have been selected, on the basis of the results, from multiple analyses of the data.
Some concerns	There is insufficient information available to exclude the possibility that reported outcome data were selected, on the basis of the results, from multiple outcome measurements (e.g. scales, definitions, time points) within the outcome domain, or from multiple analyses of the data. **Given that analysis intentions are often unavailable or not reported with sufficient detail, we anticipate that this will be the default judgement for most trials.**
High risk of bias	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 from multiple analyses of the data (or both).

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**	**Default risk of bias**
N/PN	N/PN	Low
N/PN	NI	Some concerns
NI	N/PN	Some concerns
NI	NI	Some concerns
N/PN	Y/PY	High
Y/PY	N/PN	High
Y/PY	Y/PY	High
Y/PY	NI	High
NI	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.


5.1 Reported data selected, on the basis of the results, from multiple outcome measurements?
5.2 Reported data selected, on the basis of the results, from multiple analyses of the data?
Both N/PN
Low risk
At least one NI, but neither Y/PY
Some concerns
Either Y/PY
High risk

3 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. Hill AB. Suspended judgment. Memories of the British Streptomycin Trial in Tuberculosis. The first randomized clinical trial. Control Clin Trials. 1990;11(2):77-9.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. Armijo-Olivo S, Saltaji H, da Costa BR, Fuentes J, Ha C, Cummings GG. What is the influence of randomisation sequence generation and allocation concealment on treatment effects of physical therapy trials? A meta-epidemiological study. BMJ open. 2015;5(9):e008562.

10. Bialy L, Vandermeer B, Lacaze-Masmonteil T, Dryden DM, Hartling L. A meta-epidemiological study to examine the association between bias and treatment effects in neonatal trials. Evid Based Child Health. 2014;9(4):1052-9.

11. Chaimani A, Vasiliadis HS, Pandis N, Schmid CH, Welton NJ, Salanti G. Effects of study precision and risk of bias in networks of interventions: a network meta-epidemiological study. Int J Epidemiol. 2013;42(4):1120-31.

12. 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.

13. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352(9128):609-13.

14. Papageorgiou SN, Xavier GM, Cobourne MT. Basic study design influences the results of orthodontic clinical investigations. J Clin Epidemiol. 2015.

15. Herbison P, Hay-Smith J, Gillespie WJ. Different methods of allocation to groups in randomized trials are associated with different levels of bias. A meta-epidemiological study. J Clin Epidemiol. 2011;64(10):1070-5.

16. 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.

17. 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.

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. Schulz KF, Grimes DA. Unequal group sizes in randomised trials: guarding against guessing. Lancet. 2002;359(9310):966-70.

22. 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.

23. 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.

24. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ. 2001;323(7303):42-6.

25. Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. Ann Intern Med. 2002;136(3):254-9.

26. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359(9306):614-8.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. Schulz KF. Subverting randomization in controlled trials. JAMA. 1995;274(18):1456-8.

33. 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.

34. 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.

35. 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.

36. Altman DG, Bland JM. How to randomize. BMJ. 1999;319:703-4.

37. Hernan MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. Clin Trials. 2012;9(1):48-55.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. 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.

47. 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.

48. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340:c332.

49. Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 2014;43(4):1272-83.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. Menerit CL. Clinical Trials – Design, Conduct, and Analysis. Second Edition: Oxford University Press; 2012.

57. Piantadosi S. Clinical Trials: A Methodologic perspective. Second Edition: Wiley; 2005.

58. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319(7211):670-4.

59. 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.

60. 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.

61. Abraha I, Montedori A. Modified intention to treat reporting in randomised controlled trials: systematic review. BMJ. 2010;340:c2697.

62. Vedula SS, Li T, Dickersin K. Differences in reporting of analyses in internal company documents versus published trial reports: comparisons in industry-sponsored trials in off-label uses of gabapentin. PLoS Med. 2013;10(1):e1001378.

63. 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.

64. 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.

65. 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.

66. 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.

67. 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.

68. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. J Clin Epidemiol. 2007;60(7):663-9.

69. Tierney JF, Stewart LA. Investigating patient exclusion bias in meta-analysis. Int J Epidemiol. 2005;34(1):79-87.

70. 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.

71. Fleming TR. Addressing missing data in clinical trials. Ann Intern Med. 2011;154(2):113-7.

72. Little RJA, Rubin DB. Statistical Analysis with Missing Data. Second Edition. Hoboken, NJ: Wiley; 2002.

73. 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.

74. 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.

75. 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.

76. May GS, DeMets DL, Friedman LM, Furberg C, Passamani E. The randomized clinical trial: bias in analysis. Circulation. 1981;64(4):669-73.

77. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: John Wiley & Sons; 1987.

78. 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.

79. 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.

80. 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.

81. 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.

82. 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.

83. 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.

4 Contributors

Core group: Jelena Savović, Julian Higgins, Matthew Page, Asbjørn Hróbjartsson, Isabelle Boutron, Barney Reeves, Roy Elbers, Jonathan Sterne

Allocation bias: Jelena Savović, Christopher Cates, Mark Corbett, Penny Whiting, Julian Higgins

Bias due to deviations from intended interventions: Jonathan Sterne, Ian Shrier, Peter Jüni, Jonathan Emberson, Nicky Cullum, Natalie Blencowe, Roy Elbers, Jelena Savović, Matthew Page, Julian Higgins

Bias due to missing outcome data: Asbjørn Hróbjartsson, Tianjing Li, Ian Shrier, Julian Higgins

Bias in measurement of outcomes: Isabelle Boutron, Asbjørn Hróbjartsson, Sasha Shepperd, Julian Higgins

Bias in selection of the reported result: Barney Reeves, Jamie Kirkham, Matthew Page, Lesley Stewart, Rachel Churchill, Sally Hopewell, Toby Lasserson, Sharea Ijaz, Julian Higgins

Bias in cluster randomized trials: Sandra Eldridge, Mike Campbell, Amy Drahota, Bruno Giraudeau, Jeremy Grimshaw, Barney Reeves, Nandi Siegfried, Julian Higgins

Bias in cross-over randomized trials: Julian Higgins, Doug Altman, Francois Curtin, Tianjing Li, Stephen Senn

Other contributors: Henning Keinke Andersen, Mike Clarke, Jon Deeks, Geraldine MacDonald, Richard Morris, Mona Nasser, Nishith Patel, Jani Ruotsalainen, Holger Schünemann, Jayne Tierney