BMC
Medical Research Methodology

RESEARCH ARTICLE | Open Access

SYRCLE's risk of bias tool for animal studies

Carlijn R Hooijmans[1*], Maroeska M Rovers[2], Rob BM de Vries[1], Marlies Leenaars[1], Merel Ritskes-Hoitinga[1] and Miranda W Langendam[3]

Abstract

Background: Systematic Reviews (SRs) of experimental animal studies are not yet common practice, but awareness of the merits of conducting such SRs is steadily increasing. As animal intervention studies differ from randomized clinical trials (RCT) in many aspects, the methodology for SRs of clinical trials needs to be adapted and optimized for animal intervention studies. The Cochrane Collaboration developed a Risk of Bias (RoB) tool to establish consistency and avoid discrepancies in assessing the methodological quality of RCTs. A similar initiative is warranted in the field of animal experimentation.

Methods: We provide an RoB tool for animal intervention studies (SYRCLE's RoB tool). This tool is based on the Cochrane RoB tool and has been adjusted for aspects of bias that play a specific role in animal intervention studies. To enhance transparency and applicability, we formulated signalling questions to facilitate judgment.

Results: The resulting RoB tool for animal studies contains 10 entries. These entries are related to selection bias, performance bias, detection bias, attrition bias, reporting bias and other biases. Half these items are in agreement with the items in the Cochrane RoB tool. Most of the variations between the two tools are due to differences in design between RCTs and animal studies. Shortcomings in, or unfamiliarity with, specific aspects of experimental design of animal studies compared to clinical studies also play a role.

Conclusions: SYRCLE's RoB tool is an adapted version of the Cochrane RoB tool. Widespread adoption and implementation of this tool will facilitate and improve critical appraisal of evidence from animal studies. This may subsequently enhance the efficiency of translating animal research into clinical practice and increase awareness of the necessity of improving the methodological quality of animal studies.

Keywords: Risk of bias, Methodological quality, Animal studies, Systematic reviews, Tool, Translational research

Background

The use of systematic reviews (SRs) for making evidenced-based decisions on healthcare is common practice in the clinical setting. Although most experimental animal studies aim to test safety and or efficacy of treatments to be used for human healthcare, summarizing the available evidence in an SR is far less common in the field of laboratory animal experiments. Fortunately, since an influential commentary was published in the Lancet (2002) [1], first setting out the scientific rationale for SRs of animal studies, awareness of the merits of SRs of experimental animal studies has been steadily increasing [2]. The methodology for conducting SRs of animal intervention studies is currently evolving but not yet as advanced as for clinical studies. In the clinical field, the randomized controlled trial (RCT) is considered the paradigm for evaluating the effectiveness of interventions. Animal intervention studies, like RCTs, are experimental studies, but they differ from RCTs in many respects [3] (Table 1, supporting information in Additional file 1). This means that some aspects of the systematic review process need to be adapted to the characteristics of animal intervention studies. In this paper, we focus on the methodology for assessing the risk of bias in animal intervention studies.

The extent to which an SR can draw reliable conclusions depends on the validity of the data and the results of the included studies [4-8]. Assessing the risk of bias of the individual studies, therefore, is a key feature of an SR. To assess the risk of bias of RCTs, the Cochrane Collaboration developed the Cochrane RoB Tool [9]. Such a general tool is not yet available for animal intervention

* Correspondence: Carlijn.Hooijmans@radboudumc.nl
[1]SYRCLE at Central Animal Laboratory, Radboud University Medical Center, Nijmegen, the Netherlands
Full list of author information is available at the end of the article

Table 1 Main differences between randomized clinical trials (RCTs) and animal intervention studies

RCT	Animal intervention study
Objective: demonstrating clinical efficacy	Objective: understanding disease mechanisms, suggesting intervention strategies (guiding clinical trials), examining potential efficacy, safety and toxicity of interventions
Disease naturally present	Disease often induced (with unclear/insufficient similarity to the human condition)
Timing of applying the intervention in relation to the disease onset is often heterogeneous	Intervention is often applied at a known time point in relation to the induced disease state
Often a heterogeneous group of patients (for example, lifestyle and co-morbidities)	Often a considerably homogeneous study population (e.g., comparable/controlled housing conditions and animal characteristics such as genetic backgrounds, gender, and presence of co-morbidities)
Sample size relatively large (compared to animal studies)**	Sample size relatively small (compared to RCTs) and sample size calculations often not reported
In general, relatively high internal validity because of randomization and blinding (compared to animal studies)**	In general, low internal validity (compared to RCTs) E.g., not yet standard practice to: -Randomize allocation of the animal to the intervention and control groups -Blind personnel and outcome assessors
Patients can be blinded for treatment in many situations.	Animals cannot and need not be blinded for treatment.
Relatively high external validity (extrapolation within one species)	Relatively low external validity (extrapolation between different species)
Relatively large teams involved Intervention staffs are often different from outcome assessment staff.	Relatively small teams involved One researcher is often responsible for treatment allocation and administration, outcome assessment and data analysis.
In general, no post-mortem data	In general, post-mortem material available Animals are often sacrificed at the end of the experiment.
Outcomes are often patient-relevant outcomes (compared to animal studies)	Outcomes are often surrogate outcomes, and still difficult to translate to the clinical setting even if similar to clinical outcomes
Clear guidelines for reporting and methodological quality [25]	Evolving guidelines for reporting and methodological quality [2,23,24]

**Additional file 1 provides some supportive information for this statement.
The differences described in this Table indicate general tendencies and may, therefore, not apply to all RCTs and animal intervention studies.

studies. The checklists and scales currently used for assessing study validity of animal studies [10-14] vary greatly, are sometimes designed for a specific field (i.e., toxicology) and often assess reporting quality and internal and external validity simultaneously. We believe that, although it is important to asses all aspects of study quality in an SR, the assessment and interpretation of these aspects should be conducted separately. After all, the consequences of poor reporting, methodological quality and generalizability of the results are very different. Here, the SYstematic Review Centre for Laboratory animal Experimentation (SYRCLE) presents an RoB tool for animal intervention studies: SYRCLE's RoB tool. This tool, based on the Cochrane Collaboration RoB Tool [9], aims to assess methodological quality and has been adapted to aspects of bias that play a role in animal experiments.

Methods

Development of SYRCLE's RoB tool

The Cochrane RoB Tool was the starting-point for developing an RoB tool for experimental animal studies. The Cochrane RoB Tool assesses the risk of bias of RCTs and addresses the following types of biases: selection bias, performance bias, attrition bias, detection bias and reporting bias [9]. The items in the Cochrane RoB Tool that were directly applicable to animal experiments were adopted (Table 2: items 1, 3, 8, 9 and 10).

To investigate which items in the tool might require adaptation, the differences between randomized clinical trials and animal intervention studies were set out (Table 1). Then we checked whether aspects of animal studies that differed from RCTs could cause bias in ways that had not yet been taken into account in the Cochrane RoB tool. Finally, the quality assessments of recent systematic reviews of experimental animal studies were examined to confirm that all aspects of internal validity had been taken into consideration in SYRCLE's RoB tool.

To enhance transparency and applicability, we formulated signaling questions (as used in the QUADAS tool, a tool to assess the quality of diagnostic accuracy studies [15,16]) to facilitate judgment. In order to obtain a preliminary idea of inter-observer agreement for each item in the RoB tool, Kappa statistics were determined on the basis of 1 systematic review including 32 papers.

Table 2 SYRCLE's tool for assessing risk of bias

Item	Type of bias	Domain	Description of domain	Review authors judgment
1	Selection bias	Sequence generation	Describe the methods used, if any, to generate the allocation sequence in sufficient detail to allow an assessment whether it should produce comparable groups.	Was the allocation sequence adequately generated and applied? (*)
2	Selection bias	Baseline characteristics	Describe all the possible prognostic factors or animal characteristics, if any, that are compared in order to judge whether or not intervention and control groups were similar at the start of the experiment.	Were the groups similar at baseline or were they adjusted for confounders in the analysis?
3	Selection bias	Allocation concealment	Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen before or during enrolment.	Was the allocation adequately concealed? (*)
4	Performance bias	Random housing	Describe all measures used, if any, to house the animals randomly within the animal room.	Were the animals randomly housed during the experiment?
5	Performance bias	Blinding	Describe all measures used, if any, to blind trial caregivers and researchers from knowing which intervention each animal received. Provide any information relating to whether the intended blinding was effective.	Were the caregivers and/or investigators blinded from knowledge which intervention each animal received during the experiment?
6	Detection bias	Random outcome assessment	Describe whether or not animals were selected at random for outcome assessment, and which methods to select the animals, if any, were used.	Were animals selected at random for outcome assessment?
7	Detection bias	Blinding	Describe all measures used, if any, to blind outcome assessors from knowing which intervention each animal received. Provide any information relating to whether the intended blinding was effective.	Was the outcome assessor blinded?
8	Attrition bias	Incomplete outcome data	Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized animals), reasons for attrition or exclusions, and any re-inclusions in analyses for the review.	Were incomplete outcome data adequately addressed? (*)
9	Reporting bias	Selective outcome reporting	State how selective outcome reporting was examined and what was found.	Are reports of the study free of selective outcome reporting? (*)
10	Other	Other sources of bias	State any important concerns about bias not covered by other domains in the tool.	Was the study apparently free of other problems that could result in high risk of bias? (*)

*Items in agreement with the items in the Cochrane Risk of Bias tool.

Results

SYRCLE's RoB tool

The resulting RoB tool for animal studies contains 10 entries (Table 2). These entries are related to 6 types of bias: selection bias, performance bias, detection bias, attrition bias, reporting bias and other biases. Items 1, 3, 8, 9 and 10 are in agreement with the items in the Cochrane RoB tool. The other items have either been revised or are completely new and will be discussed in greater detail below. Most of the variations between the two tools are a consequence of the differences in design between RCTs and animal studies (see also Table 1). Shortcomings in, or unfamiliarity with, specific aspects of the experimental design of animal studies compared to clinical studies also play a role.

Bias due to inadequate randomization and lack of blinding

Random allocation of animals to the experimental and control groups, firstly, is not yet standard practice in animal experiments [17]. Furthermore, as the sample size of most animal experiments is relatively small, important baseline differences may be present. Therefore, we propose to include the assessment of similarity in baseline characteristics between the experimental and control groups as a standard item. The number and type of baseline characteristics depend on the review question. Before launching a risk of bias assessment, therefore, reviewers need to discuss which baseline characteristics need to be comparable between the groups.

Secondly, we slightly adjusted the sequence allocation item, specifying that the allocation sequence should not only be adequately generated but also be adequately applied. We decided to do so because, in animal studies, diseases are often induced rather than naturally present. The timing of randomization, therefore, is more important than in a patient setting: it needs to be assessed whether the disease was induced before actual randomization and whether the order of inducement was randomly allocated. The signaling questions for judging this entry are represented in Table 3.

Thirdly, a new item pertains to randomizing the housing conditions of animals during the experiment. In animal studies, the investigators are responsible for the way the animals are housed. They determine, for example, the location of the cage in the room. As housing conditions (such as lighting, humidity, temperature, etc.) are known to influence study outcomes (such as certain biochemical parameters and behavior), it is important that the housing of these animals is randomized or, in other words, comparable between the experimental groups in order to reduce bias [18]. Animals from different treatment groups, for example, should not be housed per group on different shelves or in different rooms as the animals on the top shelf experience a higher room temperature than animals on the lowest shelf, and the temperature of the room may influence the toxicity of pharmacological agents (Table 4). When cages are not placed randomly (e.g., when animals are housed per group on different shelves), moreover, it is possible for the investigator to foresee or predict the allocation of the animals to the various groups, which might result in performance bias. Therefore, randomizing the housing conditions is also a requisite for adequately blinding the animal caregivers and investigators. Therefore, this has also been included as a signaling question in Table 3.

Fourthly, in a recent update of the Cochrane RoB tool (http://www.cochrane.org/sites/default/files/uploads/handbook/Whats%20new%20in%20Handbook%205_1_0.pdf), bias related to blinding of participants and personnel (performance bias) is assessed separately from bias related to blinding of outcome assessment (detection bias). In our tool, we followed this approach, although animals do not need to be blinded for the intervention as they do not have any expectations about the intervention. In addition, it is important to emphasize that personnel involved in the experimental animal studies should be taken to include animal caregivers. In animal studies, this group is often not taken into account when blinding the allocation of animals to various groups. If animal caregivers know that a drug might cause epileptic seizures or increases urine production, for example, they might handle the animals or clean the cages in the group receiving this drug more often, which could cause behavioral changes influencing the study results.

With regard to adequately blinding outcome assessment (entry 7), possible differences between the experimental and control groups in methods used for outcome assessment should be described and judged. It should also be determined whether or not animals were selected at random for outcome assessment, regardless of the allocation to the experimental or control group. For instance, when animals are sacrificed per group at various time points during the day, the scientist concerned might interpret the results of the groups differently because she or he can foresee or predict the allocation.

Another reason to select animals at random for outcome assessment is the presence of circadian rhythms in many biological processes (Table 4). Not selecting the animals for outcome assessment at random might influence the direction and magnitude of the effect. For example, the results of a variety of blood tests depend on their timing during the day: cholesterol levels in mice may be much higher in the morning after a meal than in the afternoon. Because of these effects, assessing whether or not animals were selected at random for outcome assessment has also been presented as a separate entry.

Reporting bias

As mentioned before, assessing reporting bias is in agreement with the Cochrane RoB tool. It is important to mention, however, that this item is quite difficult to assess in animal intervention studies at present because protocols for animal studies are not yet registered in a central, publicly accessible database. Nevertheless, many have called for registration of all animal experiments at inception [19,20], so we expect that registration of animal studies will be more common within a few years. For this reason, we already decided to include it in SYRCLE's RoB tool. Furthermore, protocols of animal studies, like those of clinical studies, can already be published in various (open access) journals, which will also help to improve the standard of research in animal sciences.

Other bias

Beyond the above-mentioned types of bias, there might be further issues that may raise concerns about the possibility of bias. These issues have been summarized in the other bias domain. The relevance of the signaling questions (Table 3) depends on the experiment. Review authors need to judge for themselves which of the items could cause bias in their results and should be assessed. In assessing entry 10 ("Was the study apparently free of other risks of bias?"), it is important to pay extra attention to the presence of unit-of-analysis errors. In animal studies, the experimental unit is often not clear, and as a consequence statistical measures are often inaccurately calculated. For example, if mice in a cage are given a treatment in their diet, it is the cage of animals rather than the individual animal that is the experimental unit. After all, the mice in the cage cannot have

Table 3 Signaling questions

The additional signaling questions are included to assist judgment. "Yes" indicates low risk of bias; "no" indicates high risk of bias; and "unclear" indicates an unclear risk of bias. If one of the relevant signaling questions is answered with "no," this indicates high risk of bias for that specific entry.

1) Was the allocation sequence adequately generated and applied?

*Did the investigators describe a random component in the sequence generation process such as: Yes/No/Unclear

- Referring to a random number table;
- Using a computer random number generator.

Additional info:

Examples of a non-random approach:

- Allocation by judgment or by investigator's preference;
- Allocation based on the results of a laboratory test or a series of tests;
- Allocation by availability of the intervention;
- Sequence generated by odd or even date of birth;
- Sequence generated by some rule based on animal number or cage number.

2) Were the groups similar at baseline or were they adjusted for confounders in the analysis?

*Was the distribution of relevant baseline characteristics balanced for the intervention and control groups? Yes/No/Unclear

*If relevant, did the investigators adequately adjust for unequal distribution of some relevant baseline characteristics in the analysis? Yes/No/Unclear

*Was the timing of disease induction adequate? Yes/No/Unclear

Additional info:

The number and type of baseline characteristics are dependent on the review question. Before starting their risk of bias assessment, therefore, reviewers need to discuss which baseline characteristics need to be comparable between the groups. In an SR investigating the effects of hypothermia on infarct size, for example, gender distribution, left ventricular weight and heart rate and blood pressure should be similar between the groups at the start of the study.

A description of baseline characteristics and/or confounders usually contains:

- The sex, age and weight of the animals
- Baseline values of the outcomes which are of interest in the study

Timing of disease induction:

In some prevention studies, the disease is induced after allocation of the intervention. For example, in an experiment on preventive probiotic supplementation in acute pancreatitis, pancreatitis is induced after allocation of the animals to the probiotic or control group. To reduce baseline imbalance, the timing of disease induction should be equal for both treatment groups.

Examples of adequate timing of disease induction:

- The disease was induced before randomization of the intervention.
- The disease was induced after randomization of the intervention, but the timing of disease induction was at random, and the individual inducing the disease was adequately blinded from knowing which intervention each animal received.

Table 3 Signaling questions *(Continued)*

3) Was the allocation to the different groups adequately concealed during?

*Could the investigator allocating the animals to intervention or control group not foresee assignment due to one of the following or equivalent methods? Yes/No/Unclear

- Third-party coding of experimental and control group allocation Central randomization by a third party

Sequentially numbered opaque, sealed envelopes

Additional info:

Examples of investigators allocating the animals being possibly able to foresee assignments:

- Open randomization schedule
- Envelopes without appropriate safeguard
- Alternation or rotation
- Allocation based on date of birth
- Allocation based on animal number
- Any other explicitly unconcealed procedure of a non-random approach

4) Were the animals randomly housed during the experiment?

*Did the authors randomly place the cages or animals within the animal room/facility? Yes/No/Unclear

- Animals were selected at random during outcome assessment (use signaling questions of entry 6).

*Is it unlikely that the outcome or the outcome measurement was influenced by not randomly housing the animals? Yes/No/Unclear

The animals from the various experimental groups live together in one cage/pasture (e.g., housing conditions are identical).

Additional info:

Examples of investigators using a non-random approach when placing the cages:

- Experimental groups were studied on various locations (e.g., group A in lab A or on shelf A; Group B in Lab B or on shelf B).

5) Were the caregivers and/or investigators blinded from knowledge which intervention each animal received during the experiment?

*Was blinding of caregivers and investigators ensured, and was it unlikely that their blinding could have been broken? Yes/No/Unclear

- ID cards of individual animals, or cage/animal labels are coded and identical in appearance.
- Sequentially numbered drug containers are identical in appearance.
- The circumstances during the intervention are specified and similar in both groups (#).
- Housing conditions of the animals during the experiment are randomized within the room (use criteria of entry 4).

Table 3 Signaling questions *(Continued)*

Additional info:	
Examples of inappropriate blinding:	
■ Colored cage labels (red for group A, yellow group B) ■ Expected differences in visible effects between control and experimental groups ■ Housing conditions of the animals are not randomized within the room during the experiment; use criteria of entry 4 ■ The individual who prepares the experiment is the same as the one who conducts and analyses the experiment ■ Circumstances during the intervention are not similar in both groups (#)	
Examples where circumstances during the intervention were not similar:	
■ Timing of administration of the placebo and exp drug was different. ■ Instruments used to conduct experiment differ between experimental and control group (e.g., experiment about effects abdominal pressure; exp group receives operation and needle to increase pressure, while control group only has the operation).	
***The relevance of the above-mentioned items depends on the experiment. Authors of the review need to judge for themselves which of the above-mentioned items could cause bias in the results when not similar. These should be assessed.*	
6) Were animals selected at random for outcome assessment?	
*Did the investigators randomly pick an animal during outcome assessment, or did they use a random component in the sequence generation for outcome assessment?	Yes/No/ Unclear
■ Referring to a random number table; ■ Using a computer random number generator; ■ Etc.	
7) Was the outcome assessor blinded?	
*Was blinding of the outcome assessor ensured, and was it unlikely that blinding could have been broken?	Yes/No/ Unclear
■ Outcome assessment methods were the same in both groups. ■ Animals were selected at random during outcome assessment (use signaling questions of entry 6).	
*Was the outcome assessor not blinded, but do review authors judge that the outcome is not likely to be influenced by lack of blinding?	Yes/No/ Unclear
(e.g., mortality)	
Additional info:	
This item needs to be assessed for each main outcome.	
8) Were incomplete outcome data adequately addressed? (*)	
*Were all animals included in the analysis?	Yes/No/ Unclear
*Were the reasons for missing outcome data unlikely to be related to true outcome? (e.g., technical failure)	Yes/No/ Unclear
*Are missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups?	Yes/No/ Unclear
*Are missing outcome data imputed using appropriate methods?	Yes/No/ Unclear
9) Are reports of the study free of selective outcome reporting? (*)	
*Was the study protocol available and were all of the study's pre-specified primary and secondary outcomes reported in the current manuscript?	Yes/No/ Unclear
*Was the study protocol not available, but was it clear that the published report included all expected outcomes (i.e. comparing methods and results section)?	Yes/No/ Unclear
Additional info:	
Selective outcome reporting:	
- Not all of the study's pre-specified primary outcomes have been reported; - One or more primary outcomes have been reported using measurements, analysis methods or data subsets (e.g., subscales) that were not pre-specified in the protocol; - One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting has been provided, such as an unexpected adverse effect); - The study report fails to include results for a key outcome that would be expected to have been reported for such a study.	
10) Was the study apparently free of other problems that could result in high risk of bias? (*)	
*Was the study free of contamination (pooling drugs)?	Yes/No/ Unclear
*Was the study free of inappropriate influence of funders?	Yes/No/ Unclear
*Was the study free of unit of analysis errors?	Yes/No/ Unclear
*Were design-specific risks of bias absent?	Yes/No/ Unclear
*Were new animals added to the control and experimental groups to replace drop-outs from the original population?	Yes/No/ Unclear
Additional info:	
The relevance of the signaling questions (Table 3) depends on the experiment. Review authors need to judge for themselves which of the items could cause bias in their results and should be assessed.	
Contamination/pooling drugs:	
Experiments in which animals receive – besides the intervention drug – additional treatment or drugs which might influence or bias the result.	
Unit of analysis errors:	
■ Interventions to parts of the body within one participant (i. e., one eye exp; one eye control). ■ All animals receiving the same intervention are caged together, but analysis was conducted as if every single animal was one experimental unit.	

Table 3 Signaling questions *(Continued)*

Design-specific risks of bias:

- Crossover design that was not suitable (intervention with no temporary effect, or the disease is not stable over time)
- Crossover design with risk of carry-over effect
- Crossover design with only first period data being available
- Crossover design with many animals not receiving 2nd or following treatment due to large number of drop-outs probably due to longer duration of study
- Crossover design in which all animals received same order of interventions
- Multi-arm study in which the same comparisons of groups are not reported for all outcomes (selective outcome reporting)
- Multi-arm study in which results of different arms are combined (all data should be presented per group)
- Cluster randomized trial not taking clustering into account during statistical analysis (unit of analysis error)
- Crossover design in which paired analysis of the results is not taken into account

different treatments, and they may be more similar than mice in different cages.

Use of SYRCLE's RoB tool

In order to assign a judgment of low, high or unclear risk of bias to each item mentioned in the tool, we have produced a detailed list with signaling questions to aid the judgment process (Table 3). It is important to emphasize that this list is not exhaustive. We recommend that people assessing the risk of bias of the included studies discuss and adapt this list to the specific needs of their review in advance. A "yes" judgement indicates a low risk of bias; a "no" judgment indicates high risk of bias; the judgment will be "unclear" if insufficient details have been reported to assess the risk of bias properly.

As a rule, assessments should be done by at least two independent reviewers, and disagreements should be resolved through consensus-oriented discussion or by consulting a third person.

We recommend that risk of bias assessment is presented in a table or figure. The investigators can present either the summary results of the risk of bias assessment or the results of all individual studies. Finally, the results of the risk of bias assessment could be used when interpreting the results of the review or a meta-analysis. For instance, sensitivity analysis can be used to show how the conclusions of the review might be affected if studies with a high risk of bias were excluded from the analysis [8,9].

We do not recommend calculating a summary score for each individual study when using this tool. A summary score inevitably involves assigning "weights" to specific domains in the tool, and it is difficult to justify the weights assigned. In addition, these weights might differ per outcome and per review.

Inter-observer variability

Inter-observer agreement was evaluated using Kappa statistics. At time of writing, the Kappa statistics could only be determined for items 1, 6, 7, 8, 9 and 10 and was based on 2 raters in one systematic review including 32 papers. For items 1, 6, 7, 8, 9 and 10, the inter-observer variability varied between 0.62 and 1.0. Kappa was for item 1: 0.87; item 6: 0.74; item 7: 0.59; item 8: 1.0; item 9: 0.62; item 10: 1.0. Kappa could not be calculated for items 2, 3, 4, and 5 as Kappa is defined for situations with at least two raters and two outcomes, and in these items we had only 1 outcome (unclear risk of bias) as a result of poor reporting.

Discussion and conclusion

In animal studies, a large variety of tools to assess study quality is currently used, but none of the tools identified so far focussed on internal validity only [11]. Most instruments assess reporting quality and internal and external validity simultaneously although consequences

Table 4 Some underlying evidence for the importance of random housing and random outcome assessment

Random housing		
Lighting	Light exposure varies with respect to rack location and position of cages within the rack	[27,28]
	Small differences in light intensity have been associated with reproductive and behavioral changes	[27,29,30]
	There can be a four-fold difference in light intensity between cages at the top or bottom of a rack	[18]
Temperature	Temperature in animal room at 1.5 m can be 3-4°C higher than at 0.5 m	[18]
	Cage temperature varies with group size	[18]
	Cage temperature varies with height of placement within the rack (top rack 5°C warmer than bottom rack)	[18,31,32]
	Small changes in temperature can influence metabolic rates and toxicity	[27,31,33]
Random outcome assessment		
Circadian rhythm	Periodic/circadian variations in lipid metabolism, neurotransmitter levels, pharmacokinetic effects, etc.	[34-37]

Suggestions for further reading: [18,27,37,38].

of poor reporting, risk of bias and generalizability of the results are very different.

Therefore, we developed SYRCLE's RoB tool to establish consistency and avoid discrepancies in assessing risk of bias in SRs of animal intervention studies. SYRCLE's RoB tool is based on the Cochrane RoB tool [9] and has been adjusted for particular aspects of bias that play a role in animal intervention studies. All items in our RoB tool can be justified from a theoretical perspective, but not all items have been validated by empirical research. However, the same holds for the original QUADAS tool (to assess the quality of diagnostic accuracy studies) and the Cochrane RoB tool [8,16]. For example, in the Cochrane RoB tool, the item on "inadequately addressing incomplete outcome data" is mainly driven by theoretical considerations [8]. In QUADAS, no empirical or theoretical evidence was available for 2 out of the 9 risk of bias items [16].

Although validation is important, providing empirical evidence for all items in this tool is not to be expected in the near future as this would require major comparative studies, which, to our knowledge, are not currently being undertaken or scheduled. Using the existing animal experimental literature is also challenging because the current reporting quality of animal studies is poor [17]; many details regarding housing conditions or timing outcome assessment are often unreported. However, we feel that publishing this tool is necessary to increase awareness of the importance of improving the internal validity of animal studies and to gather practical experience of authors using this tool.

We started to use this tool in our own SRs and hands-on training courses on conducting SRs in laboratory animal experimentation, funded by The Netherlands Organization for Health Research and Development (ZonMW). The first experiences with this tool were positive, and users found SYRCLE's RoB tool very useful. The inter-rater variability Kappa varied between 0.6 and 1 9. Users also indicated that they had to judge many entries as "unclear risk of bias". Although most users did not expect this finding, it is not altogether surprising [21,22], as a recent survey of 271 animal studies revealed that reporting experimental details on animals, methods and materials is very poor [17]. We hope and expect, therefore, that use of this tool will improve the reporting quality of essential experimental details in animal studies [23,24].

Widespread adoption and implementation of this tool will facilitate and improve critical appraisal of evidence from animal studies. This may subsequently enhance the efficiency of translating animal research results into clinical practice. Furthermore, this tool should be tested by authors of SRs of animal intervention studies to test its applicability and validity in practice. We invite users of SYRCLEs RoB tool, therefore, to provide comments and feedback via the SYRCLE LinkedIn group (risk of bias subgroup) http://www.linkedin.com/groups?gid=4301693&trk=hb_side_g. As with the QUADAS, CONSORT and PRISMA statements [15,16,25,26], we expect that user feedback and developments in this relatively new field of evidence-based animal experimentation will allow us to update this tool within a few years.

Additional file

Additional file 1: A pilot survey to provide some supportive information for some of the statements made in Table 1.

Competing interests

All authors declare that none of the authors (Hooijmans, Rovers, de Vries, Leenaars, Ritskes-Hoitinga, Langendam) have anything to disclose or any competing interests. The authors had no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years, no other relationships or activities that could appear to have influenced the submitted work.

Authors' contributions

CRH coordinated the project. CRH, MWL and MMR have made substantial contributions to the design of the RoB tool, article conception and wrote the first draft of the paper. MRH, ML and RdV provided advice on bias in animal studies (part of the discussion group). MRH and RdV revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Rikkert Stuve (www.textconsultant.nl) for copy-editing services.

Funding

The development of SYRCLE's RoB tool was partly funded by the Ministry of Health, Welfare and Sport of the government of the Netherlands (grant nr: 321200). The views expressed in this article are those of the authors and not necessarily those of the funder.

Author details

[1]SYRCLE at Central Animal Laboratory, Radboud University Medical Center, Nijmegen, the Netherlands. [2]Centre of Evidence-based Surgery, Radboud University Medical Center, Nijmegen, the Netherlands. [3]Dutch Cochrane Centre, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands.

Received: 10 September 2013 Accepted: 10 March 2014
Published: 26 March 2014

References

1. Sandercock P, Roberts I: **Systematic reviews of animal experiments.** *Lancet* 2002, **360**(9333):586.
2. Hooijmans CR, Rovers M, de Vries RB, Leenaars M, Ritskes-Hoitinga M: **An initiative to facilitate well-informed decision-making in laboratory animal research: report of the First International Symposium on Systematic Reviews in Laboratory Animal Science.** *Lab Anim* 2012, **46**(4):356–357.
3. Muhlhausler BS, Bloomfield FH, Gillman MW: **Whole animal experiments should be more like human randomized controlled trials.** *PLoS Biol* 2013, **11**(2):e1001481.
4. Chalmers TC, Celano P, Sacks HS, Smith H Jr: **Bias in treatment assignment in controlled clinical trials.** *N Engl J Med* 1983, **309**(22):1358–1361.
5. Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, Buchan A, van der Worp HB, Traystman R, Minematsu K, Donnan GA, Howells DW: **Good laboratory practice: preventing introduction of bias at the bench.** *Stroke* 2009, **40**(3):e50–e52.

6. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, Pham B, Klassen TP: **Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses.** *Health Technol Assess* 1999, **3**(12):1–4. 1–98.
7. 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):408–412.
8. *Cochrane Handbook for Systematic Reviews of Interventions.* Edited by Higgins JPT, Green S. Chichester: John Wiley & Sons; 2008.
9. Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA, Cochrane Bias Methods Group, Cochrane Statistical Methods Group: **The Cochrane Collaboration's tool for assessing risk of bias in randomised trials.** *BMJ* 2011, **343:**d5928.
10. Hooijmans CR, Pasker-de Jong PC, de Vries RB, Ritskes-Hoitinga M: **The effects of long-term omega-3 fatty acid supplementation on cognition and Alzheimer's pathology in animal models of Alzheimer's disease: a systematic review and meta-analysis.** *J Alzheimers Dis* 2012, **28**(1):191–209.
11. Krauth D, Woodruff TJ, Bero L: **Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review.** *Environ Health Perspect* 2013, **121:**985–992.
12. Sena E, van der Worp HB, Howells D, Macleod M: **How can we improve the pre-clinical development of drugs for stroke?** *Trends Neurosci* 2007, **30**(9):433–439.
13. Wever KE, Menting TP, Rovers M, van der Vliet JA, Rongen GA, Masereeuw R, Ritskes-Hoitinga M, Hooijmans CR, Warle M: **Ischemic preconditioning in the animal kidney, a systematic review and meta-analysis.** *PLoS One* 2012, **7**(2):e32296.
14. Thayer K, Rooney A, Boyles A, Holmgren S, Walker V, Kissling G: **Draft protocol for systematic review to evaluate the evidence for an association between bisphenol A (BPA) exposure and obesity.** In *National Toxicology Program.* U.S. Department of health and human services; 2013.
15. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, Group Q: **QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies.** *Ann Intern Med* 2011, **155**(8):529–536.
16. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J: **The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews.** *BMC Med Res Methodol* 2003, **3:**25.
17. Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG: **Survey of the quality of experimental design, statistical analysis and reporting of research using animals.** *PLoS One* 2009, **4**(11):e7824.
18. Beynen ACG,K, van Zutphen LFM: **Standardization of the animal and its environment.** In *Principles of Laboratory Animal Science*, Revised Edition. Edited by van Zutphen LFMB V, Beynen AC. Amsterdam and New York: Elsevier B.V.; 2001.
19. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, Macleod M, Mignini LE, Jayaram P, Khan KS: **Comparison of treatment effects between animal experiments and clinical trials: systematic review.** *BMJ* 2007, **334**(7586):197.
20. Roberts I, Kwan I, Evans P, Haig S: **Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation.** *BMJ* 2002, **324**(7335):474–476.
21. Faggion CM Jr, Giannakopoulos NN, Listl S: **Risk of bias of animal studies on regenerative procedures for periodontal and peri-implant bone defects - a systematic review.** *J Clin Periodontol* 2011, **38**(12):1154–1160.
22. Hooijmans CR, de Vries RB, Rovers MM, Gooszen HG, Ritskes-Hoitinga M: **The effects of probiotic supplementation on experimental acute pancreatitis: a systematic review and meta-analysis.** *PLoS One* 2012, **7**(11):e48811.
23. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M: **A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible.** *Altern Lab Anim* 2010, **38**(2):167–182.
24. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG: **Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research.** *PLoS Biol* 2010, **8**(6):e1000412.
25. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF: **Improving the quality of reporting of randomized controlled trials. The CONSORT statement.** *JAMA* 1996, **276**(8):637–639.
26. Moher D, Liberati A, Tetzlaff J, Altman DG: **Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement.** *PLoS Med* 2009, **6**(7):e1000097.
27. Johnston NA, Nevalainen T: **Impact of biotic and abiotic environment on animal experiments.** In *Handbook of Laboratory Animal Science (Book 1).* 3rd edition. Edited by Hau JS S. Boca Raton, Florida: CRC Press; 2010:343–369.
28. Kemppinen N, Meller A, Bjork E, Kohila T, Nevalainen T: **Exposure in the shoebox: comparison of physical environment of ivcs and open rat cages.** *Scand J Lab Anim Sci* 2008, **35**(2):97–103.
29. Donnelly H, Saibaba P: **Light-intensity and the estrous-cycle in albino and normally pigmented mice.** *Lab Anim* 1993, **27**(4):385–390.
30. Vanderschuren LJMJ, Niesink RJM, Spruijt BM, Vanree JM: **Influence of environmental-factors on social play-behavior of juvenile rats.** *Physiol Behav* 1995, **58**(1):119–123.
31. Clough G: **Environmental-effects on animals used in biomedical-research.** *Biol Rev* 1982, **57**(Aug):487–523.
32. Clough G: **Environmental factors in relation to the comfort and well-being of laboratory rats and mice.** In *Standards in Laboratory Animal Management*, Volume 1. Wheathampstead: Universities Federation for Animal Welfare (UFAW); 1984:7–24.
33. Harri MNE: **Effect of body-temperature on cardiotoxicity of isoprenaline in rats.** *Acta Pharmacol Toxicol (Copenh)* 1976, **39**(2):214–224.
34. Bruguerolle B: **Circadian changes in pharmacokinetics.** *Therapie* 1983, **38**(3):223–235.
35. Bruguerolle B, Valli M, Bouyard L, Jadot G, Bouyard P: **Effect of the hour of administration on the pharmacokinetics of lidocaine in the rat.** *Eur J Drug Metab Pharmacokinet* 1983, **8**(3):233–238.
36. Marrino P, Gavish D, Shafrir E, Eisenberg S: **Diurnal-variations of plasma-lipids, tissue and plasma-lipoprotein lipase, and VLDL secretion rates in the rat - a model for studies of VLDL metabolism.** *Biochim Biophys Acta* 1987, **920**(3):277–284.
37. Claassen V: **Circadian and other rhytms.** In *Neglected Factors in Pharmacology and Neuroscience Research Volume 12.* Edited by Huston JP. Amsterdam and New York: Elsevier Science B.V; 1994:243–381.
38. Claassen V: **Housing conditions.** In *Neglected Factors in Pharmacology and Neuroscience Research, Volume 12.* Edited by Huston JP. Amsterdam and New York: Elsevier Science B.V; 1994:223–250.

doi:10.1186/1471-2288-14-43
Cite this article as: Hooijmans *et al.*: SYRCLE's risk of bias tool for animal studies. *BMC Medical Research Methodology* 2014 **14**:43.

Appendix D. Cochrane Risk of Bias Tool

Cochrane Collaboration modified tool for assessing risk of bias for RCT's, PART I
Use this form to assess risk of bias for randomized controlled trials.

Bias is assessed as a judgement (high, low, or unclear) for individual elements from five domains (selection, performance, attrition, reporting, and other).

Risk of selection, reporting, and other bias are assessed in the **Quality Assessment Form Part I.** Risk of performance, detection, and attrition bias are assessed using the **Quality Assessment Form Part II.**

Using the guidance provided at the end of this form, select either "high", "low" or "unclear" for each judgment. When complete, proceed to **Part II of the Quality Assessment Form**

REF ID:					
Domain	**Description**	**High risk of bias**	**Low risk of bias**	**Unclear risk of bias**	**Reviewer Assessment**
Selection bias ***Random sequence generation***	Described the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups. **Reviewer Comments:**	Selection bias (biased allocation to interventions) due to inadequate generation of a randomized sequence.	Random sequence generation method should produce comparable groups	Not described in sufficient detail	**Judgement** **Random sequence generation** □ **High** □ **Low** □ **Unclear**
Selection bias ***Allocation concealment***	Described the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen in advance of, or during, enrollment. **Reviewer Comments:**	Selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment.	Intervention allocations likely could not have been foreseen in advance of, or during, enrollment	Not described in sufficient detail	**Judgement** **Allocation concealment** □ **High** □ **Low** □ **Unclear**
Reporting bias ***Selective reporting***	Stated how the possibility of selective outcome reporting was examined by the authors and what was found. **Reviewer Comments:**	Reporting bias due to selective outcome reporting.	Selective outcome reporting bias not detected	Insufficient information to permit judgement (*It is likely that the majority of studies will fall into this category.)*	**Judgement** **Selective reporting** □ **High** □ **Low** □ **Unclear**
Other bias ***Other sources of bias***	Any important concerns about bias not addressed above. If particular questions/entries were pre-specified in the study's protocol, responses should be provided for each question/entry. **Reviewer Comments:**	Bias due to problems not covered elsewhere in the table.	No other bias detected	There may be a risk of bias, but there is either insufficient information to assess whether an important risk of bias exists; or insufficient rationale or evidence that an identified problem will introduce bias.	**Judgement** **Other sources of bias** □ **High** □ **Low** □ **Unclear**

Cochrane Collaboration modified tool for assessing risk of bias for RCT's, PART II
Use this form to assess risk of bias for randomized controlled trials.

Bias is assessed as a judgement (high, low, or unclear) for individual elements from five domains of bias (selection, performance, attrition, reporting, and other).

Risk of selection, reporting, and other bias are assessed in the **Quality Assessment Form Part I.** Risk of performance, detection, and attrition bias are assessed using the **Quality Assessment Form Part II.**

Using the guidance provided at the end of this form, select either "high", "low" or "unclear" for each judgement.

Risk of bias for the domains in the Form Part II will be assessed for each main or class of outcomes. Please indicate the specific outcome and complete the assessment for each.

REF ID:					
Outcomes:					
Domain	**Description**	**High risk of bias**	**Low risk of bias**	**Unclear risk of bias**	**Reviewer Assessment**
Performance bias ***Blinding (participants and personnel)***	Described all measures used, if any, to blind study participants and personnel from knowledge of which intervention a participant received. Provided any information relating to whether the intended blinding was effective. **Reviewer Comments:**	Performance bias due to knowledge of the allocated interventions by participants and personnel during the study.	Blinding was likely effective.	Not described in sufficient detail	**Judgement** **Blinding (participants and personnel)** □ **High** □ **Low** □ **Unclear**
Detection bias ***Blinding (outcome assessment)***	Described all measures used, if any, to blind outcome assessors from knowledge of which intervention a participant received. Provided any information relating to whether the intended blinding was effective. **Reviewer Comments:**	Detection bias due to knowledge of the allocated interventions by outcome assessors.	Blinding was likely effective.	Not described in sufficient detail	**Judgement** **Blinding (outcome assessment)** □ **High** □ **Low** □ **Unclear**
Attrition bias ***Incomplete outcome data***	Described the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. Stated whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomized participants), reasons for attrition/exclusions where reported. **Reviewer Comments:**	Attrition bias due to amount, nature or handling of incomplete outcome data.	Handling of incomplete outcome data was complete and unlikely to have produced bias	Insufficient reporting of attrition/exclusions to permit judgment of 'Low risk' or 'High risk' (e.g. number randomized not stated, no reasons for missing data provided)	**Judgement** **Incomplete outcome data** □ **High** □ **Low** □ **Unclear**

Higgins JPT, Altman DG, Sterne JAC (editors). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S (editors). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 (updated March 2011). The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

Table 8.5.d: Criteria for judging risk of bias in the Cochrane 'Risk of bias' assessment tool

RANDOM SEQUENCE GENERATION **Selection bias (biased allocation to interventions) due to inadequate generation of a randomised sequence.**	
Criteria for a judgement of 'Low risk' of bias.	The investigators describe a random component in the sequence generation process such as: • Referring to a random number table; • Using a computer random number generator; • Coin tossing; • Shuffling cards or envelopes; • Throwing dice; • Drawing of lots; • Minimization*. *Minimization may be implemented without a random element, and this is considered to be equivalent to being random.
Criteria for the judgement of 'High risk' of bias.	The investigators describe a non-random component in the sequence generation process. Usually, the description would involve some systematic, non-random approach, for example: • Sequence generated by odd or even date of birth; • Sequence generated by some rule based on date (or day) of admission; • Sequence generated by some rule based on hospital or clinic record number. Other non-random approaches happen much less frequently than the systematic approaches mentioned above and tend to be obvious. They usually involve judgement or some method of non-random categorization of participants, for example: • Allocation by judgement of the clinician; • Allocation by preference of the participant; • Allocation based on the results of a laboratory test or a series of tests; • Allocation by availability of the intervention.
Criteria for the judgement of 'Unclear risk' of bias.	Insufficient information about the sequence generation process to permit judgement of 'Low risk' or 'High risk'.
ALLOCATION CONCEALMENT **Selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment.**	
Criteria for a judgement	Participants and investigators enrolling participants could not foresee

of ‘Low risk’ of bias.	assignment because one of the following, or an equivalent method, was used to conceal allocation: • Central allocation (including telephone, web-based and pharmacy-controlled randomization); • Sequentially numbered drug containers of identical appearance; • Sequentially numbered, opaque, sealed envelopes.
Criteria for the judgement of ‘High risk’ of bias.	Participants or investigators enrolling participants could possibly foresee assignments and thus introduce selection bias, such as allocation based on: • Using an open random allocation schedule (e.g. a list of random numbers); • Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or nonopaque or not sequentially numbered); • Alternation or rotation; • Date of birth; • Case record number; • Any other explicitly unconcealed procedure.
Criteria for the judgement of ‘Unclear risk’ of bias.	Insufficient information to permit judgement of ‘Low risk’ or ‘High risk’. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite judgement – for example if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.

BLINDING OF PARTICIPANTS AND PERSONNEL

Performance bias due to knowledge of the allocated interventions by participants and personnel during the study.

Criteria for a judgement of ‘Low risk’ of bias.	Any one of the following: • No blinding or incomplete blinding, but the review authors judge that the outcome is not likely to be influenced by lack of blinding; • Blinding of participants and key study personnel ensured, and unlikely that the blinding could have been broken.
Criteria for the judgement of ‘High risk’ of bias.	Any one of the following: • No blinding or incomplete blinding, and the outcome is likely to be influenced by lack of blinding; • Blinding of key study participants and personnel attempted, but likely that the blinding could have been broken, and the outcome is likely to be influenced by lack of blinding.
Criteria for the judgement of ‘Unclear risk’ of bias.	Any one of the following: • Insufficient information to permit judgement of ‘Low risk’ or ‘High risk’; • The study did not address this outcome.

BLINDING OF OUTCOME ASSESSMENT

Detection bias due to knowledge of the allocated interventions by outcome assessors.

Criteria for a judgement of ‘Low risk’ of bias.	Any one of the following: • No blinding of outcome assessment, but the review authors judge that the outcome measurement is not likely to be influenced by lack of blinding;

	• Blinding of outcome assessment ensured, and unlikely that the blinding could have been broken.
Criteria for the judgement of ‘High risk’ of bias.	Any one of the following: • No blinding of outcome assessment, and the outcome measurement is likely to be influenced by lack of blinding; • Blinding of outcome assessment, but likely that the blinding could have been broken, and the outcome measurement is likely to be influenced by lack of blinding.
Criteria for the judgement of ‘Unclear risk’ of bias.	Any one of the following: • Insufficient information to permit judgement of ‘Low risk’ or ‘High risk’; • The study did not address this outcome.
INCOMPLETE OUTCOME DATA **Attrition bias due to amount, nature or handling of incomplete outcome data.**	
Criteria for a judgement of ‘Low risk’ of bias.	Any one of the following: • No missing outcome data; • Reasons for missing outcome data unlikely to be related to true outcome (for survival data, censoring unlikely to be introducing bias); • Missing outcome data balanced in numbers across intervention groups, with similar reasons for missing data across groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk not enough to have a clinically relevant impact on the intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes not enough to have a clinically relevant impact on observed effect size; • Missing data have been imputed using appropriate methods.
Criteria for the judgement of ‘High risk’ of bias.	Any one of the following: • Reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across intervention groups; • For dichotomous outcome data, the proportion of missing outcomes compared with observed event risk enough to induce clinically relevant bias in intervention effect estimate; • For continuous outcome data, plausible effect size (difference in means or standardized difference in means) among missing outcomes enough to induce clinically relevant bias in observed effect size; • ‘As-treated’ analysis done with substantial departure of the intervention received from that assigned at randomization; • Potentially inappropriate application of simple imputation.
Criteria for the judgement of ‘Unclear risk’ of bias.	Any one of the following: • Insufficient reporting of attrition/exclusions to permit judgement of ‘Low risk’ or ‘High risk’ (e.g. number randomized not stated, no reasons for missing data provided); • The study did not address this outcome.

SELECTIVE REPORTING

Reporting bias due to selective outcome reporting.

Criteria for a judgement of 'Low risk' of bias.	Any of the following: • The study protocol is available and all of the study's pre-specified (primary and secondary) outcomes that are of interest in the review have been reported in the pre-specified way; • The study protocol is not available but it is clear that the published reports include all expected outcomes, including those that were pre-specified (convincing text of this nature may be uncommon).
Criteria for the judgement of 'High risk' of bias.	Any one of the following: • Not all of the study's pre-specified primary outcomes have been reported; • One or more primary outcomes is reported using measurements, analysis methods or subsets of the data (e.g. subscales) that were not pre-specified; • One or more reported primary outcomes were not pre-specified (unless clear justification for their reporting is provided, such as an unexpected adverse effect); • One or more outcomes of interest in the review are reported incompletely so that they cannot be entered in a meta-analysis; • The study report fails to include results for a key outcome that would be expected to have been reported for such a study.
Criteria for the judgement of 'Unclear risk' of bias.	Insufficient information to permit judgement of 'Low risk' or 'High risk'. It is likely that the majority of studies will fall into this category.

OTHER BIAS

Bias due to problems not covered elsewhere in the table.

Criteria for a judgement of 'Low risk' of bias.	The study appears to be free of other sources of bias.
Criteria for the judgement of 'High risk' of bias.	There is at least one important risk of bias. For example, the study: • Had a potential source of bias related to the specific study design used; or • Has been claimed to have been fraudulent; or • Had some other problem.
Criteria for the judgement of 'Unclear risk' of bias.	There may be a risk of bias, but there is either: • Insufficient information to assess whether an important risk of bias exists; or • Insufficient rationale or evidence that an identified problem will introduce bias.

1. Introduction

1.1. Background as provided by EFSA

Regulation (EC) No 178/2002[1] recommends that assessments undertaken by the European Food Safety Authority (EFSA) shall be conducted in an independent, objective and transparent manner, on the basis of all available scientific information and data. Moreover, the EFSA Science Strategy 2012–2016[2] emphasises the importance of strengthening the scientific basis for risk assessment and risk monitoring. Under this framework, EFSA started the PROMETHEUS project (PROmoting METHods for Evidence Use in Scientific assessments) (2014–2016), which aims to further improve the methods for "dealing with data and evidence" (i.e. collecting/extracting, validating/appraising, analysing and integrating data and evidence) in EFSA scientific assessments and to increase their consistency. In this context, Critical Appraisal Tools (CATs) providing structured and consistent guidance on how to assess the methodological quality of a study play a key role.

In the scientific community there are various on-going initiatives on CATs.[3] However, many of the available tools do not necessarily have a straightforward application for the needs of EFSA (e.g. they are developed for clinical trials or are specific for only a subset of studies relevant for EFSA). Therefore, the Assessment and Methodological support Unit (AMU) developed specific CATs for appraising: i) systematic reviews of interventions (SR), ii) randomised controlled trials (RCT) relevant for food and feed safety assessments; iii) Genetically Modified (GM) plant equivalence studies and iv) the methodological quality of extensive literature searches (ELS). These tools have been used to enhance standardisation, consistency and transparency of AMU appraisals and shall now be streamlined and published in an EFSA Technical Report in order to make them available to all EFSA units and panels.

1.2. Terms of Reference as provided by EFSA

Discuss and streamline the existing AMU CATs on systematic reviews of interventions (SR), randomised controlled trials (RCT), Genetically Modified (GM) plant equivalence studies and extensive literature searches (ELS).

Publish the revised CATs in a Technical Report.

The Technical Report should be practical and applicable to the different relevant food and feed safety fields. In particular, the report should include:

- the list of elements that are considered by AMU for critically appraising SR of interventions, RCT, GM plant equivalence studies and ELS and indications on how the Critical Appraisal Tools are currently used;
- a glossary of relevant terms.

2. Approach followed for developing the document

An internal task force was set up to discuss and streamline the CATs already developed by the AMU.

In addition to AMU staff, the task force was composed of EFSA staff from areas for which these CATs were considered particularly relevant.

With the aim of having an effective discussion, the task force members were assigned some specific CATs to review according to their background. The reviewers considered the comprehensiveness of the items included in the CATs, their clarity and the proposed ways to judge/characterise them. Each CAT was tested in parallel by two members of the task force using two concrete examples. The specific CATs were amended according to the feedback received.

[1] OJ L 31, 1.2.2002, p. 1.

[2] http://www.efsa.europa.eu/en/corporate/pub/sciencestrategy12.htm

[3] For example, SYRCLE's risk of bias tool for animal studies (Hooijmans et al., 2014); the Cochrane Collaboration's tool for assessing risk of bias in randomised trials (Higgins et al., 2011); or the OHAT risk of bias tool (Rooney et al., 2014; OHAT/NTP, 2015).

3. What is a Critical Appraisal Tool (CAT)?

A CAT allows the *methodological quality* of a study (or a process) to be assessed, which influences the *reliability* of the evidence produced by such a study. Reliability of a piece of evidence refers to: (i) *precision*, i.e. the extent to which random error is minimised and the outcome of the process is reproducible over time; and (ii) *accuracy* (also referred to as internal validity), i.e. the extent to which systematic error (bias) is minimised[4] (EFSA, 2015).

Detailed reporting of a study/process is important for assessing its methodological quality. In fact, a lack of transparent and relevant information can lead to (i) delays in the appraisal process if it is decided that clarification is required from the authors and/or (ii) increased uncertainty in the assessment when clarification is not asked for/received. However, the quality of the reporting has been addressed elsewhere by EFSA (EFSA, 2014) and it is not an aspect inherent to methodological quality; thus, it is outside the scope of the current document.

Tools for appraising the methodological quality of studies need to be design specific. For instance, the items to be considered when appraising an RCT are not the same as those to be considered for an observational study (e.g. randomisation). For the same study design, CATs should be applied by outcome or endpoint, because the methodological quality of a study in which multiple outcomes are assessed may differ depending on the outcome considered (Higgins et al., 2011). For instance, some outcomes may be analysed using appropriate methods and some others in the same study may not. CATs should be applied to each individual study included in the assessment to allow for consistent classification of the studies according to their methodological quality.

While helping to minimise subjectivity in the appraisal and maximise transparency, CATs always require domain and methodological expert judgement in their use (see also section 5 below).

4. Objective of this document

This report aims to provide a series of CATs containing a comprehensive list of items (called "appraisal questions") for appraising the following:

- systematic reviews of intervention studies (SR);
- randomised controlled trials (RCTs) in humans;
- genetically modified (GM) plant equivalence studies;
- the process for conducting extensive literature searches (ELS).

For each appraisal question, the focus is on the risk of bias/appropriateness of the design/conduct of the item under consideration and not on how the item was reported (see section 3).

However, the report does not provide detailed guidance on performing the appraisal, which may be considered by EFSA in a second step.

The CATs provided have to be considered as working documents and advances in empirical methodological research will be reflected in further improvements to these instruments. Additional practice and studies are needed with a focus on the reproducibility and construct validity[5] of these tools.

The CATs are provided as appendices to this document.

5. Intended users

Ideally the CATs should be used by a multidisciplinary team of experts that includes methodologists (e.g. experts in information science, epidemiology, statistics) and domain experts in the field of the study/process under assessment.

[4] Risk of bias also addresses aspects such as the sensitivity and specificity of the detection method used in an assessment.

[5] The extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured (Terwee et al., 2007).

The elements contained in the CATs could be relevant for various users in the EFSA context (e.g. EFSA staff, experts, applicants, contractors, Member States). Where appropriate, they could be used as a reference for developing tailor-made CATs for specific fields, being aware that such modifications could hamper comparability across assessments.

6. General structure of the CATs and legend

The CATs are provided as appendices to this document in a tabular format with five columns and the items are grouped by topic (e.g. methods, sample selection) identified by capital letters.

The Systematic Review, Randomised Controlled Trials and Extensive Literature Search CATs contain the following columns:

1. **#: the number of the item under assessment;**
2. **appraisal question:** the question to be answered to evaluate each individual item included an the CAT. It contains the description of the item to be evaluated, including some examples or a description of how the item should have been implemented in the study;
3. **information as reported:** in this column, information from the study/ELS under assessment is quoted or summarised;
4. **appraisal**: in this column, a concise answer to the appraisal question (see column 2) is provided. For each item considered in the appraisal, the appraisal scales shown in Table 1: or Table 2: will be used. The scales provide different ways to appraise the items under consideration according to their domain of pertinence (risk of bias or appropriateness). As the appraisal of the methodological quality of a study is topic specific, the guidance for judgement for each individual item should be tailored according to the topic of the study under assessment before starting the appraisal process. For instance, unblinded outcome assessors in an RCT would normally be appraised as having a high risk of bias. However, when an outcome is measured using a method that prevents subjectivity in the measurement, a lack of blinding could be judged as probably low risk of bias;

Table 1: Proposed appraisal scale for risk of bias (modified from OHAT/NTP, 2015)

Appraisal	Definition
Definitively low risk of bias	There is direct evidence in the study of low risk of bias practices.
Probably low risk of bias	There is indirect evidence in the study of low risk of bias practices OR it is deemed that deviations from low risk of bias practices for the item to be appraised would not appreciably bias the final results.
Probably high risk of bias	There is indirect evidence of high risk of bias practices OR there is insufficient information (e.g. not reported or "NR") provided about relevant risk of bias practices.
Definitively high risk of bias	There is direct evidence of high risk of bias practices.

Table 2: Proposed appraisal scale for appropriateness

Appraisal	Definition
Definitively appropriate	There is direct evidence in the study of appropriate practices.
Probably appropriate	There is indirect evidence of appropriate practices OR it is deemed that deviations from appropriate practices for the item to be appraised would not appreciably modify the final results.
Probably not appropriate	There is indirect evidence of inappropriate practices OR there is insufficient information (e.g. not reported or "NR") provided about appropriate practices.
Definitively not appropriate	There is direct evidence of inappropriate practices.
Not applicable	An item is not appraisable when a previous practice on which it is dependent was not performed. For instance, one of the questions of the SR CAT concerns the appraisal of the methodological quality of the studies. The subsequent question concerns the process that has been used when appraising the methodological quality (see items E1 and E2 of SR CAT in Appendix A –). If the methodological quality of the studies has not been appraised, the subsequent question (item) becomes not applicable.

5. **rationale for the appraisal**: in this column, the rationale supporting the appraisal is reported.

The GMO CAT contains the following columns:

1. **#**: the number of the item under assessment;
2. **appraisal question**: the question to be answered to evaluate each individual item included in the CAT. It contains the description of the item to be evaluated, including some examples or a description of how the item shall be implemented in the study;
3. **rationale of the appraisal and possible consequences of flaws identified:**
   - information from the review under assessment is quoted or summarised under the sub-heading "Information as provided";
   - the rationale of the assessment and the possible consequences of flaws identified are illustrated under the sub-heading "Explanation of the assessment and possible consequences". Weaknesses and unclear/missing items along with strengths are also described.
4. **answer to the appraisal question**: concise answer to the appraisal question (see column 2):
   - 'yes, fully'. The study is well performed as far as this item is concerned;
   - 'no'. The study is NOT at all well performed as far as this item is concerned;
   - 'partially'. The study is well performed as far as some aspects related to this item are concerned. However, other aspects are not handled appropriately or adequately;
   - 'unclear'. This option is applicable in two cases:
     - the item is poorly reported or not reported at all, making it difficult to assess it (and further information is needed from authors—see "action required" below);

     OR

     - the item seems appropriate from a methodological point of view, but domain expertise is not present in the team and confirmation is needed (see "action required" below).
5. **action required**: the action required, i.e. whether or not it is necessary to consult domain experts and/or the authors of the review under assessment:
   - *none*. Sufficient information is available for performing the assessment of this item;

- *ask confirmation from domain experts*. Sufficient information is available for performing the assessment of this item from a methodological point of view. However, confirmation from domain experts is needed;
- *ask authors for further information*. Insufficient information is available for performing the assessment of this item; therefore, there is a need for clarifications or additional information from the authors of the study.

7. Recommendations

It is recommended that:

- detailed guidelines be developed for applying the CATs;
- approaches be explored and implemented on how to prioritise the items to assess (i.e. according to the specific design and context, some items can be more relevant in terms of risk of bias and/or precision than others and consequently have a higher impact on the overall assessment).

References

EFSA (European Food Safety Authority), 2014. Guidance on Statistical Reporting. EFSA Journal 2014;12(12):3908, 18 pp. doi:10.2903/j.efsa.2014.3908

EFSA (European Food Safety Authority), 2015. Principles and process for dealing with data and evidence in scientific assessments. EFSA Journal 2015;13(5):4121, 36 pp. doi:10.2903/j.efsa.2015.4121

Higgins JPT, Altman DG and Sterne JAC, 2011. Assessing risk of bias in included studies. In: Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. Eds Higgins JPT and Green S. The Cochrane Collaboration. Available online: www.cochrane-handbook.org

Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M and Langendam MW, 2014. SYRCLE's risk of bias tool for animal studies. BMC Medical Research Methodology, 14, 43.

OHAT/NTP (Office of Health Assessment and Translation/National Toxicology Program), 2015. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Available online: http://ntp.niehs.nih.gov/ntp/ohat/pubs/handbookjan2015_508.pdf

Rooney AA, Boyles AL, Wolfe MS, Bucher JR and Thayer KA, 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environmental Health Perspectives, 122, 711–718.

Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM and de Vet HC, 2007. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology, 60, 34–42.

Glossary and Abbreviations

TERM **DEFINITION**

Bias A systematic error or deviation from the truth, in results or inferences

Bibliographic record An entry representing a specific item in a library catalogue or bibliographic database, containing all the data elements necessary for a full description, presented in a specific bibliographic format. In modern cataloguing, the standard format is machine-readable (*example*: the MARC record), but prior to the use of computers, the traditional format was the catalogue card

Boolean operator Boolean operators are used to combine terms when conducting electronic searches. Examples include "AND" (used to narrow a search), "OR" (used to broaden a search) and "NOT" (used to exclude terms from a search)

Controlled terms See Controlled vocabulary

Controlled vocabulary An established list of preferred terms from which a cataloguer or indexer must select when assigning subject headings or descriptors in a bibliographic record, to indicate, for example, the content of the work in a library catalogue, index, or bibliographic database. Synonyms are included as lead-in vocabulary, with instructions to see or use the authorised heading. For example, if the authorised subject heading for works about dogs is "Dogs," then all items about dogs will be assigned the heading "Dogs," including a work titled All about Canines. A cross-reference to the heading "Dogs" will be made from the term "Canines" to ensure that anyone looking for information about dogs under "Canines" will be directed to the correct heading. Controlled vocabulary is usually listed alphabetically in a subject headings list or thesaurus of indexing terms. The process of creating and maintaining a list of preferred indexing terms is called vocabulary control

Endpoint See Outcome(s)

Extensive literature search (ELS) A literature search process structured in a way to identify as many studies relevant to a review question as needed. It is tailored in order to address the trade-off between sensitivity and specificity depending on the context of the review question. The fundamental characteristics of an ELS are: (1) use of tailored search strings, and (2) tailored use of literature sources (i.e. bibliographic databases and other sources accessed via electronic or hand-searching – for example, websites, journal tables of content, theses repositories, etc.)

Free text search A search of a bibliographic database in which natural language words and phrases appearing in the text of the documents indexed, or in their bibliographic descriptions, are used as search terms, rather than terms selected from a list of controlled vocabulary (authorised subject headings or descriptors)

GMO An organism or microorganism whose genetic material has been altered by means of genetic engineering. Techniques of genetic engineering to produce a genetically modified organism (GMO) are described in Annex 1 A of Directive 2001/18/EC

Grey literature Types of publication which are less systematically recorded in bibliographic tools such as catalogues and databases than journals and books

Intervention questions A question that seeks to assess the effect of an intervention, which is the factor(s) to which the population is exposed (e.g. an additive in food or feed, a vaccine, or a disinfection or eradication method)

Key elements of a question
Elements of a review question that, if well defined, help to answer it (e.g. selecting the eligibility criteria for studies, developing the search strategy, selecting the studies, or collecting the data). The key elements vary depending on the question type. For questions about effects of an intervention or exposure, the key elements are the population (P), the intervention (I) or exposure (E), the comparator (C) and the outcome (O) (together represented as PICO or PECO). For test accuracy question, the key elements are the population (P), the index test (I) and the target condition (T) (together PIT). For descriptive questions (prevalence, incidence, occurrence and consumption), the key elements are the population (P) and the condition of interest (O) (together PO)

Metadata
It means, “data about data.” Structured information describing information resources/objects for a variety of purposes

Outcome(s)
Variable(s) for which data are collected to enable the questions of the study to be answered

PECO(S)
Acronym summarising the population (P), exposure (E), comparator (C) and outcome (O) in a question about an exposure effect. S stands for study design. See also “Key elements of a question”

PICO(S)
Acronym summarising the population (P), intervention (I), comparator (C) and outcome (O) in a question about an intervention effect. S stands for study design. See also ‘Key elements of a question’

PIT
Acronym summarising the population (P), index test (I), and target population (T) in a question about test accuracy. See also “Key elements of a question”.

PO
Acronym summarising the population (P) and outcome (O) in a descriptive question. See also “Key elements of a question”

Primary research study
The original study in which data were collected. The term is sometimes used to distinguish such studies from secondary studies that re-examine previously collected data (e.g. a review).

Record
Summary information about a full-text document or conference presentation, typically included in a bibliographic database, which may include a bibliographic reference and one or more of the following: an abstract or summary of the scientific content, additional categorisations or indexing terms

Reference
A conventional word or phrase used in a work to refer the reader to another part of the text (see above or see below) or a similar word or phrase used in an index, catalogue, or reference work to direct the user from one heading or entry to another (see or see also). Also refers to any Latin phrase used in footnotes, endnotes, and bibliographies to refer the reader to works previously quoted or cited, for example, ibid. and op. cit. Sometimes used synonymously with citation.

Also refers to a letter written in support of a person’s application for employment or housing, usually by someone familiar with the applicant’s qualifications or reputation, or to a person who agrees to be contacted for such a recommendation, usually by telephone.

Appendix B – Critical appraisal tool for assessing quality of Randomised Controlled Trials in humans (RCT CAT)

1. Randomised Controlled Trials: main definitions and the EFSA context

A Randomised Controlled Trial (RCT from now onward) is a study in which experimental units are allocated at random to interventions. One of these interventions is the standard of comparison or control. The control may be a standard practice, a placebo, or no intervention at all. RCTs seek to measure and compare the outcomes after the experimental units receive the interventions. Because the outcomes are measured, RCTs are quantitative studies. A RCT is considered the gold standard methodology to determining whether a cause–and–effect relationship exists between an intervention and an outcome (modified from Webster's New World™ Medical Dictionary, 3rd Edition, 2008).

In the EFSA context, RCTs are mainly encountered in the context of human nutrition. The current CAT does not take into account the relevance of the study to the assessment question instead it is specifically aimed at appraising the intrinsic methodological quality of the RCT.

2. RCT CAT

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
A.	**METHODS- Overview**			
A1	**Are general and specific objectives of the trial consistent with the research question?** General objective: testing for difference, non-inferiority or equivalence. Specific objectives: hypotheses to be tested. In general a study should report first the research question, then the general and specific objectives. In case the research question is not directly mentioned in the paper the reply to this question should be "Not Applicable".	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
A2	**Is the trial design appropriate to meet the objective?** E.g. cross-over/parallel, completely randomised/randomised block design, study duration etc.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
B.	**METHODS - Participants**			
B1	**Are the eligibility (inclusion/exclusion) criteria for participants consistent with the study objectives?** For instance the objective of a study may be to examine the effect on a specific outcome of the daily intake of a particular substance in diabetic patients. In this case the participants should be enrolled among people affected by diabetes mellitus as confirmed by specific criteria.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
B2	**Are the experimental settings appropriate to meet the objectives of the trial?** Examples of the items to be considered are: • location; • start and end dates; • interruptions to the trial; • delayed start at some locations/centres.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
C.	**METHODS – Interventions**			
C1	**Were the interventions (including administration route and dosage) defined consistently with the study objectives and appropriate?** Examples of the items to be considered are: • treatment and control arms; • dose; • administration route; • administration timing; • duration.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
D.	**METHODS – Outcomes**			
D1	**Are the outcomes appropriate to meet the objectives of the RCT?** Examples of the items to be considered are: • list of all outcomes measured in the trial; • primary outcomes; • secondary outcomes; • how were they measured? • when were they measured? • any changes with reasons after trial commenced.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
E.	**METHODS - Sample size calculation**			
E1	**Is sample size calculation appropriate for primary objective(s) and estimates/hypotheses to be tested?** *In case of hypotheses testing:* • power analysis carried out a priori considering the following elements: o desired confidence level; o desired minimum power of the test; o minimum effect size considered biologically relevant? o expected variability of the effect. • is the issue of multiplicity (if any) addressed while planning the sample size? • are drop out considered for sample size calculation?	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
F.	**METHODS – Randomisation (sequence generation and allocation concealment)**			
F1	**Is the random allocation appropriate?** Examples of the items to be considered are: • methods for generating random allocation sequence e.g. ○ algorithm used should generate an unpredictable allocation sequence; ○ type of randomisation should prevent unbalances in the allocation (e.g. completely randomised design, randomised block design). • implementation of the randomisation: the eligibility check of the subjects to be enrolled in the study should be done before randomisation.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively low risk of bias □ Probably low risk of bias □ Probably high risk of bias □ Definitively high risk of bias	*Please provide the main supporting information for your appraisal*
F2	**Is allocation concealment appropriate?** Assignment mechanism should prevent foreknowledge of treatment (i.e. prevent participants selection bias) until allocation. Note that allocation concealment can always be successfully implemented.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively low risk of bias □ Probably low risk of bias □ Probably high risk of bias □ Definitively high risk of bias	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
G.	**METHODS – Blinding**			
G1	**Were the participants and trial personnel appropriately blinded?** • assignment mechanism should prevent knowledge of treatment after allocation (i.e. prevent performance bias) by participants and trial personnel; • trial can be open label (no blinding), single blinded (only participants) or double blinded (participants and trial personnel). Note that it is not always possible to carry out a blinded study. For example, when the intervention consists of a substance with peculiar organoleptic characteristics (e.g. colour, taste or smell) which cannot be masked or replicate in the control or when the substance has remarkably visible effects (e.g. changes in the urine, faeces, skin) for which participants and trial personnel can assume to which group the subjects were randomised to. If an open label trial is performed an adequate justification should be provided by the authors.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively low risk of bias □ Probably low risk of bias □ Probably high risk of bias □ Definitively high risk of bias	*Please provide the main supporting information for your appraisal*
G2	**Were the outcome assessors appropriately blinded?** Assignment mechanism should prevent knowledge of treatment after allocation (i.e. prevent performance bias) by outcome assessors. Examples of outcome assessors are: • primary data collectors (e.g. interview staff responsible for measurement or collection of outcome data); • secondary assessors (e.g. external outcome adjudication committees).	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively low risk of bias □ Probably low risk of bias □ Probably high risk of bias □ Definitively high risk of bias	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
H.	**METHODS – Analysis populations**			
H1	**Was the population approach taken to analyse data appropriate? Was the approach planned a priori?** • intention to treat (ITT8); • per protocol (PP9) • complete cases.10 A justification should be provided for the choice.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

[8] Intention-to-treat: analysis carried out including all randomised participants and retaining all of them in the group to which they were allocated.

[9] Per protocol: analysis carried out by excluding participants who did not adequately adhere to the protocol (e.g. those who did not meet the inclusion, did not take all the intended treatment, or received a different treatment or no intervention).

[10] Complete cases: analysis carried out only on those whose outcome is known (in case on missing outcome data and no imputation envisaged).

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
I.	**METHODS - Data analysis**			
I1	**Was the need for any data transformation appropriately evaluated? If transformation was applied, is it appropriate? Was it planned a priori?** • Is there any evidence that transformation should have been used/not used? • If transformation applied, was the correct formula used? • If transformation applied, were data back-transformed appropriately?	Please insert a quotation or a brief summary of what has been reported in the RCT	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
I2	**Were appropriate methods used to detect and handle outliers? Were criteria for detection defined a priori?** • Was the presence of outliers appropriately checked (at least with graphical methods)? • Were outliers appropriately handled?	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
I3	**If any model was used to analyse data, is it appropriate? Was it planned a priori?** • assumptions underlying the analysis; • specification of the model consistent with variable types, objectives, assumptions etc.; • estimates of the effect; • indications of the goodness of fit of the model (if applicable).	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
I4	**If any method was used to handle missing data, is it appropriate? Was it planned a priori?** • Are the assumptions on the missing mechanism realistic in the specific context (e.g. missing at random – MAR, MNAR)? • Is the method used to handle missing data consistent with the assumptions?	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
I5	**Were appropriate tests used to assess the effects of the intervention? Was it planned a priori? Was the issue of multiple testing (multiplicity) appropriately handled?** Test suitable for: • type of variables in the study; • more than 2 intervention groups (if it is the case); • multiple primary endpoints (if it is the case); multiple measurements of primary endpoints (if it is the case).	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
I6	**Were potential discrepancies among groups at baseline adequately investigated and appropriately taken into consideration in the analyses?**	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively low risk of bias □ Probably low risk of bias □ Probably high risk of bias □ Definitively high risk of bias	*Please provide the main supporting information for your appraisal*
I7	**If any, were the additional analyses (e.g subgroups analysis, interim analyses, sensitivity analysis) performed appropriately? Were they planned a priori?** • foreseen prior to the start of the experiment; • consistently with objectives of the study.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
I8	*Only for crossover trials* **Are the following issues adequately addressed?** • carry over effect; • period effect; • sequence effect; • treatment by period interactions.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
J.	**RESULTS**			
J1	**Is the number of sampling units and measurements taken on them used in the analyses consistent with the approach to the analysis declared in the study?** The number of observations actually used in each analysis should be consistent with what occurred to the subjects, the measurement of endpoints along the study (subjects randomised, who received intended treatment, drop-outs etc…) and the approach to the analysis. E.g. if an intention to treat approach was declared, the number of observations used in the analysis must be consistent with the subjects allocated to the treatment groups.	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively low risk of bias □ Probably low risk of bias □ Probably high risk of bias □ Definitively high risk of bias	*Please provide the main supporting information for your appraisal*
J2	**Were the results appropriately presented?** • Was the uncertainty around the estimates appropriately investigated and appropriately taken into account in the conclusions? • Was the appropriate effect type taken into consideration in drawing conclusions (e.g. absolute and/or relative effect)? • Was the full list of endpoints taken into consideration for the conclusions? • Are the results of the analyses conducted post-hoc identifiable?	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
J3	**Were the results of the analysis interpreted in line with the stated hypotheses and objectives of the study? Are the conclusions consistent with the actual evidence that was produced?**	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
J4	**Was biological relevance of the results clearly discussed?**	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*

#	Appraisal question	Information as reported	Appraisal	Rationale for the appraisal
K.	**Additional considerations**			
K1	**Was the Protocol registered before the starting of the trial and was the trial performed and reported accordingly?**	*Please insert a quotation or a brief summary of what has been reported in the RCT*	□ Definitively appropriate □ Probably appropriate □ Probably not appropriate □ Definitively not appropriate □ Not Applicable	*Please provide the main supporting information for your appraisal*
K2	**Have any competing interests been identified?**	*Please insert a quotation or a brief summary of what has been reported in the RCT*		
K3	**Add here any aspects that should be outlined and are not covered above.**	• Add • Add • etc		

3. References

The RCT CAT was developed using and integrating the following documents:

Altman DG, 1996. Better reporting of randomised controlled trials: the CONSORT statement. BMJ, 313, 570–571.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D and Stroup DF, 1996. Improving the quality of reporting of randomised controlled trials. The CONSORT statement. JAMA, 276, 637–639.

Bornhoft G, Maxion-Bergemann S, Wolf U, Kienle GS, Michalsen A, Vollmar HC, Gilbertson S and Matthiessen PF, 2006. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. BMC Med Res Methodol, 6, 56.

Chalmers TC, Smith H, Jr., Blackburn B, Silverman B, Schroeder B, Reitman D and Ambroz A, 1981. A method for assessing the quality of a randomised control trial. Control Clin Trials, 2, 31–49.

Chalmers TC, Celano P, Sacks HS and Smith H, Jr., 1983. Bias in treatment assignment in controlled clinical trials. N Engl J Med, 309, 1358–1361.

Cho MK and Bero LA, 1996. The quality of drug studies published in symposium proceedings. Ann Intern Med, 124, 485–489.

Colditz GA, Miller JN and Mosteller F, 1989. How study design affects outcomes in comparisons of therapy. I: Medical. Statistics in medicine, 8, 441–454.

Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C and Antes G, 1997. Language bias in randomised controlled trials published in English and German. Lancet, 350, 326–329.

Gluud LL, 2006. Bias in clinical intervention research. Am J Epidemiol, 163, 493–501.

Greenhalgh T, 1997. How to read a paper. Papers that report diagnostic or screening tests. BMJ: British Medical Journal, 315, 540.

Higgins JPT and Green S, 2011. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. www.cochrane-handbook.org

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ and McQuay HJ, 1996. Assessing the quality of reports of randomised clinical trials: is blinding necessary? Control Clin Trials, 17, 1–12.

Moher D, Jadad AR, Nichol G, Penman M, Tugwell P and Walsh S, 1995. Assessing the quality of randomised controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials, 16, 62–73.

Moher D, Jadad AR and Tugwell P, 1996. Assessing the quality of randomised controlled trials. Current issues and future directions. Int J Technol Assess Health Care, 12, 195–208.

Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M and Altman DG, 2010. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ, 340, c869.

Ohlsson A and Lacy JB, 1995. Quality assessments of randomised controlled trials: evaluation by the Chalmers versus the Jadad method. Proceedings of the 3rd Annual Cochrane Colloquium.

Rothwell PM, 2005. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet, 365, 82–93.

Schulz KF, Chalmers I, Hayes RJ and Altman DG, 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA, 273, 408–412.

Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ and Sterne JA, 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ, 336, 601–605.

QUADAS-2: Background Document

QUADAS-2

QUADAS-2 is designed to assess the quality of primary diagnostic accuracy studies; it is not designed to replace the data extraction process of the review and should be applied in addition to extracting primary data (e.g. study design, results etc) for use in the review. It consists of four key domains covering patient selection, index test, reference standard, and flow of patients through the study and timing of the index test(s) and reference standard ("flow and timing") (Table 1). The tool is completed in four phases: 1) state the review question; 2) develop review specific guidance; 3) review the published flow diagram for the primary study or construct a flow diagram if none is reported; 4) judgement of bias and applicability. Each domain is assessed in terms of the *risk of bias* and the first three are also assessed in terms of *concerns regarding applicability*. To help reach a judgement on the risk of bias, *signalling questions* are included. These flag aspects of study design related to the potential for bias and aim to help reviewers make risk of bias judgements.

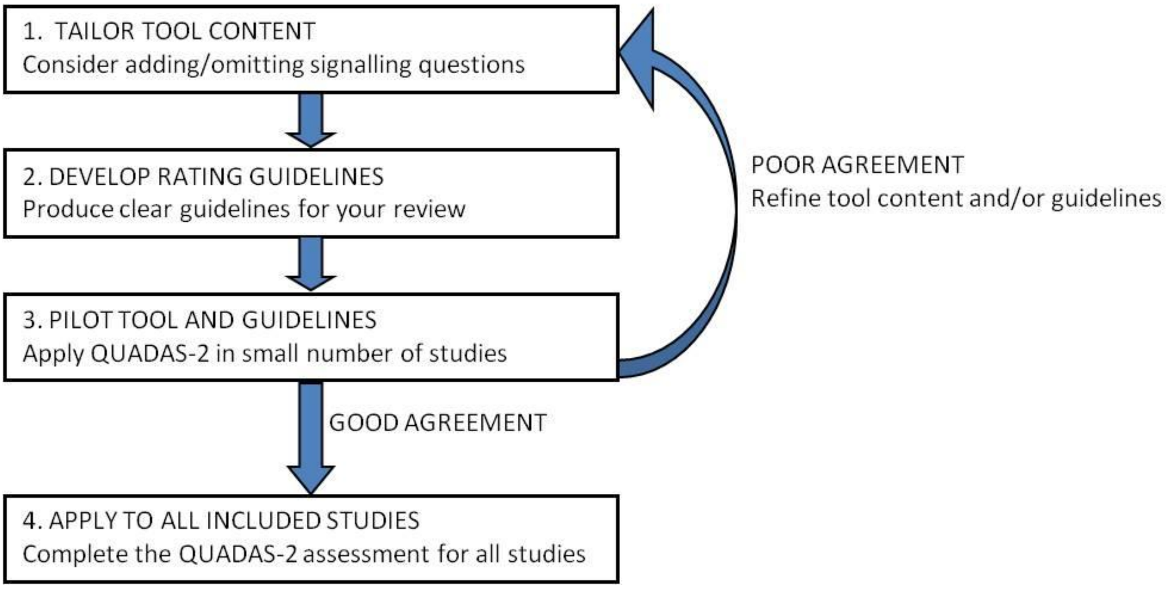
Phase 1: Review Question

Review authors are first asked to report their systematic review question in terms of patients, index test(s), and reference standard and target condition. As the accuracy of a test may depend on where in the diagnostic pathway it will be used, review authors are asked to describe patients in terms of setting, intended use of the index test, patient presentation and prior testing.(1;2)

Phase 2: Review Specific Tailoring (Figure 1)

It is essential to tailor QUADAS-2 to each review by adding or omitting signalling questions and developing review-specific guidance on how to assess each signalling question and use this information to judge the risk of bias. The first step is to consider whether any signalling question does not apply to the review or whether any specific issues for the review are not adequately covered by the core signalling questions. For example, for a review of an objective index test it may be appropriate to omit the signalling question relating to blinding of the test interpreter to results of the reference standard. Review authors should avoid complicating the tool by adding too many signalling questions. Once tool content has been agreed, review-specific rating guidance should be developed. The tool should be piloted independently by at least two people. If agreement is good, the tool can be used to rate all included studies. If agreement is poor, further refinement may be needed.

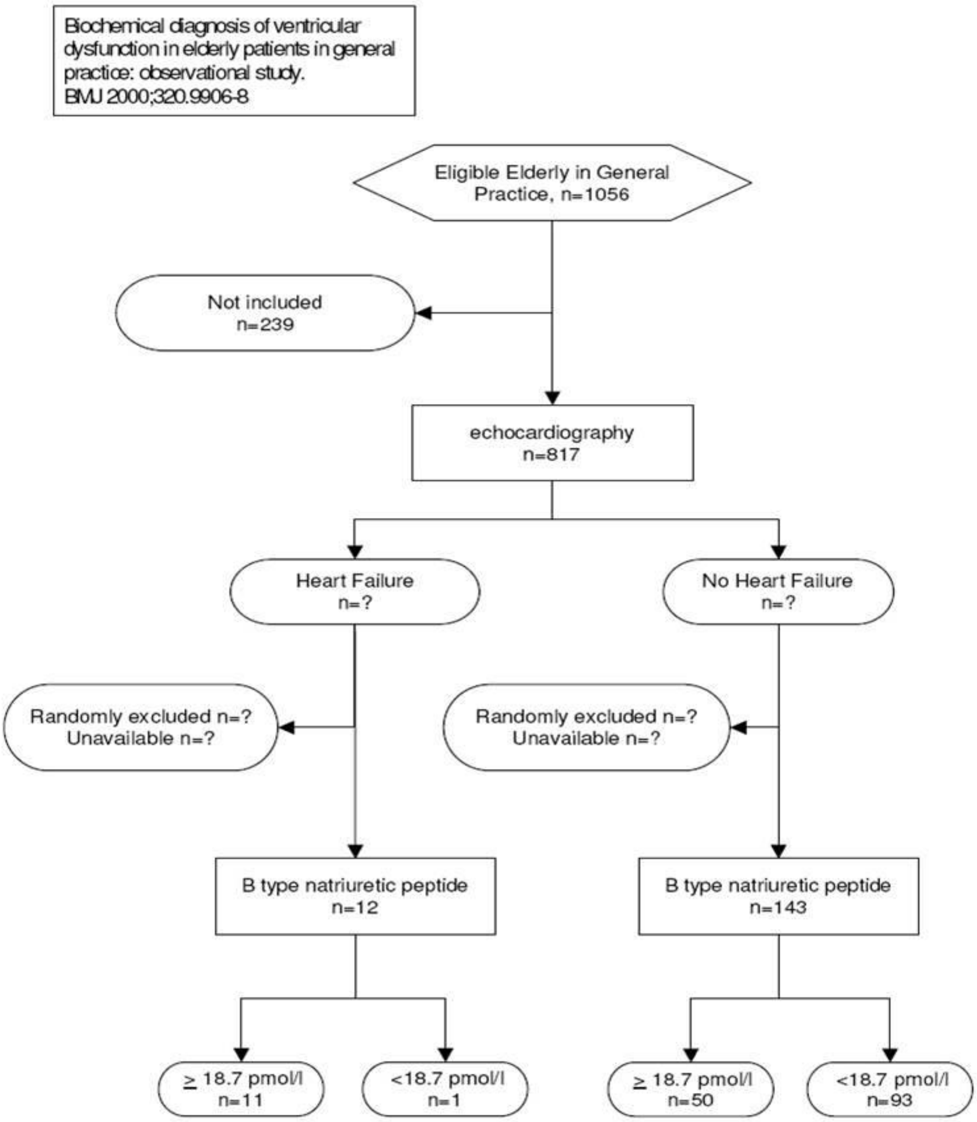
Figure 1: Process for tailoring QUADAS-2 to your systematic review


1. TAILOR TOOL CONTENT
Consider adding/omitting signalling questions
2. DEVELOP RATING GUIDELINES
Produce clear guidelines for your review
3. PILOT TOOL AND GUIDELINES
Apply QUADAS-2 in small number of studies
GOOD AGREEMENT
4. APPLY TO ALL INCLUDED STUDIES
Complete the QUADAS-2 assessment for all studies
POOR AGREEMENT
Refine tool content and/or guidelines

Phase 3: Flow Diagram

The next stage is to review the published flow diagram for the primary study or to draw one if none is reported or the published diagram is not adequate. The flow diagram will facilitate judgments of risk of bias, and should provide information about the method of recruitment of patients (e.g. based on a consecutive series of patients with specific symptoms suspected of having the target condition, or of cases and controls), the order of test execution, and the number of patients undergoing the index test and the reference standard. A hand drawn diagram is sufficient as this step does not need to be reported as part of the QUADAS-2 assessment. Figure 2 shows an example based on a primary study of B type natriuretic peptide for the diagnosis of heart failure.

Figure 2: Flowchart based on diagnostic cohort study of BNP for diagnosing heart failure


Biochemical diagnosis of ventricular dysfunction in elderly patients in general practice: observational study. BMJ 2000;320.9906-8
Eligible Elderly in General Practice, n=1056
Not included n=239
echocardiography n=817
Heart Failure n=?
No Heart Failure n=?
Randomly excluded n=? Unavailable n=?
Randomly excluded n=? Unavailable n=?
B type natriuretic peptide n=12
B type natriuretic peptide n=143
≥ 18.7 pmol/l n=11
<18.7 pmol/l n=1
≥ 18.7 pmol/l n=50
<18.7 pmol/l n=93

Phase 4: Judgments on bias and applicability

Risk of bias

The first part of each domain concerns bias and comprises three sections: 1) information used to support the risk of bias judgment, 2) signalling questions, and 3) judgment of risk of bias. By recording the information used to reach the judgment ("support for judgment"), we aim to make the rating transparent and facilitate discussion between review authors completing assessments independently.(3) The additional signalling questions are included to assist judgments. They are answered as “yes”, “no”, or “unclear”, and are phrased such that “yes” indicates low risk of bias.

Risk of bias is judged as “low”, “high”, or “unclear”. If all signalling questions for a domain are answered “yes” then risk of bias can be judged “low”. If any signalling question is answered “no” this flags the potential for bias. Review authors then need to use the guidelines developed in phase 2 to judge risk of bias. The “unclear” category should be used only when insufficient data are reported to permit a judgment.

Applicability

Applicability sections are structured in a similar way to the bias sections, but do not include signalling questions. Review authors are asked to record the information on which the judgment of applicability is made and then to rate their concern that the study does not match the review question. Concerns regarding applicability are rated as “low”, “high” or “unclear”. Applicability judgments should refer to the first phase, where the review question was recorded. Again, the “unclear” category should only be used when insufficient data are reported.

The following sections provide brief explanations of the signalling questions and risk of bias/concerns regarding applicability questions for each domain.

DOMAIN 1: PATIENT SELECTION

Risk of bias: Could the selection of patients have introduced bias?

Signalling question 1: Was a consecutive or random sample of patients enrolled?

Signalling question 2: Was a case-control design avoided?

Signalling question 3: Did the study avoid inappropriate exclusions?

A study should ideally enrol all consecutive, or a random sample of, eligible patients with suspected disease – otherwise there is potential for bias. Studies that make inappropriate exclusions, e.g. excluding “difficult to diagnose” patients, may result in overoptimistic estimates of diagnostic accuracy. In a review of anti-CCP antibodies for the diagnosis of rheumatoid arthritis, we found that some studies enrolled consecutive patients who had confirmed diagnoses. These studies showed greater sensitivity of the anti-CCP test than studies that included patients with suspected disease but in whom the diagnosis had not been confirmed – “difficult to diagnose” patients.(4) Similarly, studies enrolling patients with known disease and a control group without the condition may exaggerate diagnostic accuracy.(5;6) Exclusion of patients with “red flags” for the target condition, who may be easier to diagnose, may lead to underestimation of diagnostic accuracy.

Applicability: Are there concerns that the included patients and setting do not match the review question?

There may be concerns regarding applicability if patients included in the study differ, compared to those targeted by the review question, in terms of severity of the target condition, demographic features, presence of differential diagnosis or co-morbidity, setting of the study and previous testing protocols. For example, larger tumours are more easily seen with imaging tests than smaller ones, and larger myocardial infarctions lead to higher levels of cardiac enzymes than small infarctions making them easier to detect and so increasing estimates of sensitivity.(7)

DOMAIN 2: INDEX TEST

Risk of Bias: Could the conduct or interpretation of the index test have introduced bias?

Signalling question 1: Were the index test results interpreted without knowledge of the results of the reference standard?

This item is similar to “blinding” in intervention studies. Interpretation of index test results may be influenced by knowledge of the reference standard.(6) The potential for bias is related to the subjectivity of index test interpretation and the order of testing. If the index test is always conducted and interpreted prior to the reference standard,this item can be rated “yes”.

Signalling question 2: If a threshold was used, was it pre-specified?

Selecting the test threshold to optimise sensitivity and/or specificity may lead to overoptimistic estimates of test performance, which is likely to be poorer in an independent sample of patients in whom the same threshold is used.(8)

Applicability: Are there concerns that the index test, its conduct, or interpretation differ from the review question?

Variations in test technology, execution, or interpretation may affect estimates of its diagnostic accuracy. If index tests methods vary from those specified in the review question there may be concerns regarding applicability. For example, a higher ultrasound transducer frequency has been shown to improve sensitivity for the evaluation of patients with abdominal trauma.(9)

DOMAIN 3: REFERENCE STANDARD

Risk of Bias: Could the reference standard, its conduct, or its interpretation have introduced bias?

Signalling question 1: Is the reference standard likely to correctly classify the target condition?

Estimates of test accuracy are based on the assumption that the reference standard is 100% sensitive and specific disagreements between the reference standard and index test are assumed to result from incorrect classification by the index test.(10;11)

Signalling question 2: Were the reference standard results interpreted without knowledge of the results of the index test?

This item is similar to the signalling question related to interpretation of the index test. Potential for bias is related to the potential influence of prior knowledge on the interpretation of the reference standard.(6)

Applicability: Are there concerns that the target condition as defined by the reference standard does not match the question?

The reference standard may be free of bias but the target condition that it defines may differ from the target condition specified in the review question. For example, when defining urinary tract infection the reference standard is generally based on specimen culture but the threshold above which a result is considered positive may vary.(12)

DOMAIN 4: FLOW AND TIMING

Risk of Bias: Could the patient flow have introduced bias?

Signalling question 1: Was there an appropriate interval between index test and reference standard?

Ideally results of the index test and reference standard are collected on the same patients at the same time. If there is a delay or if treatment is started between index test and reference standard, misclassification may occur due to recovery or deterioration of the condition. The length of interval leading to a high risk of bias will vary between conditions. A delay of a few days may not be a problem for chronic conditions, while for acute infectious diseases a short delay may be important. Conversely, when the reference standard involves follow-up a minimum follow-up period may be required to assess the presence or absence of the target condition. For example, for the evaluation of magnetic resonance imaging for the early diagnosis of multiple sclerosis, a minimum follow-up period of around 10 years is required to be confident that all patients who will go on to fulfil diagnostic criteria for multiple sclerosis will have done so.(13)

Signalling question 2: Did all patients receive the same reference standard?

Verification bias occurs when not all of the study group receive confirmation of the diagnosis by the same reference standard. If the results of the index test influence the decision on whether to perform the reference standard or which reference standard is used, estimated diagnostic accuracy may be biased.(5;14) For example, a study evaluating the accuracy of the D-dimer test for the diagnosis of pulmonary embolism carried out ventilation perfusion scans (reference standard 1) in those testing positive and used clinical follow-up to determine whether or not those testing negative had a pulmonary embolism (reference standard 2). This may result in misclassifying some of the false negatives as true negatives as some patients who had a pulmonary embolism but were index test negative may be missed by clinical follow-up and so be classified as not having a pulmonary embolism. This misclassification will overestimate sensitivity and specificity.

Signalling question 3: Were all patients included in the analysis?

All patients who were recruited into the study should be included in the analysis.(15) There is a potential for bias if the number of patients enrolled differs from the number of patients included in the 2x2 table of results, for example because patients lost to follow-up differ systematically from those who remain.

Incorporating QUADAS-2 assessments in diagnostic accuracy reviews

We emphasise that QUADAS-2 should not be used to generate a summary “quality score”, because of the well-known problems associated with such scores.(16;17) If a study is judged as “low” on all domains relating to bias or applicability then it is appropriate to have an overall judgment of “low risk of bias” or “low concern regarding applicability” for that study. If a study is judged "high" or "unclear" on one or more domains then it may be judged “at risk of bias” or as having “concerns regarding applicability”.

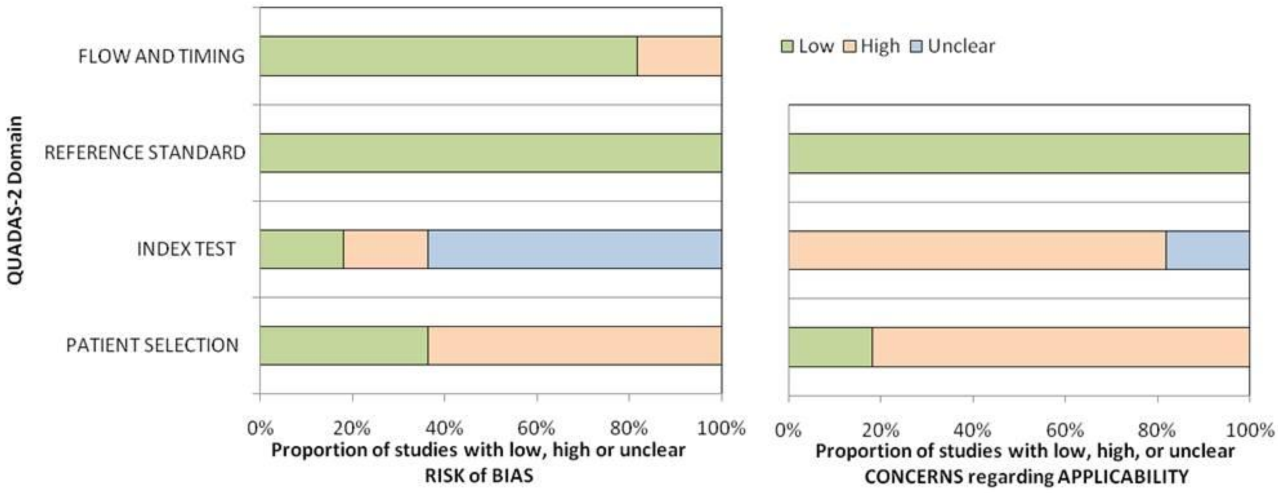
At minimum, reviews should present a summary of the results of the QUADAS-2 assessment for all included studies. This could include summarising the number of studies that found low, high or unclear risk of bias/concerns regarding applicability for each domain. If studies are found to consistently rate well or poorly on particular signalling questions then reviewers may choose to highlight these. Tabular (Table) and graphical (Figure 3) displays are helpful to summarise QUADAS-2 assessments.

Table: Suggested tabular presentation for QUADAS-2 results

Study	RISK OF BIAS				APPLICABILITY CONCERNS		
	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD	FLOW AND TIMING	PATIENT SELECTION	INDEX TEST	REFERENCE STANDARD
Study 1	☺	☺	☺	☺	☹	☺	☺
Study 2	☺	☺	☺	☺	☹	☺	☺
Study 3	☹	☹	☺	☺	☹	☺	☺
Study 4	☹	☹	☺	☺	☹	☺	☺
Study 5	☹	?	☺	☺	☹	☺	☺
Study 6	☹	?	☺	☺	☹	?	☺
Study 7	☹	?	☺	☺	☹	☺	☺
Study 8	☹	?	☺	☺	☹	?	☺
Study 9	☹	?	☺	☺	☹	☺	☺
Study 10	☹	?	☺	☹	☹	☺	☺
Study 11	☺	?	☺	☹	☺	☺	☺

☺Low Risk ☹High Risk ? Unclear Risk

Figure 3: Suggested Graphical Display for QUADAS-2 results



Review authors may choose to restrict the primary analysis so that only studies at low risk of bias and/or low concern regarding applicability for all or specified domains are included. It may be appropriate to restrict inclusion to the review based on similar criteria, but it is often preferable to review all relevant evidence and then investigate possible reasons for heterogeneity.(13;18) Subgroup and or sensitivity analysis can be conducted by investigating how estimates of accuracy of the index test vary between studies rated as high, low, or unclear on all or selected domains. Domains or signalling questions can be included as items in meta-regression analyses, to investigate their association with estimated accuracy.

Website

The QUADAS website (www.quadas.org) contains QUADAS-2, information on training, a bank of additional signalling questions, more detailed guidance for each domain, examples of completed QUADAS-2 assessments, and downloadable resources including a Microsoft Access™ database for data extraction, an Excel™ spreadsheet to produce graphical displays of results, and templates for Word™ tables to summarise results.

References

(1) Bossuyt PM, Leeflang MMG. Chapter 6: Developing Criteria for Including Studies. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration; 2009.

(2) Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149(12):889-897.

(3) Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD et al. The Cochrane Collaboration's tool for assessing risk of bias in randomized trials. BMJ. In press 2011.

(4) Whiting PF, Smidt N, Sterne JA, Harbord R, Burton A, Burke M et al. Systematic review: accuracy of anti-citrullinated Peptide antibodies for diagnosing rheumatoid arthritis. Ann Intern Med 2010; 152(7):456-464.

(5) Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282(11):1061-1066.

(6) Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140(3):189-202.

(7) Reitsma J, Rutjes A, WP, Vlassov V, Leeflang M, Deeks J. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration; 2009.

(8) Leeflang MM, Moons KG, Reitsma JB, Zwinderman AH. Bias in sensitivity and specificity caused by data-driven selection of optimal cutoff values: mechanisms, magnitude, and solutions. Clinical Chemistry 2008; 54(4):729-737.

(9) Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between compliance with methodological standards of diagnostic research and reported test accuracy: meta-analysis of focused assessment of US for trauma. Radiology 2005; 236(1):102-111.

(10) Biesheuvel C, Irwig L, Bossuyt P. Observed differences in diagnostic test accuracy between patient subgroups: is it real or due to reference standard misclassification? Clin Chem 2007; 53(10):1725-1729.

(11) van Rijkom HM, Verdonschot EH. Factors involved in validity measurements of diagnostic tests for approximal caries--a meta-analysis. Caries Research 1995; 29(5)):364-70.

(12) Whiting P, Westwood M, Bojke L, Palmer S, Richardson G, Cooper J et al. Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model. Health Technol Assess 2006; 10(36):iii-xiii, 1.

(13) Whiting P, Harbord R, Main C, Deeks JJ, Filippini G, Egger M et al. Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review. BMJ 2006; 332(7546):875-884.

(14) Rutjes A, Reitsma J, Di NM, Smidt N, Zwinderman A, Van RJ et al. Bias in estimates of diagnostic accuracy due to shortcomings in design and conduct: empirical evidence [abstract]. XI Cochrane Colloquium: Evidence, Health Care and Culture; 2003 Oct 26 31; Barcelona, Spain 2003;45.

(15) Macaskill P, Gatsonis C, Deeks JJ, Harbord R, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration; 20010.

(16) Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999; 282(11):1054-1060.

(17) Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 2005; 5:19.

(18) Whiting PF, Weswood ME, Rutjes AW, Reitsma JB, Bossuyt PN, Kleijnen J et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Medical Research Methodology 2006; 6:9.

QUADAS-2

Phase 1: State the review question:

Patients (setting, intended use of index test, presentation, prior testing):
Index test(s):
Reference standard and target condition:

Phase 2: Draw a flow diagram for the primary study

Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

DOMAIN 1: PATIENT SELECTION

A. Risk of Bias

Describe methods of patient selection:

- ❖ Was a consecutive or random sample of patients enrolled? Yes/No/Unclear
- ❖ Was a case-control design avoided? Yes/No/Unclear
- ❖ Did the study avoid inappropriate exclusions? Yes/No/Unclear

Could the selection of patients have introduced bias? RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability

Describe included patients (prior testing, presentation, intended use of index test and setting)**:**

Is there concern that the included patients do not match the review question? CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 2: INDEX TEST(S)

If more than one index test was used, please complete for each test.

A. Risk of Bias

Describe the index test and how it was conducted and interpreted:

- ❖ Were the index test results interpreted without knowledge of the results of the reference standard? Yes/No/Unclear
- ❖ If a threshold was used, was it pre-specified? Yes/No/Unclear

Could the conduct or interpretation of the index test have introduced bias? RISK: LOW /HIGH/UNCLEAR

B. Concerns regarding applicability

Is there concern that the index test, its conduct, or interpretation differ from the review question? CONCERN: LOW /HIGH/UNCLEAR

DOMAIN 3: REFERENCE STANDARD

A. Risk of Bias

Describe the reference standard and how it was conducted and interpreted:

- ❖ Is the reference standard likely to correctly classify the target condition? Yes/No/Unclear
- ❖ Were the reference standard results interpreted without knowledge of the results of the index test? Yes/No/Unclear

Could the reference standard, its conduct, or its interpretation have introduced bias? **RISK: LOW /HIGH/UNCLEAR**

B. Concerns regarding applicability

Is there concern that the target condition as defined by the reference standard does not match the review question? **CONCERN: LOW /HIGH/UNCLEAR**

DOMAIN 4: FLOW AND TIMING

A. Risk of Bias

Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):

Describe the time interval and any interventions between index test(s) and reference standard:

- ❖ Was there an appropriate interval between index test(s) and reference standard? Yes/No/Unclear
- ❖ Did all patients receive a reference standard? Yes/No/Unclear
- ❖ Did patients receive the same reference standard? Yes/No/Unclear
- ❖ Were all patients included in the analysis? Yes/No/Unclear

Could the patient flow have introduced bias? **RISK: LOW /HIGH/UNCLEAR**

Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance

Edited by Jonathan AC Sterne, Julian PT Higgins, Roy G Elbers and Barney C Reeves
on behalf of the development group for ROBINS-I

Updated 20 October 2016

To cite the ROBINS-I tool: Sterne JAC, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan AW, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Ramsay CR, Regidor D, Rothstein HR, Sandhu L, Santaguida PL, Schünemann HJ, Shea B, Shrier I, Tugwell P, Turner L, Valentine JC, Waddington H, Waters E, Wells GA, Whiting PF, Higgins JPT. ROBINS-I: a tool for assessing risk of bias in non-randomized studies of interventions. *BMJ* 2016; **355**; i4919.

To cite this document: Sterne JAC, Higgins JPT, Elbers RG, Reeves BC and the development group for ROBINS-I. Risk Of Bias In Non-randomized Studies of Interventions (ROBINS-I): detailed guidance, updated 12 October 2016. Available from http://www.riskofbias.info [accessed {date}]

Contents

1 Contributors

(Listed alphabetically within category)

Core group: Julian Higgins, Barney Reeves, Jelena Savović, Jonathan Sterne, Lucy Turner.

Additional core research staff: Roy Elbers, Alexandra McAleenan, Matthew Page.

Bias due to confounding: Nancy Berkman, Miguel Hernán, Pasqualina Santaguida, Jelena Savović, Beverley Shea, Jonathan Sterne, Meera Viswanathan.

Bias in selection of participants into the study: Nancy Berkman, Miguel Hernán, Pasqualina Santaguida, Jelena Savović, Beverley Shea, Jonathan Sterne, Meera Viswanathan.

Bias due to departures from intended interventions: David Henry, Julian Higgins, Peter Jüni, Lakhbir Sandhu, Pasqualina Santaguida, Jonathan Sterne, Peter Tugwell.

Bias due to missing data: James Carpenter, Julian Higgins, Terri Piggott, Hannah Rothstein, Ian Shrier, George Wells.

Bias in measurement of outcomes or interventions: Isabelle Boutron, Asbjørn Hróbjartsson, David Moher, Lucy Turner.

Bias in selection of the reported result: Doug Altman, Mohammed Ansari, Barney Reeves, An-Wen Chan, Jamie Kirkham, Jeffrey Valentine.

Cognitive testing leads: Nancy Berkman, Meera Viswanathan.

Piloting and cognitive testing participants: Katherine Chaplin, Hannah Christensen, Maryam Darvishian, Anat Fisher, Laura Gartshore, Sharea Ijaz, J Christiaan Keurentjes, José López-López, Natasha Martin, Ana Marušić, Anette Minarzyk, Barbara Mintzes, Maria Pufulete, Stefan Sauerland, Jelena Savović, Nandi Seigfried, Jos Verbeek, Marie Wetwood, Penny Whiting.

Other contributors: Belinda Burford, Rachel Churchill, Jon Deeks, Toby Lasserson, Yoon Loke, Craig Ramsay, Deborah Regidor, Jan Vandenbroucke, Penny Whiting.

2 Background

The goal of a systematic review of the effects of an intervention is to determine its causal effects on one or more outcomes. When the included studies are randomized trials, causality can be inferred if the trials are methodologically sound, because successful randomization of a sufficiently large number of individuals should result in intervention and comparator groups that have similar distributions of both observed and unobserved prognostic factors. However, evidence from randomized trials may not be sufficient to answer questions of interest to patients and health care providers, and so systematic review authors may wish to include non-randomized studies of the effects of interventions (NRSIs) in their reviews.

Our ROBINS-I tool ("Risk Of Bias In Non-randomized Studies - of Interventions") is concerned with evaluating the risk of bias (RoB) in the results of NRSIs that compare the health effects of two or more interventions. The types of NRSIs that can be evaluated using this tool are quantitative studies estimating the effectiveness (harm or benefit) of an intervention, which did not use randomization to allocate units (individuals or clusters of individuals) to comparison groups. This includes studies where allocation occurs during the course of usual treatment decisions or peoples' choices: such studies are often called "observational". There are many types of such NRSIs, including cohort studies, case-control studies, controlled before-and-after studies, interrupted-time-series studies and controlled trials in which intervention groups are allocated using a method that falls short of full randomization (sometimes called "quasi-randomized" studies). This document provides guidance for using the ROBINS-I tool specifically for studies with a cohort-type of design, in which individuals who have received (or are receiving) different interventions are followed up over time.

The ROBINS-I tool is based on the Cochrane RoB tool for randomized trials, which was launched in 2008 and modified in 2011 (Higgins et al, 2011). As in the tool for randomized trials, risk of bias is assessed within specified bias domains, and review authors are asked to document the information on which judgements are based. ROBINS-I also builds on related tools such as the QUADAS 2 tool for assessment of diagnostic accuracy studies (Whiting et al, 2011) by providing signalling questions whose answers flag the potential for bias and should help review authors reach risk of bias judgements. Therefore, the ROBINS-I tool provides a systematic way to organize and present the available evidence relating to risk of bias in NRSI.

2.1 Context of the tool

Evaluating risk of bias in a systematic review of NRSI requires both methodological and content expertise. The process is more involved than the process of evaluating risk of bias in randomized trials, and typically involves three stages.

- First, at the planning stage, the review question must be clearly articulated, and important potential problems in NRSI should be identified. This includes a preliminary specification of key confounders (see the discussion below Table 1, and section 4.1) and co-interventions (see section 4.4).
- Second, each study should be carefully examined, considering all the ways in which it might be put at risk of bias. The assessment must draw on the preliminary considerations, to identify important issues that might not have been anticipated. For example, further key confounders, or problems with definitions of interventions, or important co-interventions, might be identified.
- Third, to draw conclusions about the extent to which observed intervention effects might be causal, the studies should be compared and contrasted so that their strengths and weaknesses can be considered jointly. Studies with different designs may present different types of bias, and "triangulation" of findings across these studies may provide assurance either that the biases are minimal or that they are real.

This document primarily addresses the second of these stages, by proposing a tool for assessing risk of bias in a NRSI. Some first-stage considerations are also covered, since these are needed to inform the assessment of each study.

2.2 Assessing risk of bias in relation to a target trial

Both the ROBINS-I tool and the Cochrane RoB tool for randomized trials focus on a study's **internal validity**, For both types of study, we define bias as a tendency for study results to differ systematically from the results expected from a randomized trial, conducted on the same participant group that had no flaws in its conduct. This would typically be a large trial that achieved concealment of randomized allocation; maintained blinding of patients, health care professionals and outcome assessors to intervention received throughout follow up; ascertained outcomes in all randomized participants; and reported intervention effects for all measured outcomes. Defined in this way, bias is distinct from **issues of generalizability (applicability or transportability)** to types of individual who were not included in the study. For example, restricting the study sample to individuals free of comorbidities may limit the utility of its findings because they cannot be generalized to clinical practice, where comorbidities are common.

Evaluations of risk of bias in the results of NRSIs are therefore facilitated by considering each NRSI as an attempt to emulate (mimic) a hypothetical trial. This is the hypothetical pragmatic randomized trial that compares the health effects of the same interventions, conducted on the same participant group and without features putting it at risk of bias (Hernán 2011; Institute of Medicine 2012). **We refer to such a hypothetical randomized trial as a "target" randomized trial** (see section 3.1.1 for more details). Importantly, a target randomized trial need not be feasible or ethical.

ROBINS-I requires that review authors explicitly identify the interventions that would be compared in the target trial that the NRSI is trying to emulate. Often the description of these interventions will require subject-matter knowledge, because information provided by the investigators of the observational study is insufficient to define the target trial. For example, authors may refer to "use of therapy [A]," which does not directly correspond to the intervention "initiation of therapy [A]" that would be tested in an intention-to-treat analysis of the target trial. Meaningful assessment of risk of bias is problematic in the absence of well-defined interventions. For example, it would be harder to assess confounding for the effect of obesity on mortality than for the effect of a particular weight loss intervention (e.g., caloric restriction) in obese people on mortality.

To keep the analogy with the target trial, **this document uses the term "intervention" groups to refer to "treatment" or "exposure" groups in observational studies** even though in such studies no actual intervention was implemented by the investigators.

2.3 Domains of bias

The ROBINS-I tool covers seven domains through which bias might be introduced into a NRSI. These domains provide a framework for considering any type of NRSI, and are summarized in Table 1. The first two domains address issues before the start of the interventions that are to be compared ("baseline") and the third domain addresses classification of the interventions themselves. The other four domains address issues after the start of interventions. For the first three domains, risk of bias assessments for NRSIs are mainly distinct from assessments of randomized trials because randomization protects against biases that arise before the start of intervention. However, randomization does not protect against biases that arise after the start of intervention. Therefore, there is substantial overlap for the last four domains between bias assessments in NRSI and randomized trials.

Variation in terminology between contributors and between research areas proved a challenge to development of ROBINS-I and to writing guidance. 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. Table 1 explains the terms that we have chosen to describe each bias domain, and related terms that are sometimes used. The term *selection bias* is a particular source of confusion. It is often used as a synonym for *confounding* (including in the current Cochrane tool for assessing RoB in randomized trials), which occurs when one or more prognostic factors also predict whether an individual receives one or the other intervention of interest. We restrict our use of the term selection bias to refer to a separate type of bias that occurs when some eligible participants, or the initial follow up time of some participants, or some outcome events, are excluded in a way that leads to the association between intervention and outcome differing from the association that would have been observed in complete follow up of the target trial. *We discourage the use of the term selection bias to refer to confounding*, although we have done this in the past, for example in the context of the RoB tool for randomized trials. Work is in progress to resolve this difference in terminology between the ROBINS-I tool and the current Cochrane tool for assessing RoB in randomized trials.

By contrast with randomized trials, in NRSIs the characteristics of study participants will typically differ between intervention groups. The assessment of the risk of bias arising from uncontrolled confounding is therefore a major component of the ROBINS-I assessment. **Confounding** of intervention effects occurs when one or more prognostic factors (factors that predict the outcome of interest) also predict whether an individual receives one or the other intervention of interest. As an example, consider a cohort study of HIV-infected patients that compares the risk of death from initiation of antiretroviral therapy A versus antiretroviral therapy B. If confounding is successfully controlled, the effect estimates from this observational study will be identical, except for sampling variation, to those from a trial that randomly assigns individuals in the same study population to either intervention A or B. However, failure to control for key confounders may violate the expectation of comparability between those receiving therapies A and B, and thus result in bias. A detailed discussion of assessment of confounding appears in section 4.1

Selection bias may arise when the analysis does not include all of the participants, or all of their follow-up after initiation of intervention, that would have been included in the target randomized trial. The ROBINS-I tool addresses two types of selection bias: (1) bias that arises when either all of the follow-up or a period of follow-up following initiation of intervention is missing for some individuals (for example, bias due to the inclusion of prevalent users rather than new users of an intervention), and (2) bias that arises when later follow-up is missing for individuals who were initially included and followed (for example, bias due to differential loss to follow-up that is affected by prognostic factors).We consider the first type of selection bias under "Bias in selection of participants into the study" (section 4.2), and aspects relating to loss to follow up are covered under "Bias due to missing data" (section 4.5). Examples of these types of bias are given within the relevant sections.

Table 1. Bias domains included in the ROBINS-I tool

Domain	Related terms	Explanation	
Pre-intervention			Pre-intervention or at-intervention domains for which risk of bias assessment is mainly distinct from assessments of randomized trials
Bias due to confounding	Selection bias *as it is sometimes used in relation to clinical trials* (and currently in widespread use within Cochrane); Allocation bias; Case-mix bias; Channelling bias.	Baseline confounding occurs when one or more prognostic variables (factors that predict the outcome of interest) also predicts the intervention received at baseline. ROBINS-I can also address time-varying confounding, which occurs when individuals switch between the interventions being compared and when post-baseline prognostic factors affect the intervention received after baseline.	
Bias in selection of participants into the study	Selection bias *as it is usually used in relation to observational studies and sometimes used in relation to clinical trials*; Inception bias; Lead-time bias; Immortal time bias. Note that this bias specifically excludes lack of external validity, which is viewed as a failure to generalize or transport an unbiased (internally valid) effect estimate to populations other than the one from which the study population arose.	When exclusion of some eligible participants, or the initial follow up time of some participants, or some outcome events, is related to both intervention and outcome, there will be an association between interventions and outcome even if the effects of the interventions are identical. This form of selection bias is distinct from confounding. A specific example is bias due to the inclusion of prevalent users, rather than new users, of an intervention.	
At intervention			
Bias in classification of interventions	Misclassification bias; Information bias; Recall bias; Measurement bias; Observer bias.	Bias introduced by either differential or non-differential misclassification of intervention status. Non-differential misclassification is unrelated to the outcome and will usually bias the estimated effect of intervention towards the null. Differential misclassification occurs when misclassification of intervention status is related to the outcome or the risk of the outcome, and is likely to lead to bias.	

Post-intervention			Post-intervention domains for which there is substantial overlap with assessments of randomized trials
Bias due to deviations from intended interventions	Performance bias; Time-varying confounding	Bias that arises when there are systematic differences between experimental intervention and comparator groups in the care provided, which represent a deviation from the intended intervention(s). Assessment of bias in this domain will depend on the type of effect of interest (either the effect of assignment to intervention or the effect of starting and adhering to intervention).	
Bias due to missing data	Attrition bias; Selection bias *as it is sometimes used in relation to observational studies*	Bias that arises when later follow-up is missing for individuals initially included and followed (e.g. differential loss to follow-up that is affected by prognostic factors); bias due to exclusion of individuals with missing information about intervention status or other variables such as confounders.	
Bias in measurement of outcomes	Detection bias; Recall bias; Information bias; Misclassification bias; Observer bias; Measurement bias	Bias introduced by either differential or non-differential errors in measurement of outcome data. Such bias can arise when outcome assessors are aware of intervention status, if different methods are used to assess outcomes in different intervention groups, or if measurement errors are related to intervention status or effects.	
Bias in selection of the reported result	Outcome reporting bias; Analysis reporting bias	Selective reporting of results in a way that depends on the findings.	

2.4 Study designs

This document relates most closely to NRSIs with **cohort-like designs**, such as cohort studies, quasi-randomized trials and other concurrently controlled studies. Much of the material is also relevant to designs such as case-control studies, cross-sectional studies, interrupted time series and controlled before-after studies, although we are currently considering whether modifications to the signalling questions are required for these other types of studies.

2.5 Risk of bias assessments should relate to a specified intervention effect

This section relates to the effect of intervention that a study aims to quantify. The effect of interest in the target trial will be either

- the effect of **assignment** to the intervention at baseline (start of follow-up), regardless of the extent to which the intervention was received during follow-up (sometimes referred to as the "intention-to-treat" effect in the context of randomized trials); or
- the effect of **starting and adhering** to the intervention as specified in the trial protocol (sometimes referred to as the "per-protocol" effect in the context of randomized trials).

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 *starting and adhering to* the treatment according to a specified protocol, compared with a specified comparator. Review authors need to define the intervention effect of interest to them in each NRSI, and apply the risk of bias tool appropriately to this effect. Issues relating to the choice of intervention effect are discussed in more detail in Section 3.2.2 below.

Note that in the context of ROBINS-I, specification of the intervention effect does not relate to choice of a relative or absolute measures, nor to specific PICO (patient, intervention, comparator, outcome) elements of the review question.

2.6 Structure of this document

Sections 3 and 4 of this document provide detailed guidance on use of ROBINS-I. This includes considerations during the process of writing the review protocol (section 3.1), issues in specifying the effect of interest (section 3.2.2), the use of signalling questions in assessments of risk of bias (section 3.3), the requirement for domain-level bias judgements (section 3.4), how these are used to reach an overall judgement on risk of bias (section 3.5) and the use of outcome-level assessments (section 3.6). Detailed guidance on bias assessments for each domain is provided in Section 4.

3 Guidance for using the tool: general considerations

3.1 At protocol stage

3.1.1 Specifying the research question

The research question follows directly from the objective(s) of the review. It addresses the population, experimental intervention, comparator and outcomes of interest. The comparator could be no intervention, usual care, or an alternative intervention.

A review of NRSI should begin with consideration of what problems might arise, in the context of the research question, in making a causal assessment of the effect of the intervention(s) of interest on the basis of NRSI. It is helpful to think about what is to be studied, why it is to be studied, what types of study are likely to be found, and what problems are likely to be encountered in those studies. Identification of the problems that might arise will be based in part on subject matter experts' knowledge of the literature: the team should also address whether conflicts of interest might affect experts' judgements.

Features of the research question may highlight difficulties in defining the intervention being evaluated in a NRSI, or complexities that may arise with respect to the tools used to measure an outcome domain or the timing of measurements. Ideally, the protocol will specify how the review authors plan to accommodate such complexities in their conduct of the review as well as in preparing for the risk of bias assessment.

3.1.2 Listing the confounding domains relevant to all or most studies eligible for the review

Relevant confounding domains are the prognostic factors that predict whether an individual receives one or the other intervention of interest. They 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. These issues are discussed further in section 4.1.

3.1.3 Listing the possible co-interventions that could differ between intervention groups and have an impact on study outcomes

Relevant co-interventions are the interventions or exposures that individuals might receive after or with initiation of the intervention of interest, which are related to the intervention received and which are prognostic for the outcome of interest. These are also 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. These issues are discussed further in section 4.4.

Box 1: The ROBINS-I tool (Stage 1): At protocol stage

Specify the review question

Participants	
Experimental intervention	
Comparator	
Outcomes	

List the confounding domains relevant to all or most studies

List co-interventions that could be different between intervention groups and that could impact on outcomes

3.2 Preliminary considerations for each study

3.2.1 Specifying a target trial specific to the study

Evaluations of risk of bias are facilitated by considering the NRSI as an attempt to emulate a pragmatic randomized trial, which we refer to as the **target trial**. The first part of a ROBINS-I assessment for a particular study is to specify a target trial (Box 2). The target trial is the hypothetical randomized trial whose results should be the same as those from the NRSI under consideration, in the absence of bias. Its key characteristics are the types of participant (including exclusion/inclusion criteria) and a description of the experimental and comparator interventions. These issues are considered in more detail by Hernán (2001). The differences between the target trial for the individual NRSI and the generic research question of the review relate to issues of heterogeneity and/or generalizability rather than risk of bias.

Because it is hypothetical, ethics and feasibility need not be considered when specifying the target trial. For example there would be no objection to a target trial that compared individuals who did and did not start smoking, even though such a trial would be neither ethical nor feasible in practice.

Selection of a patient group that is eligible for a target trial may require detailed consideration, and lead to exclusion of many patients. For example, Magid et al, (2010) studied the comparative effectiveness of ACE inhibitors compared to beta-blockers as second-line treatments for hypertension. From an initial cohort of 1.6m patients, they restricted the analysis population to (1) persons with incident hypertension, (2) who were initially treated with a thiazide agent, and (3) who had one of the two drugs of interest added as a second agent for uncontrolled hypertension, and (4) who did not have a contraindication to either drug. Their "comparative effectiveness" cohort included 15,540 individuals: less than 1% of the original cohort.

3.2.2 Specifying the effect of interest

In the target trial, the effect of interest for any specific research question will be either the effect of **assignment** to the interventions at baseline, regardless of the extent to which the interventions were received during the follow-up, or the effect of **starting and adhering to** the interventions as specified in the protocol (Box 2). The choice between these effects is a decision of the review authors, and is not determined by the choice of analyses made by authors of the NRSI. However, the analyses of an NRSI may correspond more closely to one of the effects of interest, and therefore be biased with respect to the other one.

In the context of randomized trials, the effect of assignment to intervention can be estimated via an **intention-to-treat (ITT) analysis**, in which participants are analysed in the intervention groups to which they were randomized. In the presence of non-adherence to randomized intervention, an ITT analysis of a placebo-controlled trial underestimates the intervention effect that would have been seen if all participants had adhered to the randomized allocation. 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, and are problematic for non-inferiority or equivalence studies, or for estimating harms.

Patients and other stakeholders are often interested in the effect of starting and adhering to the intervention as described in the trial protocol (sometimes referred to as the **per protocol effect**). This is also the effect that is likely to be of interest when considering adverse (or unintended) effects of interventions. It is possible to use data from randomized trials to estimate the effect of starting and adhering to intervention. However, approaches used to do so in papers reporting on randomized trials are often problematic. In particular, unadjusted analyses based on the treatment actually received, or naïve "per protocol" analyses restricted to individuals in each intervention group who (or the follow up during which they) adhered to the trial protocol can be biased, if prognostic factors influenced treatment received. Advanced statistical methods permit appropriate adjustment for such bias, although applications of such methods are relatively rare. Alternative methods that use randomization status as an instrumental variable bypass the need to adjust for such prognostic factors, but they are not always applicable.

Analogues of these effects can be defined for NRSI. For example, the intention-to-treat effect can be approximated by the effect of *starting* experimental intervention versus *starting* comparator intervention, which corresponds to the intention-to-treat effect in a trial in which participants assigned to an intervention always start that intervention). This differs slightly from the ITT effect in randomized trials, because some individuals randomly assigned to a particular intervention may never initiate it. An analogue of the effect of starting and adhering to the intervention as described in the trial protocol is *starting and adhering to* experimental intervention versus *starting and adhering to* comparator intervention unless medical reasons (e.g. toxicity) indicate discontinuation.

For example, in a study of cancer screening the effect of interest might relate either to receipt (or not) of an invitation to screening (the effect estimated in an ITT analysis of a randomized trial of screening), or to uptake (or not) of an invitation to screening.

For both randomized trials and NRSI, unbiased estimation of the effect of starting and adhering to intervention requires appropriate adjustment for prognostic factors that predict deviations from the intended interventions ("time-varying confounders", see detailed discussion in sections 4.1.9 and 4.4). Review authors should seek specialist advice when assessing intervention effects estimated using methods that adjust for time-varying confounding.

In both randomized trials and NRSI, risk of bias assessments should be in relation to a specified effect of interest. **When the effect of interest is that of assignment to the intervention at baseline (randomized trials) or starting intervention at baseline (NRSI), risk of bias assessments for both types of study need not be concerned with post-baseline deviations from intended interventions** that reflect the natural course of events (for example, a departure from randomized intervention that was clinically necessary because of a sudden worsening of the patient's condition) rather than potentially biased actions of researchers. When the effect of interest is starting and adhering to the intended intervention, risk of bias assessments of both randomized trials and NRSI may have to consider adherence and differences in additional interventions ("co-interventions") between intervention groups. More detailed discussions of these issues are provided in sections 4.1.8, 4.1.9 and 4.4.

3.2.3 Preliminary considerations of confounders and co-interventions

We recommend that the study be examined in detail in two key areas before completing the tool proper (Box 3). These two areas are confounders and co-interventions. The process should determine whether the critical confounders and co-interventions as specified in the protocol were measured or administered in the study at hand, and whether additional confounders and co-interventions were identified in the study. Further guidance and a structure for the assessment is provided in sections 4.1 and 4.4.

Box 2: The ROBINS-I tool (Stage 2, part 1): For each study: setting up the assessment

Specify a target randomized trial specific to the study

Design	Individually randomized / Cluster randomized / Matched (e.g. cross-over)
Participants	
Experimental intervention	
Comparator	

Is your aim for this study…?

☐ to assess the effect of *assignment to* intervention

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

Specify the outcome

Specify which outcome is being assessed for risk of bias (typically from among those earmarked for the Summary of Findings table). Specify whether this is a proposed benefit or harm of intervention.

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.

Box 3: The ROBINS-I tool (Stage 2, part 2): For each study: evaluation of confounding domains and co-interventions

Preliminary consideration of confounders

Complete a row for each important confounding domain (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as potentially important.

"Important" confounding domains are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention. "Validity" refers to whether the confounding variable or variables fully measure the domain, while "reliability" refers to the precision of the measurement (more measurement error means less reliability).

(i) Confounding domains listed in the review protocol				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favour the experimental intervention or the comparator?
			Yes / No / No information	Favour experimental / Favour comparator / No information

(ii) Additional confounding domains relevant to the setting of this particular study, or which the study authors identified as important				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favour the experimental intervention or the comparator?
			Yes / No / No information	Favour experimental / Favour comparator / No information

* In the context of a particular study, variables can be demonstrated not to be confounders and so not included in the analysis: (a) if they are not predictive of the outcome; (b) if they are not predictive of intervention; or (c) because adjustment makes no or minimal difference to the estimated effect of the primary parameter. Note that "no statistically significant association" is not the same as "not predictive"

Preliminary consideration of co-interventions

Complete a row for each important co-intervention (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as important.
"Important" co-interventions are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention.

(i) Co-interventions listed in the review protocol		
Co-intervention	Is there evidence that controlling for this co-intervention was unnecessary (e.g. because it was not administered)?	Is presence of this co-intervention likely to favour outcomes in the experimental intervention or the comparator
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information

(ii) Additional co-interventions relevant to the setting of this particular study, or which the study authors identified as important		
Co-intervention	Is there evidence that controlling for this co-intervention was unnecessary (e.g. because it was not administered)?	Is presence of this co-intervention likely to favour outcomes in the experimental intervention or the comparator
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information

3.3 Signalling questions

A key feature of the tool is the inclusion of signalling questions within each domain of bias. These are reasonably factual in nature and aim 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; and
(5) No information.

One exception to this system is the opening signalling question (1.1, in the assessment of bias due to confounding) does not have a "No information" option.

Some signalling questions are only answered in certain circumstances, for example if the response to a previous question is "Yes" or "Probably yes" (or "No" or "Probably no"). When questions are not to be answered, a response option of "Not applicable" may be selected. Responses underlined in green in the tool are potential markers for low risk of bias, and responses in red are potential markers for a risk of bias. Where questions relate only to sign posts to other questions, no formatting is used.

Responses of "Yes" and "Probably yes" (also of "No" and "Probably no") have similar implications, but allow for a distinction between something that is known and something that is likely to be the case. The former would imply that firm evidence is available in relation to the signalling question; the latter would imply that a judgement has been made. If measures of agreement are applied to answers to the signalling questions, we recommend grouping these pairs of responses.

3.3.1 Free-text boxes alongside signalling questions

There is space for free text alongside each signalling question. This should be used to provide support for each answer. Brief direct quotations from the text of the study report should be used when possible to support responses.

3.4 Domain-level judgements about risk of bias

ROBINS-I is conceived hierarchically: responses to signalling questions (relatively factual, "what happened" or "what researchers did") provide the basis for domain-level judgements about RoB, which then provide the basis for an overall RoB judgement for a particular outcome. Use of the word "judgement" to describe the second and third stages is very important, since the review author needs to consider both the severity of the bias in a particular domain and the relative consequences of bias in different domains. The key to applying the tool is to make domain-level judgements about risk of bias that mean the same across 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 RoB for a particular outcome is relatively straightforward (see 3.5).

Criteria for reaching risk of bias judgements for the seven domains are provided. If none of the answers to the signalling questions for a domain suggest a potential problem then risk of bias for the domain can be judged to be low. Otherwise, potential for bias exists. Review authors must then make a judgement on the extent to which the results of the study are 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 valid conclusions from the study: a serious risk of a very small degree of bias should not be considered "Serious risk" of bias

The "no information" category should be used only when insufficient data are reported to permit a judgment.

The **response options for each domain-level RoB judgement** are:

(1) Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain);
(2) Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial);

(3) Serious risk of bias (the study has some important problems in this domain);
(4) Critical risk of bias (the study is too problematic in this domain to provide any useful evidence on the effects of intervention); *and*
(5) No information on which to base a judgement about risk of bias for this domain.

The "low risk of bias" category exists to emphasize the distinction between randomized trials and non-randomized studies. These distinctions apply in particular to the "pre-intervention" and "at-intervention" domains (see Table 1). In particular, we anticipate that only rarely design features of a non-randomized study will lead to a classification of low risk of bias due to confounding. Randomization does not protect against post-intervention biases, and we therefore expect more overlap between assessments of randomized trials and assessments of NRSI for the post-intervention domains. However other features of randomized trials, such as blinding of participants, health professionals or outcome assessors, may protect against post-intervention biases.

3.4.1 Free-text boxes alongside risk of bias judgements

There is space for free text alongside each RoB judgement to explain the reasoning that underpins the judgement. It is essential that the reasons are provided for any judgements of "Serious" or "Critical" risk of bias.

3.4.2 Direction of bias

It would be highly desirable to know the magnitude and direction of any potential biases identified, but this is considerably more challenging than judging the risk 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, suspicion of selective non-reporting of statistically non-significant results would suggest bias against the null. However, for other domains (in particular confounding, selection bias and forms of measurement bias such as differential misclassification), the bias needs to be thought of not in relation to the null, but as an increase or decrease in the effect estimate (i.e. to favour either the experimental intervention or comparator). For example, confounding bias that decreases the effect estimate would be towards the null if the true risk ratio were greater than 1, and away from the null if the risk ratio were less than 1. **If review authors do not have a clear rationale for judging the likely direction of the bias, they should not attempt to guess it.**

3.5 Reaching an overall judgement about risk of bias

The **response options for an overall RoB judgement** are:

(1) Low risk of bias (the study is comparable to a well-performed randomized trial);
(2) Moderate risk of bias (the study provides sound evidence for a non-randomized study but cannot be considered comparable to a well-performed randomized trial);
(3) Serious risk of bias (the study has some important problems);
(4) Critical risk of bias (the study is too problematic to provide any useful evidence and should not be included in any synthesis); *and*
(5) No information on which to base a judgement about risk of bias

Table 2 shows the basic approach to be used to map RoB judgements within domains to a single RoB judgement across domains for the outcome.

Table 2. Reaching an overall RoB judgement for a specific outcome.

RESPONSE OPTION	CRITERIA
Low risk of bias (the study is comparable to a well-performed randomized trial);	The study is judged to be at **low risk of bias for all domains**.
Moderate risk of bias (the study appears to provide sound evidence for a non-randomized study but cannot be considered comparable to a well-performed randomized trial);	The study is judged to be at **low or moderate risk of bias for all domains**.
Serious risk of bias (the study has some important problems);	The study is judged to be at **serious risk of bias** in at least one domain, but not at critical risk of bias in any domain.
Critical risk of bias (the study is too problematic to provide any useful evidence and should not be included in any synthesis);	The study is judged to be at **critical risk of bias in at least one domain.**
No information on which to base a judgement about risk of bias.	There is no clear indication that the study is at serious or critical risk of bias *and* there is a lack of information in one or more key domains of bias (*a judgement is required for this*).

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 outcome being assessed). Therefore, a judgement of "Serious risk of bias" within any domain should have similar implications for the study as a whole, irrespective of which domain is being assessed.

Because it will be rare that an NRSI is judged as at low risk of bias due to confounding, we anticipate that most NRSI will be judged as at least at moderate overall risk of bias.

The mapping of domain-level judgements to overall judgements described in Table 2 is a programmable algorithm. However, in practice some "Serious" risks of bias (or "Moderate" risks of bias) might be considered to be additive, so that "Serious" risks of bias in multiple domains can lead to an overall judgement of "Critical" risk of bias (and, similarly, "Moderate" risks of bias in multiple domains can lead to an overall judgement of "Serious" risk of bias).

3.6 Assessing risk of bias for multiple outcomes in a review

ROBINS-I addresses the risk of bias in a specific result from a NRSI. The risk of bias in the effect of an intervention may be very different for different analyses of the same outcome (e.g. when different analyses adjust for different confounders), as well as for different outcomes. NRSI included in systematic reviews will frequently (if not usually) contribute results for multiple outcomes, so several risk of bias assessments may be needed for each study. Table 3 shows examples of possible assessments for a hypothetical NRSI that addresses three outcomes, O1 (e.g. mortality), O2 (e.g. viral load) and O3 (e.g. quality of life).

Table 3. Reaching an overall RoB judgement for a specific outcome.

Domain	Assessments by outcome	Comment
Bias due to confounding	O1: Serious risk	e.g. only counts available (no adjustment for confounders)
	O2: Moderate risk	e.g. appropriately adjusted
	O3: Serious risk	e.g. insufficient adjustment
Bias in selection of participants into the study	Grouped (O1, O2, O3): Low risk	e.g. same issues thought to apply to all
Bias in classification of interventions	Grouped (O1, O2, O3): Low risk	e.g. same issues thought to apply to all
Bias due to deviations from intended interventions	Grouped (O1, O2, O3): Moderate risk	e.g. same issues thought to apply to all
Bias due to missing data	O1: Low risk	e.g. everyone followed up through records
	Grouped (O2, O3): No information	e.g. due to attrition; same participants
Bias in measurement of outcomes	Grouped (O1, O2): Low risk	e.g. both objective measures
	O3: Serious risk	e.g. prone to biases due to lack of blind outcome assessment
Bias in selection of the reported result	O1: Moderate risk	e.g. unlikely to be manipulated
	O2: Moderate risk	e.g. unlikely to be manipulated
	O3: Serious risk	e.g. cut-point used without justification

This would give us the RoB profiles (which might accompany meta-analyses and/or GRADE assessments) shown in Table 4.

Table 4. Illustration of different RoB judgements for different outcomes

Domain	O1	O2	O3
Bias due to confounding	Serious risk	Moderate risk	Serious risk
Bias in selection of participants into the study	Low risk	Low risk	Low risk
Bias in classification of interventions	Low risk	Low risk	Low risk
Bias due to deviations from intended interventions	Moderate risk	Moderate risk	Moderate risk
Bias due to missing data	Low risk	No info	No info
Bias in measurement of outcomes	Low risk	Low risk	Serious risk
Bias in selection of the reported result	Moderate risk	Moderate risk	Serious risk
Overall*	*Serious risk*	*Moderate risk*	*Serious risk*

4 Guidance for using the tool: detailed guidance for each bias domain

4.1 Detailed guidance: Bias due to confounding

4.1.1 *Background*

A confounding domain is a pre-intervention prognostic factor that predicts whether an individual receives one or the other intervention of interest. Some common examples are severity of pre-existing disease, presence of comorbidities, health care utilization, adiposity, and socioeconomic status. Confounding domains can be characterised by measuring one or more of a range of specific variables. The relevant confounding domains vary across study settings. For example, socioeconomic status might not introduce confounding in studies conducted in countries in which access to the interventions of interest is universal and therefore socioeconomic status does not influence intervention received.

The tool addresses two types of confounding: baseline confounding and time-varying confounding.

4.1.2 *Baseline confounding*

Baseline confounding occurs when one or more pre-intervention prognostic factors predict the intervention received at start of follow up. A **pre-intervention variable** is one that is measured before the start of interventions of interest. For example, a non-randomized study comparing two antiretroviral drug regimens should control for CD4 cell count measured before the start of antiretroviral therapy, because this is strongly prognostic for AIDS and death and is likely to influence choice of regimen. Baseline confounding is likely to be an issue in most or all NRSI.

4.1.3 *Time-varying confounding*

Time-varying confounding occurs when the intervention received can change over time (for example, if individuals switch between the interventions being compared), and when post-baseline prognostic factors affect the intervention received after baseline. A **post-baseline variable** is one that is measured after baseline: for example CD4 cell count measured 6 months after initiation of therapy. **Time-varying confounding needs to be considered in studies that partition follow-up time for individual participants into time spent in different intervention groups**.

For example, suppose a study of patients treated for HIV partitions follow-up time into periods during which patients were receiving different antiretroviral regimens and compares outcomes during these periods in the analysis. CD4 cell count (as a post-baseline prognostic variable) might influence switches between the regimens of interest. When post-baseline prognostic variables are affected by the interventions themselves (for example, antiretroviral regimen may influence post-baseline CD4 count), conventional adjustment for them in statistical analyses is not appropriate as a means of controlling for confounding. For example, CD4 count measured after start of antiretroviral therapy (a post-baseline prognostic variable) might influence switches between the regimens of interest (Hernán et al, 2002). When post-baseline prognostic variables are affected by the interventions themselves (for example, antiretroviral regimen may influence post-baseline CD4 count), conventional adjustment for them in statistical analyses is not appropriate as a means of controlling for confounding (Hernán et al, 2002; Hernán et al, 2004). Note that when individuals switch between the interventions being compared the effect of interest is that of starting and adhering to intervention, not the effect of assignment to intervention.

As a further example, a large open comparative NRSI compared cardiovascular events in patients while taking a new medication for diabetes with those in control patients while receiving older therapies. Research evidence published during the study's follow up period suggested that the new diabetes medication increased the risk of vascular events. Patients whose blood pressure or lipid levels deteriorated after study entry were switched away from the new drug by physicians concerned about the cardiovascular risk. Because blood pressure and lipid levels were prognostic for cardiovascular events **and** predicted the intervention received after baseline, the study was at risk of bias due to time-varying confounding. These issues are discussed in sections 4.1.8 and 4.1.9.

4.1.4 *Identifying confounding domains*

Important confounding domains should be pre-specified in the protocol of a review of NRSI. The identification of potential confounding domains requires subject-matter knowledge. For example, in an observational study comparing minimally invasive and open surgical strategies, lack of adjustment for pre-intervention fitness for surgery (comorbidity), measured by American Society of Anesthesiologists (ASA) class or Charlson index, would result in confounding if this factor predicted choice of surgical strategy. Experts on surgery are best-placed to identify prognostic factors that are likely to be related to choice of surgical strategy. The procedures described below are therefore designed to be used by raters who have good knowledge of the subject matter under study. **We recommend that subject-matter experts be included in the team writing the review protocol, and encourage the listing of confounding domains in the review protocol, based on initial discussions among the review authors and existing knowledge of the literature.**

It is likely that new ideas relating to confounding and other potential sources of bias will be identified after the drafting of the review protocol, and even after piloting data collection from studies selected for inclusion in the systematic review. For example, such issues may be identified because they are mentioned in the introduction and/or discussion of one or more papers. This could be addressed by explicitly recording whether potential confounders or other sources of bias are mentioned in the paper, as a field for data collection.

For rare or unusual adverse effects the underlying risk factors may not be known, and it may prove difficult to identify sources of confounding beforehand. For instance, nephrogenic systemic fibrosis is a rare, recently discovered adverse event where the aetiological factors and natural history have yet to be elucidated. In this specific situation, review authors may not be able to specify relevant sources of confounding beforehand or to judge if studies assessing this adverse event have adequately addressed confounding. On the other hand, review authors could judge confounding to be implausible if they believed that those assigning interventions were not aware of the possibility of an adverse effect and so unlikely to make treatment decisions based on risk factors for that adverse effect. Note that if the adverse effect is a result of, or correlated with, a known adverse event (for example, poor kidney function in the nephrogenic systemic fibrosis example above) of treatment, then confounding may still be present.

4.1.5 Residual and unmeasured confounding

Because confounding domains may not be directly measured, investigators measure specific variables (often referred to as confounders) in an attempt to fully or partly adjust for these confounding domains. For example, baseline CD4 cell count and recent weight loss may be used to adjust for disease severity; hospitalizations and number of medical encounters in the 6 months preceding baseline may be used to adjust for healthcare utilization; geographic measures to adjust for physician prescribing practices; body mass index and waist-to-hip ratio to adjust for adiposity; and income and education to adjust for socioeconomic status.

We can identify two broad reasons that confounding is not fully controlled. **Residual confounding** occurs when a confounding domain is measured with error, or when the relation between the confounding domain and the outcome or exposure (depending on the analytic approach being used) is imperfectly modelled. For example, in a NRSI comparing two antihypertensive drugs, we would expect residual confounding if pre-intervention blood pressure was measured 3 months before the start of intervention, but the blood pressures used by clinicians to decide between the drugs at the point of intervention were not available in our dataset. **Unmeasured confounding** occurs when a confounding domain has not been measured, or when it is not controlled in the analysis. This would be the case if no pre-intervention blood pressure measurements were available, or if the analysis failed to control for pre-intervention blood pressure despite it being measured.

Note that when intervention decisions are made by health professionals, measurement error in the information available to them does not necessarily introduce residual confounding. For example, pre-intervention blood pressure will not perfectly reflect underlying blood pressure. However, if intervention decisions were made based on two pre-intervention measurements, and these measurements were available in our dataset, it would be possible to adjust fully for the confounding.

For some review questions the confounding may be intractable, because it is not possible to measure all the confounding domains that influence treatment decisions. For example, consider a study of the effect of treating type 2 diabetes with insulin when oral antidiabetic drugs fail. The patients are usually older, and doctors may, without recording their decisions, prescribe insulin treatment mostly to those without cognitive impairment and with sufficient manual dexterity. This creates potentially strong confounding that may not be measurable.

4.1.6 Control of confounding

When all confounders are measured without error, confounding may be controlled either by design (for example by restricting eligibility to individuals who all have the same value of the baseline confounders) or through statistical analyses that adjust ("control") for the confounding factor(s). If, in the context of a particular study, a confounding factor is unrelated to intervention or unrelated to outcome, then there is no need to control for it in the analysis. It is however important to note that in this context "unrelated" means "not associated" (for example, risk ratio close to 1) and does not mean "no statistically significant association".

Appropriate control of confounding requires that the variables used are valid and reliable measures of the confounding domains. In this context, "validity" refers to whether the variable or variables fully measures the domain, while "reliability" refers to the precision of the measurement (more measurement error means less reliability) (Streiner and Norman, 2003). For some topics, a list of valid and reliable measures of confounding domains will be available in advance and should be specified in the review protocol. For other topics, such a list may not be available. Study authors may cite references to support the use of a particular measure: reviewers can then base their judgment of the validity and reliability of the measure based on these citations (Cook and Beckman, 2006). Some authors may control for confounding variables with no indication of their validity or reliability. In such instances, review authors should pay attention to the subjectivity of the measure. Subjective measures based on self-report may tend to have lower validity and reliability relative to objective measures such as clinical reports and lab findings (Cook et al, 1990).

It is important to consider whether inappropriate adjustments were made. In particular, **adjusting for post-intervention variables is usually not appropriate**. Adjusting for **mediating variables** (those on the causal pathway from intervention to outcome) restricts attention to the effect of intervention that does not go via the mediator (the "direct effect") and may introduce confounding, even for randomized trials. Adjusting for **common effects** of intervention and outcome causes bias. For example, in a study comparing different antiretroviral drug combinations it will usually be essential to adjust for pre-intervention CD4 cell count, but it would be inappropriate to adjust for CD4 cell count 6 months after initiation of therapy.

4.1.7 Negative controls

Use of a "**negative control**" – exploration of an alternative analysis in which no association should be observed – can sometimes address the likelihood of unmeasured confounding. Lipsitch *et al* (2010) discussed this issue, and distinguished two types of negative controls: exposure controls and outcome controls. One example discussed by these authors relates to observational studies in elderly persons that have suggested that vaccination against influenza is associated with large reductions in risk of pneumonia/influenza hospitalization and in all-cause mortality. To test this hypothesis, Jackson *et al* (2006) reproduced earlier estimates of the protective effect of influenza vaccination, then repeated the analysis for two sets of negative control outcomes. First, they compared the risk of pneumonia/influenza hospitalization and all-cause mortality in vaccinated and unvaccinated persons before, during, and after influenza season ("exposure control"). They reasoned that if the effect measured in previous studies was causal, it should be most prominent during influenza season. Despite efforts to control for confounding, they observed that the protective effect was actually greatest before, intermediate during, and least after influenza season. They concluded that this is evidence that confounding, rather than protection against influenza, accounts for a substantial part of the observed "protection." Second, they postulated that the protective effects of influenza vaccination, if real, should be limited to outcomes plausibly linked to influenza. They repeated their analysis, but substituted hospitalization for injury or trauma as the end point ("outcome control"). They found that influenza vaccination was also "protective" against injury or trauma hospitalization. This, too, was interpreted as evidence that some of the protection observed for pneumonia/influenza hospitalization or mortality was due to inadequately controlled confounding. A second example of "outcome control" is that studies of smoking and suicide also found an association between smoking and homicide (Davey Smith et al, 1992).

4.1.8 Switches between interventions

In some (perhaps many) NRSI, particularly those based on routinely collected data, the intervention received by participants may change, during follow up, from the intervention that they received at baseline to another of the interventions being compared in the review. This may result in "**switches between interventions of interest**", a phenomenon that we consider here under the confounding domain (see "time-varying confounding" below). If one of the intervention groups being compared is no intervention, then such switches include discontinuation of active intervention, or starting active treatment for individuals assigned to control. On the other hand, change from the baseline intervention may result in switching to an intervention that is not of interest to the review question. We consider switches of this kind under "Deviations from intended intervention".

For studies in which participants switch between interventions, risk of bias assessments will depend on the effect of interest. There are two broad approaches:

1. The effect of interest is the effect of assignment to (or starting) experimental intervention versus assignment to (or starting) comparator intervention) and participants are analysed in groups defined by the initial intervention received. In this circumstance, switches between interventions during follow up do not cause bias. For example, consider a study in which men with screen-detected localized prostate cancer are assigned to either immediate surgery or active monitoring of their cancer. Some men subsequently receive surgery, but they would be analysed according to the initial intervention. As another example, a study examining the effect of women's choice of oral contraceptive on their subsequent risk of breast cancer would include all follow-up time, regardless of whether women stopped using contraception because they wished to conceive.
2. The effect of interest is the effect of starting and adhering to intervention, and follow-up time is split into time during which different interventions were received. For example, in a 12-month study comparing two selective serotonin-reuptake inhibitors (SSRIs) A and B with no intervention, a patient might spend 6 months on A, two months on no intervention and four months on B, and these follow up periods are assigned to the different interventions in the analysis. Such studies and analyses depend on an assumption that the risk of the outcome of interest changes soon after change of intervention: for example the study authors may believe that any change in the risk of venous thrombosis that is associated with the use of a particular oral contraceptive stops soon after the use of that intervention. By contrast, study authors may believe that changes in the risk of breast cancer are sustained for a considerable period after cessation of the oral contraceptive: estimation of "per protocol" effects would then be very difficult because it would be necessary to make strong assumptions about the contributions of previous and current interventions to breast cancer risk during a particular period of follow-up.

4.1.9 Time-varying confounding

When follow-up time is split according to the intervention received we need to assess the risk of bias due to time-varying confounding. If the values of factors that are prognostic for the outcome of interest and predict a switch of intervention also change with time, then adjusting only for baseline confounding is insufficient. For example, in a study comparing the effect of non-steroidal anti-inflammatory drugs (NSAIDS) on mortality, in which participants switched during follow-up between the NSAIDS being compared, time-varying confounding would occur if episodes of gastrointestinal bleeding during follow up were prognostic for mortality and also predicted switches between NSAIDS.

4.1.10 Technical note: adjusting for time-varying confounding

Time-varying confounding occurs when time-varying factors that predict the outcome also affect changes of intervention. If, in addition, past changes to intervention affect subsequent values of the same factors, standard statistical methods (such as Cox regression models including the time-varying factor) are not able to adjust appropriately for the confounding, even if the factor concerned is perfectly measured and its effect perfectly modelled. Studies of antiretroviral therapy (ART) for HIV infection provide an example: CD4 cell count is a prognostic factor for AIDS that might predict adherence to ART, and adherence to ART affects subsequent CD4 counts. In these circumstances, estimation of the effect of continuous intervention is in principle possible, but requires use of methods that can deal with time-varying confounding. A commonly used method is inverse probability weighting of marginal structural models, but their implementation is technically challenging. Specialist advice should be sought for risk of bias assessments of studies employing these methods.

4.1.11 Risk of bias assessment for bias due to confounding

The signalling questions and risk of bias assessments are given in Box 4 and Table 5. If there is potential for confounding, risk of bias judgements should be based on answers to questions 1.4 to 1.6 for studies in which participants remained in their initial intervention group during follow up or for which time-varying confounding is not expected, and on answers to questions 1.7 and 1.8 for studies in which participants switched between interventions of interest and time-varying confounding is expected.

Box 4: The ROBINS-I tool (Stage 2, part 3): Risk of bias due to confounding

Signalling questions	**Elaboration**	**Response options**
1.1 Is there potential for confounding of the effect of intervention in this study? **If N/PN to 1.1:** the study can be considered to be at low risk of bias due to confounding and no further signalling questions need be considered **If Y/PY to 1.1**: determine whether there is a need to assess time-varying confounding:	In rare situations, such as when studying harms that are very unlikely to be related to factors that influence treatment decisions, no confounding is expected and the study can be considered to be at low risk of bias due to confounding, equivalent to a fully randomized trial. There is no NI (No information) option for this signalling question.	Y / PY / PN / N
1.2. Was the analysis based on splitting participants' follow up time according to intervention received? **If N/PN**, answer questions relating to baseline confounding (1.4 to 1.6) **If Y/PY**, go to question 1.3.	If participants could switch between intervention groups then associations between intervention and outcome may be biased by time-varying confounding. This occurs when prognostic factors influence switches between intended interventions.	NA / Y / PY / PN / N / NI
1.3. Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome? **If N/PN**, answer questions relating to baseline confounding (1.4 to 1.6) **If Y/PY**, answer questions relating to both baseline and time-varying confounding (1.7 and 1.8)	If intervention switches are unrelated to the outcome, for example when the outcome is an unexpected harm, then time-varying confounding will not be present and only control for baseline confounding is required.	NA / Y / PY / PN / N / NI

Signalling questions	Elaboration	Response options
Questions relating to baseline confounding only		
1.4. Did the authors use an appropriate analysis method that controlled for all the important confounding domains?	Appropriate methods to control for measured confounders include stratification, regression, matching, standardization, and inverse probability weighting. They may control for individual variables or for the estimated propensity score. Inverse probability weighting is based on a function of the propensity score. Each method depends on the assumption that there is no unmeasured or residual confounding.	NA / Y / PY / PN / N / NI
1.5. **If Y/PY to 1.4**: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?	Appropriate control of confounding requires that the variables adjusted for are valid and reliable measures of the confounding domains. For some topics, a list of valid and reliable measures of confounding domains will be specified in the review protocol but for others such a list may not be available. Study authors may cite references to support the use of a particular measure. If authors control for confounding variables with no indication of their validity or reliability pay attention to the subjectivity of the measure. Subjective measures (e.g. based on self-report) may have lower validity and reliability than objective measures such as lab findings.	NA / Y / PY / PN / N / NI
1.6. Did the authors control for any post-intervention variables that could have been affected by the intervention?	Controlling for post-intervention variables that are affected by intervention is not appropriate. Controlling for mediating variables estimates the direct effect of intervention and may introduce bias. Controlling for common effects of intervention and outcome introduces bias.	NA / Y / PY / PN / N / NI
Questions relating to baseline and time-varying confounding		
1.7. Did the authors use an appropriate analysis method that controlled for all the important confounding domains and for time-varying confounding?	Adjustment for time-varying confounding is necessary to estimate the effect of starting and adhering to intervention, in both randomized trials and NRSI. Appropriate methods include those based on inverse probability weighting. Standard regression models that include time-updated confounders may be problematic if time-varying confounding is present.	NA / Y / PY / PN / N / NI
1.8. **If Y/PY to 1.7**: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?	See 1.5 above.	NA / Y / PY / PN / N / NI

Risk of bias judgement Optional: What is the predicted direction of bias due to confounding?	See Table 5. Can the true effect estimate be predicted to be greater or less than the estimated effect in the study because one or more of the important confounding domains was not controlled for? Answering this question will be based on expert knowledge and results in other studies and therefore can only be completed after all of the studies in the body of evidence have been reviewed. Consider the potential effect of each of the unmeasured domains and whether all important confounding domains not controlled for in the analysis would be likely to change the estimate in the same direction, or if one important confounding domain that was not controlled for in the analysis is likely to have a dominant impact.	Low / Moderate / Serious / Critical / NI Favours experimental / Favours comparator / Unpredictable

Table 5: Reaching risk of bias judgements for bias due to confounding

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	No confounding expected.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Confounding expected, all known important confounding domains appropriately measured and controlled for; *and* (ii) Reliability and validity of measurement of important domains were sufficient, such that we do not expect serious residual confounding.
Serious risk of bias (the study has some important problems)	(i) At least one known important domain was not appropriately measured, or not controlled for; *or* (ii) Reliability or validity of measurement of an important domain was low enough that we expect serious residual confounding.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) Confounding inherently not controllable *or* (ii) The use of negative controls strongly suggests unmeasured confounding.
No information on which to base a judgement about risk of bias for this domain	No information on whether confounding might be present.

4.2 Detailed guidance: Bias in selection of participants into the study

4.2.1 Introduction

Selection bias occurs when some eligible participants, or the initial follow up time of some participants, or some outcome events, are excluded in a way that leads to the association between intervention and outcome differing from the association that would have been observed in the target trial. As explained in section 2.3, **this phenomenon is distinct from that of confounding**, although the term selection bias is sometimes used to mean confounding.

Our use of the term "selection bias" is intended to refer only to **biases that are internal to the study**, and **not to issues of indirectness** (**generalizability**, **applicability** or **transferability** to people who were excluded from the study) (Schunemann et al, 2013). For example, restricting the study sample to individuals free of comorbidities may limit the generalizability of its findings to clinical practice, where comorbidities are common. However it does not bias the estimated effect of intervention in individuals free of comorbidities.

4.2.2 When selection of participants into the study may introduce bias

Selective recruitment of participants into a study does not necessarily cause bias. For example, consider a study that selected (at random) only half of eligible men, but included all eligible women. The effect of intervention in men will be less precisely estimated than if all men had been included, but if the true effect of the intervention in women is the same as in men, the effect estimate will not be biased by the selection.

Selection bias occurs when selection of participants is **related to both intervention and outcome**. For example, studies of folate supplementation to prevent neural tube defects were biased because they were restricted to live births (Hernán et al, 2002). The bias arises because stillbirths and therapeutic abortions (which were excluded from the sample), are related to both the intervention and the outcome (Velie and Shaw, 1996, Hernán et al, 2002).

For the same reason, selection bias can occur when some follow up time is excluded from the analysis. For example, consider the potential for bias when prevalent, rather than new (incident), users of the intervention are included in analyses. This is analogous to starting the follow-up of the target trial some time after the start of intervention, so that some individuals who experienced the outcome after starting the intervention will have been excluded. This is a type of selection bias that has also been termed **inception bias** or **lead time bias**. If participants are not followed from the start of the intervention (inception), as they would be in a randomized trial, then a period of follow up has been excluded, and individuals who experienced the outcome soon after intervention will be missing from analyses. The key problem is that there is no reason to expect the effect of the intervention to be constant over time. Therefore, excluding follow up immediately after intervention may bias the estimated effect either upwards or downwards. Studies that report estimates of the effect of intervention stratified into follow up periods may provide information on the extent to which the effect of intervention varies with time since the start of intervention.

For example, analyses based on prevalent users of a drug may tend to select those who tolerate the drug well: "depletion of the susceptible" will already have taken place. As a result we may underestimate the rate of adverse effects in the intervention group: pharmacoepidemiological studies therefore often specify that there should have been no record of use of the drug in the previous 12 months. For example, there was an apparently increased risk of venous thromboembolism with the newer oral contraceptive progestogens when investigated in NRSI (Ray *et al*, 2003; Suissa *et al*, 2000).

Users of the newer agents had started treatment more recently than users of older agents and the risk of venous thromboembolism is greatest early in the course of treatment. Contemporary methodological standards emphasize the importance both of identifying cohorts of new users of health technologies and of commencing follow-up from the date of the treatment decision, not commencement of treatment, in order to avoid biases like this (Ray *et al*, 2003; Suissa, 2008).

A related bias – **immortal time bias** – occurs when the interventions are defined in such a way that there is a period of follow up during which the outcome cannot occur. For example, a study followed cohorts of subjects with chronic obstructive pulmonary disease or chronic heart failure and considered them to be in two groups according to whether they received telehomecare or standard care. However, to get telehomecare, patients had to survive for several weeks after the index hospitalization: therefore the time between hospitalization and start of telehomecare was "immortal time". Exclusion of this follow up period, and of the deaths that occur during the period, will bias the study towards finding that telehomecare reduces mortality. Comparison with a target trial should facilitate identification of such bias, because in a trial participants would be followed from the time of randomization even if implementation of intervention occurred some time later.

4.2.3 Technical note: adjusting for selection bias

There are analytic approaches to adjust for these types of selection bias and statistical analyses that protect against selection bias. The key issue is whether measured variables that permit meaningful adjustment (for example via inverse-probability-weighting) are available. In many situations this will not be the case, so that these design issues will lead to a classification of serious or critical risk of bias.

4.2.4 Risk of bias assessment for bias in selection of participants into the study

The signalling questions and risk of bias assessments are given in Box 5 and Table 6.

Box 5: The ROBINS-I tool (Stage 2, part 4): Risk of bias in selection of participants into the study

Signalling questions	Elaboration	Response options
2.1. Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention? **If N/PN to 2.1:** go to 2.4	This domain is concerned only with selection into the study based on participant characteristics observed *after* the start of intervention. Selection based on characteristics observed *before* the start of intervention can be addressed by controlling for imbalances between experimental intervention and comparator groups in baseline characteristics that are prognostic for the outcome (baseline confounding).	Y / PY / PN / N / NI
2.2. **If Y/PY to 2.1**: Were the post-intervention variables that influenced selection likely to be associated with intervention?	Selection bias occurs when selection is related to an effect of either intervention or a cause of intervention **and** an effect of either the outcome or a cause of the outcome. Therefore, the result is at risk of selection bias if selection into the study is related to both the intervention and the outcome.	NA / Y / PY / PN / N / NI
2.3 **If Y/PY to 2.2**: Were the post-intervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome?		NA / Y / PY / PN / N / NI
2.4. Do start of follow-up and start of intervention coincide for most participants?	If participants are not followed from the start of the intervention then a period of follow up has been excluded, and individuals who experienced the outcome soon after intervention will be missing from analyses. This problem may occur when prevalent, rather than new (incident), users of the intervention are included in analyses.	Y / PY / PN / N / NI
2.5. **If Y/PY to 2.2 and 2.3, or N/PN to 2.4**: Were adjustment techniques used that are likely to correct for the presence of selection biases?	It is in principle possible to correct for selection biases, for example by using inverse probability weights to create a pseudo-population in which the selection bias has been removed, or by modelling the distributions of the missing participants or follow up times and outcome events and including them using missing data methodology. However such methods are rarely used and the answer to this question will usually be "No".	NA / Y / PY / PN / N / NI
Risk of bias judgement	See Table 6.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to selection of participants into the study?	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 6: Reaching risk of bias judgements in selection of participants into the study

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) All participants who would have been eligible for the target trial were included in the study; *and* (ii) For each participant, start of follow up and start of intervention coincided.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Selection into the study may have been related to intervention and outcome; *and* The authors used appropriate methods to adjust for the selection bias; *or* (ii) Start of follow up and start of intervention do not coincide for all participants; *and* (a) the proportion of participants for which this was the case was too low to induce important bias; *or* (b) the authors used appropriate methods to adjust for the selection bias; *or* (c) the review authors are confident that the rate (hazard) ratio for the effect of intervention remains constant over time.
Serious risk of bias (the study has some important problems)	(i) Selection into the study was related (but not very strongly) to intervention and outcome; *and* This could not be adjusted for in analyses; *or* (ii) Start of follow up and start of intervention do not coincide; *and* A potentially important amount of follow-up time is missing from analyses; *and* The rate ratio is not constant over time.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) Selection into the study was very strongly related to intervention and outcome; *and* This could not be adjusted for in analyses; *or* (ii) A substantial amount of follow-up time is likely to be missing from analyses; *and* The rate ratio is not constant over time.
No information on which to base a judgement about risk of bias for this domain	No information is reported about selection of participants into the study or whether start of follow up and start of intervention coincide.

4.3 Detailed guidance: Bias in classification of interventions

4.3.1 Introduction

Bias may be introduced if intervention status is misclassified. This is seldom a problem in randomized trials and other experimental studies, because interventions are actively assigned by the researcher and their accurate recording is a key feature of the study. However, in observational studies information about interventions allocated or received needs to be collected.

Possible methods for data collection include:

- systematic assessment of patients (clinical examinations, interviews, diagnostic tests);
- administrative or in-house databases (prospective recording of data with no pre-specified purpose);
- extraction from medical records; and
- organizational records or policy documents (e.g. for organizational or public health interventions).

4.3.2 Differential and non-differential misclassification

Misclassification of intervention status may be non-differential or differential. **Non-differential misclassification** is unrelated to the outcome: for example in a comparison of smoke alarm installation with no smoke alarm installation, receipt of intervention may be incompletely recorded so that some people who installed a smoke alarm are incorrectly allocated to the "no alarm" group. Provided that such misclassification is unrelated to subsequent outcomes (e.g. risk of fire-related injury is unrelated to the reasons for failing to identify smoke alarm installation), the misclassification is non-differential and will usually bias the estimated effect of intervention towards the null (no intervention effect or no difference between interventions).

Differential misclassification occurs when misclassifications of intervention status is related to subsequent outcome or to the risk of the outcome. It is important that, wherever possible, interventions are defined and categorized without knowledge of subsequent outcomes. A well-known example of differential misclassification, when this might not be the case, is **recall bias** in a case-control study, whereby knowledge of case-control status affects recall of previous intervention: typically the cases are more likely than controls to recall potentially important events.

Differential misclassification can occur in cohort studies, if information about intervention status is obtained retrospectively. This can happen if the information (or availability of information) on intervention status is influenced by outcomes: for example a cohort study in elderly people in which the outcome is dementia, and participants' recall of past intervention status at study inception was affected by pre-existing mild cognitive impairment. Alternatively, a research assistant may search more diligently for past intervention status when the participant has dementia. Other mechanisms may lead to differential misclassification of intervention status. For instance, information on the vaccination status of children in parts of Africa is collected by examining vaccination cards on periodic visits to family homes, and if no card is found, a child is assumed to be unvaccinated. In some cultures, vaccination cards are destroyed if a child dies. Vaccination status for such children may be differentially misclassified if they are analysed as unvaccinated in studies of the effect of vaccination on mortality. Such problems can be avoided if information about intervention status is collected at the time of the intervention and the information is complete and accessible to those undertaking the NRSI.

4.3.3 Risk of bias assessment for bias in classification of interventions

The signalling questions and risk of bias assessments are given in Box 6 and Table 7.

Box 6: The ROBINS-I tool (Stage 2, part 5): Risk of bias in classification of interventions

Signalling questions	Elaboration	Response options
3.1 Were intervention groups clearly defined?	A pre-requisite for an appropriate comparison of interventions is that the interventions are well defined. Ambiguity in the definition may lead to bias in the classification of participants. For individual-level interventions, criteria for considering individuals to have received each intervention should be clear and explicit, covering issues such as type, setting, dose, frequency, intensity and/or timing of intervention. For population-level interventions (e.g. measures to control air pollution), the question relates to whether the population is clearly defined, and the answer is likely to be 'Yes'.	Y / PY / PN / N / NI
3.2 Was the information used to define intervention groups recorded at the start of the intervention?	In general, if information about interventions received is available from sources that could not have been affected by subsequent outcomes, then differential misclassification of intervention status is unlikely. Collection of the information at the time of the intervention makes it easier to avoid such misclassification. For population-level interventions (e.g. measures to control air pollution), the answer to this question is likely to be 'Yes'.	Y / PY / PN / N / NI
3.3 Could classification of intervention status have been affected by knowledge of the outcome or risk of the outcome?	Collection of the information at the time of the intervention may not be sufficient to avoid bias. The way in which the data are collected for the purposes of the NRSI should also avoid misclassification.	Y / PY / PN / N / NI
Risk of bias judgement	See Table 7.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to measurement of outcomes or 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 in classification of interventions

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) Intervention status is well defined; *and* (ii) Intervention definition is based solely on information collected at the time of intervention.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Intervention status is well defined; *and* (ii) Some aspects of the assignments of intervention status were determined retrospectively.
Serious risk of bias (the study has some important problems)	(i) Intervention status is not well defined; *or* (ii) Major aspects of the assignments of intervention status were determined in a way that could have been affected by knowledge of the outcome.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(Unusual) An extremely high amount of misclassification of intervention status, e.g. because of unusually strong recall biases.
No information on which to base a judgement about risk of bias for this domain	No definition of intervention or no explanation of the source of information about intervention status is reported.

4.4 Detailed guidance: Bias due to deviations from intended interventions

4.4.1 Introduction

We consider in this domain biases that arise **when there are systematic differences between the care provided to experimental intervention and comparator groups, beyond the assigned interventions**. These differences reflect additional aspects of care, or intended aspects of care that were not delivered.

It is important to distinguish between:

(a) deviations from intended intervention that arise because of knowledge of the intervention applied and **because of expectation of finding a difference between experimental intervention and comparator consistent with the hypothesis being tested in the study**. Such deviations are not part of usual practice.

(b) deviations from intended intervention that happen during usual clinical care following the intervention (for example, cessation of a drug intervention because of acute toxicity); and

4.4.2 The importance of the nature of the effect of interest

The extent to which these considerations are associated with bias depends on the nature of the effect of interest.

Deviations of the first type, (a) above, are always of concern. For example, a study compared infection rates after insertion by cardiologists of two different permanent cardiac pacemaker devices. It was not routine to give prophylactic antibiotics at the participating institutions. Blinding was not feasible. Some cardiologists believed that device A would have a higher infection rate than device B and, as a result, administered antibiotics to patients receiving device A more often than to patients receiving device B. These deviations from intended intervention did not reflect usual clinical care (type b) – they reflected cardiologists' expectations of differences in infection rates between the two devices (type a). The result of the study is at risk of bias, whichever the effect of interest.

The importance of deviations of type (b) depends on the nature of the effect of interest. If the goal is the unbiased estimation of the effect of *assignment to* (or starting) intervention, then there will be no bias due to deviation from the intended interventions for deviations of type (b). Specifically, if all deviations from intended intervention are part of usual practice, then we can still evaluate the effect of assignment to intervention, regardless of the actual implementation of the interventions.

On the other hand, if the goal is the unbiased estimation of *starting and adhering* to intervention, then all deviations from the target "protocol" will lead to bias. For example, 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. 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. These deviations from the intended interventions put the result of the study at risk of bias in relation to the effect of starting and adhering to the intended interventions.

4.4.3 Types of deviations from intended intervention

Biases that arise due to deviations from intended interventions are sometimes referred to as **performance biases**. They arise, in both randomized trials and NRSI, when systematic differences between the care provided to experimental intervention and comparator groups **occur after the start of intervention**, and the participant continues (for analysis purposes) to be part of the intended intervention group.

Technical aside: Considerations of the risk of performance bias are thus distinct from confounding. Note that methods that adjust for time-varying confounding may be used to adjust both for switches between interventions of interest (addressed under confounding in ROBINS-I) and for deviations from intended interventions (addressed in this bias domain). Such methods can rely on sufficient data on predictors of switches between or deviations from interventions having been collected.

In randomized trials, performance bias can sometimes be reduced or avoided by **blinding** of participants and healthcare providers. Blinding does not generally occur in NRSI: thus both patients and healthcare providers are typically aware of the interventions that are being implemented.

Knowledge of the intervention assignment may influence the likelihood of **co-interventions** (receipt of interventions other than the studied interventions, whose frequency may differ between intervention groups), compromised fidelity of **implementation** (i.e. failure to implement some or all of the intervention as intended by the health care professionals delivering care during the trial), and **adherence** to the intervention by patients or participants. Failures in implementation or adherence include *contamination* (the inadvertent application of one of the studied interventions in participants intended to receive the other), and *switches* from the intended interventions to other interventions or to none.

4.4.4 Deviations from intended intervention when assessing the effect of starting and adhering to intervention

Consideration of co-interventions, implementation of the intervention and adherence by participants should be assessed only when interest is in the effect of starting and adhering to the intervention.

4.4.4.1 Considerations for co-interventions

Co-interventions are a potentially important source of bias. For example, consider an observational study comparing rates of post-operative infection in patients who received one of two surgical procedures A and B. If antibiotic prophylaxis was provided for patients receiving A but not those receiving B, lower rates of post-operative infection observed in patients receiving A might be attributable to antibiotic prophylaxis rather than to the surgical procedure, and there is a potential for bias. This is not the case if the specified target trial compares intervention A plus antibiotics with intervention B without antibiotics. Similarly, a "pragmatic" target trial might allow for opportunistic use of antibiotics as medically indicated, whereas an "explanatory" trial comparing the effects of A versus B alone might aim for balance in the use of antibiotics.

A co-intervention is defined as a new intervention that is **not** part of intended intervention. It is important to consider what is normal or usual practice for the intended intervention before determining the presence of co-interventions. For example, the normal administration of a drug treatment for diabetes may require monitoring to allow for adjustments to the dose or addition of another drug. These adjustments are therefore not a deviation from the intended intervention. Similarly, addition of other treatments aimed at diabetes control may be pre-specified as part of usual clinical practice in the context of the intended intervention.

In some instances the protocol for the intended intervention specifies that the addition of other treatments is at the discretion of physicians, but such additions differ between the intervention groups. For example, consider a cohort study comparing rates of gastrointestinal ulcers in aspirin users and non-users. The use of proton pump inhibitors (PPIs) to prevent bleeding in those taking aspirin is part of usual practice. If their use is pre-specified in the study protocol then the comparison is of aspirin plus PPIs (as necessary) with non-use of aspirin, and the PPIs should not be considered a co-intervention. However if the study aims to compare aspirin use with no use, then PPIs may be considered a co-intervention because their greater use in the aspirin group leads to an underestimate of the effect of aspirin on gastrointestinal ulcers. Similarly, PPIs cause diarrhoea, and a higher frequency of diarrhoea in aspirin users may be due to proton pump inhibitor co-intervention, rather than the aspirin.

Review authors should make in advance a list of important co-interventions that could differ between intervention groups and could have an impact on study outcomes (see section 3.1.3). They should then consider whether they are likely to be administered in the context of each particular study.

We suggest that review authors consider whether the critical co-interventions are balanced between intervention groups. If effective co-interventions are **not** balanced, performance bias is probably present. However, if the co-interventions are balanced across intervention groups, there is still a risk that intervention groups will differ in their management or behaviour beyond the intervention comparison of interest. This will be the case if a co-intervention interacts with the experimental and comparator intervention in different ways (for example, it enhances the effect of one intervention, but has no effect on the other).

4.4.4.2 Considerations for fidelity of implementation of intended interventions

Reasons that an intervention is not implemented as intended by the study investigators include problems with (i) adherence to protocols by investigators; (ii) technical problems with the intervention if it is complex or relies on operator skill; and (iii) differences in the context of the study. Problems with fidelity of implementation can occur in one or any of the intervention groups. The implementation of the intended treatment protocol and adherence to intervention by study participants cannot always be disentangled, as one may influence the other. For example, an intervention can be administered in a manner that does not encourage adherence by study participants.

An example of unsuccessful implementation of an intervention is provided by a study comparing three complex interventions for adolescents with depression and a history of substance abuse: (1) antidepressant A combined with cognitive behavioural therapy (CBT); (2) drug B combined with CBT; and (3) CBT alone. Although therapists were sent the protocol for the CBT, there was no specific training with regards to the administration of CBT, which led to potential between-therapist differences in the content and coverage of the CBT intervention. Neither was there a checklist for the content of CBT sessions. In addition, different therapists, with differing degrees of experience and professional training with CBT, administered the CBT within each of the intervention groups. These problems mean that the study is at risk of bias due to lack of fidelity of implementation of the intended interventions.

Review authors should consider the details of the intervention with respect to how, when, where and to whom it is applied. Some key features to consider when assessing the risk of bias associated with lack of implementation fidelity with the intended intervention includes evaluation of the following:

a) ***Practitioner:*** characteristics of those administering the intervention (e.g. staff characteristics, level of expertise and training, potential therapeutic alliances) and opportunities for those implementing the intervention to modify the intended protocol (e.g. physician will decrease dose because of potential for adverse events);

b) ***Intervention:*** core components of the intervention or comparator as it was intended to be delivered within the context of the primary study (e.g. content of the intervention); this includes the complexity of the intervention (e.g. it has multiple components that may adversely affect adherence), the sequence and order of how it is delivered, the dosage, duration, or format (e.g. phone follow-up rather than face to face meeting), and operational definitions of treatment;

c) ***Context:*** characteristics of the healthcare setting (e.g. public outpatient versus hospital outpatient), organizational service structure (e.g. managed care or publically funded program), geographical setting (e.g. rural vs urban), and cultural setting and the legal environment where the intervention is implemented.

The assessment aims to determine whether the intervention and comparator were implemented as intended by the study investigators. This should be evaluated for each intervention group. Consider also if the study design and analysis attempts to minimize the impact of inadvertent application of the unintended intervention.

Differences in how patients are monitored may affect the fidelity of the intervention. For instance, in a cohort study assessing adverse effects of spironolactone compared to its non-use, physicians who recognize that spironolactone can increase serum potassium may choose more frequent monitoring of serum potassium with subsequent dose adjustment before serum potassium reaches abnormal levels.

4.4.4.3 *Considerations for adherence to intervention*

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. The likelihood of non-adherence will differ according to the nature of the interventions being evaluated. For example, poor adherence to pharmaceutical interventions can be frequent, and multiple switches between interventions (or between taking and not taking a pharmaceutical) can occur within one individual. However, time-varying non-adherence is unlikely in comparisons of surgical interventions such as heart valves or joint prostheses.

Users of the tool should consider the interventions being compared. Is there a potential for people receiving one intervention to receive more or less than was intended, to stop intervention, or to switch to other interventions? Are multiple switches possible or likely? It is important to consider the overall rates of non-adherence within each group and determine if this may impact on the study outcomes. The threshold of non-adherence likely to impact the outcomes will vary with the type of intervention and the study design.

4.4.4.4 *Technical note: adjustment for departures from intended interventions*

In Section 4.1.9 we briefly described statistical approaches to estimate the effect of starting and adhering to the intervention, allowing for switches between the interventions of interest, using methods that adjust for time-varying confounding. Related methods can be used to allow for deviations from intended interventions. One approach is to **censor follow up at the time that the deviation occurs**. It is necessary to use statistical methods that avoid the bias (technically, this is a type of selection bias) that can result from such censoring, for example through inverse-probability weighting. For example, consider a 12 month study examining the effect of starting and adhering to use of selective serotonin-reuptake inhibitor (SSRI) A versus no intervention. If some patients switched to SSRI B during follow up then there is a risk of performance bias. This might be dealt with by censoring follow up on receipt of SSRI B, but the analysis would then need to allow for the possibility that patients who switch to SSRI B are systematically different from those who remain on SSRI A.

Alternatively, consider a study that aims to estimate the effect of starting and adhering to continuous intervention with A versus continuous intervention with B, in which no participants change from A to B or from B to A after baseline. If a greater proportion of those assigned to A also take concomitant intervention C during parts of the follow up then there is time-varying confounding by C. Providing that the prognostic factors that predict intervention with C are measured over time, methods that adjust for time-varying confounding (see section 4.1.9) can be used to adjust for the bias due to time-varying confounding that is caused by the imbalance in use of intervention C.

Typically for ethical reasons, a study protocol will permit some changes in the intervention based on participants' health status during follow up; this occurs even in trials where randomization occurs. Such changes could include an alteration of the dose or type of intervention that is provided throughout the remainder of the study. For example, in a study comparing two medications to control blood glucose in diabetic patients, there are repeated measurements of glycaemic response over time (e.g. HBA1c, blood pressure) to monitor response to intervention. If during the normal course of treatment these show poor control, then the clinician will alter the medication (for example, increasing medication dose if HBA1c is too high). However glycaemic response may also influence the primary outcome (e.g. cardiovascular mortality). No statistical adjustment is necessary if the protocol for the target trial allows for modification of dose in response to glycaemic index. However specialist statistical methods (see above) are required to deal with the problem of time-varying confounding by glycaemic index in studies estimating of the effect of continuous treatment with the initial medication dose.

4.4.5 Risk of bias assessment for bias due to deviations from intended interventions

The signalling questions and risk of bias assessments are given in Box 7 and Table 8.

We are aware that review authors would find it extremely useful if we could provide guidance on criteria that should be used to judge co-interventions to be "balanced", or the amount of adherence to intervention to be "high". Unfortunately, we do not believe that simple guidance will be generally applicable: a small absolute difference in the numbers of patients receiving an important co-intervention might lead to substantial bias if the co-intervention strongly influenced the outcome and patients in whom the outcome occurred were usually those who received the co-intervention.

We recommend that review teams ensure that judgements of "balanced" co-intervention, "successful" implementation and lack of adherence are applied consistently across the studies included in their review.

Box 7: The ROBINS-I tool (Stage 2, part 6): Risk of bias due to deviations from intended interventions

Signalling questions	Elaboration	Response options
If your aim for this study is to assess the effect of assignment to intervention, answer questions 4.1 and 4.2		
4.1. Were there deviations from the intended intervention beyond what would be expected in usual practice?	Deviations that happen in usual practice following the intervention (for example, cessation of a drug intervention because of acute toxicity) are part of the intended intervention and therefore do not lead to bias in the effect of assignment to intervention. Deviations may 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). Such deviations are not part of usual practice, so may lead to biased effect estimates. However these are not expected in observational studies of individuals in routine care.	Y / PY / PN / N / NI
4.2. **If Y/PY to 4.1**: 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
If your aim for this study is to assess the effect of starting and adhering to intervention, answer questions 4.3 to 4.6		
4.3. 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.	Y / PY / PN / N / NI
4.4. Was the intervention implemented successfully for most participants?	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

Signalling questions	Elaboration	Response options
4.5. 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. We distinguish between analyses where follow-up time after interventions switches (including cessation of intervention) is assigned to (1) the new intervention or (2) the original intervention. (1) is addressed under time-varying confounding, and should not be considered further here.	Y / PY / PN / N / NI
4.6. **If N/PN to 4.3, 4.4 or 4.5**: 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. Specialist advice may be needed to assess studies that used these approaches. If everyone in one group received a co-intervention, adjustments cannot be made to overcome this.	NA / Y / PY / PN / N / NI
Risk of bias judgement	See Table 8.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to deviations from the 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 8: Reaching risk of bias judgements for bias due to deviations from intended interventions

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	**Effect of assignment to intervention:** (i) Any deviations from intended intervention reflected usual practice; *or* (ii) Any deviations from usual practice were unlikely to impact on the outcome. **Effect of starting and adhering to intervention:** 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.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	**Effect of assignment to intervention:** There were deviations from usual practice, but their impact on the outcome is expected to be slight. **Effect of starting and adhering to intervention:** (i) There were deviations from intended intervention, but their impact on the outcome is expected to be slight. *or* (ii) The important co-interventions were not balanced across intervention groups, or there were deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *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.
Serious risk of bias (the study has some important problems)	**Effect of assignment to intervention:** There were deviations from usual practice that were unbalanced between the intervention groups and likely to have affected the outcome. **Effect of starting and adhering to intervention:** (i) The important co-interventions were not balanced across intervention groups, or there were deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *and* (ii) The analysis was not 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.

Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	**Effect of assignment to intervention:** There were substantial deviations from usual practice that were unbalanced between the intervention groups and likely to have affected the outcome. **Effect of starting and adhering to intervention:** (i) There were substantial imbalances in important co-interventions across intervention groups, or there were substantial deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *and* (ii) The analysis was not 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.
No information on which to base a judgement about risk of bias for this domain	No information is reported on whether there is deviation from the intended intervention.

4.5 Detailed guidance: Bias due to missing data

4.5.1 Introduction

Missing data may arise, among other reasons, through attrition (loss to follow up), missed appointments, incomplete data collection and by participants being excluded from analysis by primary investigators. In NRSI, data may be missing for baseline characteristics (including interventions received or baseline confounders), for outcome measurements, for other variables involved in the analysis or a combination of these. A general rule for consideration of bias due to missing data is that we should assume that an analysis using the data we intended to collect (were they available) would produce an unbiased effect estimate, so that we concentrate only on biases that might be introduced by the missing data.

The starting point for considering risk of bias due to missing outcome data is to clarify the nature of the comparison of interest, particularly with regard to the distinction between *assignment to* (or start of) intervention and *starting and adhering to* intervention (see section 3.2.2). For example, the "complete" data set would be different for a comparison between those who were and were not *offered* screening and a comparison between those who did and did not *attend* screening. Therefore the definition of missing data would also be different. In order to consider missing outcome data, it is therefore important that a study sample is clearly defined at the outset. This can be achieved through consideration of the target randomized trial.

4.5.2 Differential missingness

Specific considerations for missing data broadly follow those established for randomized trials and described in the existing Cochrane RoB tool for randomized trials. Differentials in missing data between intervention groups are key, along with the reasons for data being missing. If (i) the proportion of missing data and (ii) the reasons for missing data are similar across intervention groups, then there would typically be only limited bias in the effect estimate, so that risk of bias would be considered low or moderate (see section 3.4). As the proportion of missing data increases, differences in response to intervention may increase concerns about the potential for bias. While (i) can usually be established from the reported data, (ii) is typically a judgement of the review authors. Given this, balance in proportions of missingness across intervention groups alone provides only moderate reassurance about the risk of bias.

4.5.3 Adverse effects

When looking at unintended effects, an important consideration is whether the review authors are satisfied that follow-up has not systematically excluded non-trivial proportions of individuals in whom adverse effects may be prevalent. For instance, if older people drop out (or miss appointments) more, and also have more adverse events, then a large proportion of adverse events may be missing from the analysis. This will not necessarily introduce bias, although bias would result if the older people are more likely to drop out of one intervention group than the other. This might occur, for example, in a comparison of exercise versus crossword puzzles to prevent cognitive decline.

4.5.4 Risk of bias assessment for bias due to missing data

The signalling questions and risk of bias assessments are given in Box 8 and Table 9.

We are aware that review authors would find it extremely useful if we could provide guidance on the extent of missing data that should lead to the conclusion that a result is at moderate or high risk of bias. For example, a criterion of less than 80% completeness of follow up has been used as a threshold in some guidance. Unfortunately, we do not believe that a single threshold can be meaningfully defined: for example a result based on 95% complete outcome data might still be at high risk of bias if the outcome was rare and if reasons for missing outcome data were strongly related to intervention group.

Box 8: The ROBINS-I tool (Stage 2, part 7): Risk of bias due to missing data

Signalling questions	Elaboration	Response options
5.1 Were outcome data available for all, or nearly all, participants?	"Nearly all" should be interpreted as "enough to be confident of the findings", and a suitable proportion depends on the context. In some situations, availability of data from 95% (or possibly 90%) of the participants may be sufficient, providing that events of interest are reasonably common in both intervention groups. One aspect of this is that review authors would ideally try and locate an analysis plan for the study.	Y / PY / PN / N / NI
5.2 Were participants excluded due to missing data on intervention status?	Missing intervention status may be a problem. This requires that the *intended* study sample is clear, which it may not be in practice.	Y / PY / PN / N / NI
5.3 Were participants excluded due to missing data on other variables needed for the analysis?	This question relates particularly to participants excluded from the analysis because of missing information on confounders that were controlled for in the analysis.	Y / PY / PN / N / NI
5.4 **If PN/N to 5.1, or Y/PY to 5.2 or 5.3**: Are the proportion of participants and reasons for missing data similar across interventions?	This aims to elicit whether either (i) differential proportion of missing observations or (ii) differences in reasons for missing observations could substantially impact on our ability to answer the question being addressed. "Similar" includes some minor degree of discrepancy across intervention groups as expected by chance.	NA / Y / PY / PN / N / NI
5.5 **If PN/N to 5.1, or Y/PY to 5.2 or 5.3**: Is there evidence that results were robust to the presence of missing data?	Evidence for robustness may come from how missing data were handled in the analysis and whether sensitivity analyses were performed by the investigators, or occasionally from additional analyses performed by the systematic reviewers. It is important to assess whether assumptions employed in analyses are clear and plausible. Both content knowledge and statistical expertise will often be required for this. For instance, use of a statistical method such as multiple imputation does not guarantee an appropriate answer. Review authors should seek naïve (complete-case) analyses for comparison, and clear differences between complete-case and multiple imputation-based findings should lead to careful assessment of the validity of the methods used.	NA / Y / PY / PN / N / NI
Risk of bias judgement	See Table 9.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to missing 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 data

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) Data were reasonably complete; *or* (ii) Proportions of and reasons for missing participants were similar across intervention groups; *or* (iii) The analysis addressed missing data and is likely to have removed any risk of bias.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) Proportions of and reasons for missing participants differ slightly across intervention groups; *and* (ii) The analysis is unlikely to have removed the risk of bias arising from the missing data.
Serious risk of bias (the study has some important problems)	(i) Proportions of missing participants differ substantially across interventions; *or* Reasons for missingness differ substantially across interventions; *and* (ii) The analysis is unlikely to have removed the risk of bias arising from the missing data; *or* Missing data were addressed inappropriately in the analysis; *or* The nature of the missing data means that the risk of bias cannot be removed through appropriate analysis.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) (Unusual) There were critical differences between interventions in participants with missing data; *and* (ii) Missing data were not, or could not, be addressed through appropriate analysis.
No information on which to base a judgement about risk of bias for this domain	No information is reported about missing data or the potential for data to be missing.

4.6 Detailed guidance: Bias in measurement of outcomes

4.6.1 Introduction

Bias may be introduced if outcomes are misclassified or measured with error. Possible methods for data collection include:

- systematic assessment of patients (clinical examinations, interviews, diagnostic tests);
- administrative or in-house databases (prospective recording of data with no pre-specified purpose); and
- extraction from medical records; and
- organizational records or policy documents (e.g. for organizational or public health outcomes).

4.6.2 Differential and non-differential measurement error

Misclassification or measurement error of outcomes may be non-differential or differential. **Non-differential measurement error** is unrelated to the intervention received. It can be systematic (for example when measurement of blood pressure is consistently 5 units too high in every participant) – in which case it will not affect precision or cause bias; or it can be random (for example when measurement of blood pressure is sometimes too high and sometimes too low in a manner that does not depend on the intervention or the outcome) – in which case it will affect precision without causing bias.

Differential measurement error is measurement error related intervention status. It will bias the intervention-outcome relationship. This is often referred to as **detection bias**. Examples of situations in which detection bias can arise are (i) if outcome assessors are aware of intervention status (particularly when the outcome is subjective); (ii) different methods (or intensities of observation) are used to assess outcomes in the different intervention groups; and (iii) measurement errors are related to intervention status (or to a confounder of the intervention-outcome relationship).

Blinding of outcome assessors aims to prevent systematic differences in measurements between intervention groups. However, blinding is frequently not possible or not performed for practical reasons. It is also much less frequent in NRSI than in randomized trials.

The signalling questions include consideration of the comparability of data collection methods and of whether measurement errors may be related to intervention status. If data collection methods are very well standardized, the risk of bias may be lower. It is important also to consider the intensity of investigation across intervention groups. For example, in a study evaluating whether corticosteroid drugs are associated with an increased risk of type 2 diabetes mellitus, diagnosis of type 2 diabetes mellitus may be based on regular blood tests. If patients taking corticosteroids have more frequent blood tests than non-users of corticosteroids (possibly because of their underlying disease), then diabetes is more likely to be detected, introducing a bias against corticosteroids. Another example of detection bias despite standardized assessment of a reasonably objective outcome measures relates to the use of a regular size cuff for measuring blood pressure, which may overestimate the true blood pressure in obese patients. If intervention is also related to body mass then the measurement error will introduce bias, and this bias may be present even if outcomes are measured blind to intervention status.

4.6.3 Risk of bias assessment for bias in measurement of outcomes

The signalling questions and risk of bias assessments are given inBox 4 Box 9 and Table 10.

Box 9: The ROBINS-I tool (Stage 2, part 8): Risk of bias in measurement of outcomes

Signalling questions	Elaboration	Response options
6.1 Could the outcome measure have been influenced by knowledge of the intervention received?	Some outcome measures involve negligible assessor judgment, e.g. all-cause mortality or non-repeatable automated laboratory assessments. Risk of bias due to measurement of these outcomes would be expected to be low.	Y / PY / PN / N / NI
6.2 Were outcome assessors aware of the intervention received by study participants?	If outcome assessors were blinded to intervention status, the answer to this question would be 'No'. In other situations, outcome assessors may be unaware of the interventions being received by participants despite there being no active blinding by the study investigators; the answer this question would then also be 'No'. In studies where participants report their outcomes themselves, for example in a questionnaire, the outcome assessor is the study participant. In an observational study, the answer to this question will usually be 'Yes' when the participants report their outcomes themselves.	Y / PY / PN / N / NI
6.3 Were the methods of outcome assessment comparable across intervention groups?	Comparable assessment methods (i.e. data collection) would involve the same outcome detection methods and thresholds, same time point, same definition, and same measurements.	Y / PY / PN / N / NI
6.4 Were any systematic errors in measurement of the outcome related to intervention received?	This question refers to differential misclassification of outcomes. Systematic errors in measuring the outcome, if present, could cause bias if they are related to intervention or to a confounder of the intervention-outcome relationship. This will usually be due either to outcome assessors being aware of the intervention received or to non-comparability of outcome assessment methods, but there are examples of differential misclassification arising despite these controls being in place.	Y / PY / PN / N / NI
Risk of bias judgement	See Table 10.	Low / Moderate / Serious / Critical / NI
Optional: What is the predicted direction of bias due to measurement of outcomes?	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 10: Reaching risk of bias judgements for bias in measurement of outcomes

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	(i) The methods of outcome assessment were comparable across intervention groups; *and* (ii) The outcome measure was unlikely to be influenced by knowledge of the intervention received by study participants (i.e. is objective) or the outcome assessors were unaware of the intervention received by study participants; *and* (iii) Any error in measuring the outcome is unrelated to intervention status.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) The methods of outcome assessment were comparable across intervention groups; *and* (ii) The outcome measure is only minimally influenced by knowledge of the intervention received by study participants; *and* (iii) Any error in measuring the outcome is only minimally related to intervention status.
Serious risk of bias (the study has some important problems)	(i) The methods of outcome assessment were not comparable across intervention groups; *or* (ii) The outcome measure was subjective (i.e. vulnerable to influence by knowledge of the intervention received by study participants); *and* The outcome was assessed by assessors aware of the intervention received by study participants; *or* (iii) Error in measuring the outcome was related to intervention status.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	The methods of outcome assessment were so different that they cannot reasonably be compared across intervention groups.
No information on which to base a judgement about risk of bias for this domain	No information is reported about the methods of outcome assessment.

4.7 Detailed guidance: Bias in selection of the reported result

4.7.1 Introduction

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 treatment) 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).

4.7.2 Exclusion of outcome non-reporting bias from the risk of bias tool

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 (Kirkham et al, 2010). Outcome reporting bias (ORB) arises when the outcome domain is not reported or partially reported based on the direction, magnitude or statistical significance of its association with intervention group. The presence of such bias in one or more of the studies included in a systematic review puts the treatment effect estimate reported by the systematic review at risk of bias (usually in the direction of exaggeration of the magnitude of effect).

The proposed new structure of the RoB tool considers this kind of selective outcome reporting as analogous to publication bias. Therefore, it is proposed to appraise this kind of selective outcome reporting using a different mechanism (e.g. as part a GRADE assessment in the Summary of Findings), not as part of the RoB tool. **This is a notable departure from the current Cochrane RoB tool for randomized trials**. We therefore do not include signalling questions for selective non-reporting (or insufficient reporting) of outcome domains in this document. We recommend the Kirkham et al (2010) framework for considering this kind of selective outcome reporting.

4.7.3 Selective reporting of a result contributing to the synthesis

We consider here the selective reporting of **fully reported results**, that is results that are sufficiently reported to allow the estimate 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; (ii) **selective reporting of a particular analysis** from multiple analyses of a specific outcome measurement; and (iii) **selective reporting of a subset of the participants**. These types of selective reporting put effect estimates from individual primary studies at risk of bias in the same way as other bias domains considered in the ROBINS-I tool. Selective reporting will lead to bias if it is based on the direction, magnitude or statistical significance of intervention effect estimates.

Selective outcome reporting occurs when the effect estimate for an outcome measurement was selected from among analyses of multiple outcome measurements for the outcome domain. Examples include: use of multiple measurement instruments (e.g. pain scales) and reporting only the most favourable result; reporting only the most favourable subscale (or a subset of subscales) for an instrument when measurements for other subscales were available; reporting only one or a subset of time points for which the outcome was measured.

Selective analysis reporting occurs when results are selected from intervention effects estimated in multiple ways: e.g. carrying out analyses of both change scores and post-intervention scores adjusted for baseline; multiple analyses of a particular measurement with and without transformation; 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 with different cut-points; multiple composite outcomes analysed for one outcome domain, but results were reported only for one (or a subset) of the composite outcomes. (Reporting an effect estimate for an unusual composite outcome might be evidence of such selective reporting.)

Selection of a subgroup from a larger cohort: The cohort for analysis may have been selected from a larger cohort for which data were available on the basis of a more interesting finding. Subgroups defined in unusual ways (e.g. an unusual classification of subgroups by dose or dose frequency) may provide evidence of such selective reporting.

Selective reporting can arise for both harms and benefits, although the motivations (and direction of bias) underlying selective reporting of effect estimates for harms and benefits may differ. Selective reporting 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.

These types of selective reporting apply to all cohort study designs, irrespective of whether they involve clustering.

Selective reporting is more likely to arise in studies which have exploratory objectives because, by their nature, such studies often involve inspecting many associations between multiple interventions or multiple outcomes. However, an exploratory study that fully reported all associations investigated would not be at risk of selective reporting: it is selective reporting that it is the problem, not the exploratory nature of the objective *per se*.

4.7.4 *Evidence of selective reporting*

Papers can provide evidence of selective reporting in many ways, too numerous to catalogue. Congruence between outcome measurements and analyses specified in a protocol or statistical analysis plan, before analyses were carried out, is required in order to assign low risk of bias.

Indirect evidence that selective reporting may not be a serious problem can be gleaned from: consistency (not as strong a requirement as congruence) between the reported outcome measurements and analyses and an a priori plan, or clearly defined outcome measurements and analyses that are internally consistent across Methods and Results in the paper, and externally consistent with other papers reporting the study. To assign moderate risk of bias there should also be no indication of selection of the reported analysis from among multiple analyses and no indication of selection of the cohort or subgroups for analysis and reporting on the basis of the results.

Inconsistency (internally or externally) in outcome measurements, analyses or analysis cohorts (e.g. a large difference between the size of cohort of eligible participants and the size of the cohort analysed) should indicate a serious risk of selective reporting, especially if all reported results are statistically significant. Some circumstances increase the risk of selective reporting from among multiple analyses, e.g. substantial imbalance in prognostic variables at baseline, without describing a strategy to minimize this risk (e.g. criteria for including covariates in a multiple regression model).

Direct proof or strong suspicion of selective reporting (indicative of *critical risk of bias*, see below) can sometimes be found in the text of a paper. Examples of the kinds of statements that should cause alarm include: (a) “the results for outcome X [relevant to the systematic review outcome domain D] were more favourable than for outcome Y [also relevant to the same systematic review outcome domain D]”; (b) “various cut-off criteria for dichotomizing/classifying a continuous variable were ‘tried out’”; (c) “change scores were also analysed but not reported because the effect was not significant”. The specific text provoking a judgement of critical bias must be recorded in the free text box.

4.7.5 *Risk of bias assessment for bias in selection of the reported result*

The signalling questions and risk of bias assessments are given in Box 10 and Table 11.

Box 10: The ROBINS-I tool (Stage 2, part 9): Risk of bias in selection of the reported result

Signalling questions	Elaboration	Response options
Is the reported effect estimate likely to be selected, on the basis of the results, from...		
7.1. ... multiple outcome *measurements* within the outcome domain?	For a specified outcome domain, it is possible to generate multiple effect estimates for different measurements. If multiple measurements were made, but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
7.2 ... multiple *analyses* of the intervention-outcome relationship?	Because of the limitations of using data from non-randomized studies for analyses of effectiveness (need to control confounding, substantial missing data, etc), analysts may implement different analytic methods to address these limitations. Examples include unadjusted and adjusted models; use of final value vs change from baseline vs analysis of covariance; different transformations of variables; a continuously scaled outcome converted to categorical data with different cut-points; different sets of covariates used for adjustment; and different analytic strategies for dealing with missing data. Application of such methods generates multiple estimates of the effect of the intervention versus the comparator on the outcome. If the analyst does not pre-specify the methods to be applied, and multiple estimates are generated but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
7.3 ... different *subgroups*?	Particularly with large cohorts often available from routine data sources, it is possible to generate multiple effect estimates for different subgroups or simply to omit varying proportions of the original cohort. If multiple estimates are generated but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
Risk of bias judgement	See Table 11.	Low / Moderate / Serious / Critical / NI
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 11: Reaching risk of bias judgements for bias in selection of the reported result

Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	There is clear evidence (usually through examination of a pre-registered protocol or statistical analysis plan) that all reported results correspond to all intended outcomes, analyses and sub-cohorts.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial)	(i) The outcome measurements and analyses are consistent with an *a priori* plan; or are clearly defined and both internally and externally consistent; *and* (ii) There is no indication of selection of the reported analysis from among multiple analyses; *and* (iii) There is no indication of selection of the cohort or subgroups for analysis and reporting on the basis of the results.
Serious risk of bias (the study has some important problems)	(i) Outcomes are defined in different ways in the methods and results sections, or in different publications of the study; *or* (ii) There is a high risk of selective reporting from among multiple analyses; *or* (iii) The cohort or subgroup is selected from a larger study for analysis and appears to be reported on the basis of the results.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention)	(i) There is evidence or strong suspicion of selective reporting of results; *and* (ii) The unreported results are likely to be substantially different from the reported results.
No information on which to base a judgement about risk of bias for this domain.	There is too little information to make a judgement (for example if only an abstract is available for the study).

5 References

Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine* 2006; **119**: 166.e7-166.e16.

Cook TD, Campbell DT, Peracchio L. Quasi experimentation. In: Dunnette MD, Hough LM (Eds.), *Handbook of Industrial and Organizational Psychology* (pp. 491-567). Palo Alto, CA, US: Consulting Psychologists Press, 1990.

Davey Smith G, Phillips AN, Neaton JD. Smoking as "independent" risk factor for suicide: illustration of an artifact from observational epidemiology? *Lancet* 1992; **340**: 709-12.

Feinstein AR. An additional basic science for clinical medicine: II. The limitations of randomized trials. *Annals of Internal Medicine* 1983;99:544-50.

Hernán MA. With great data comes great responsibility. Publishing comparative effectiveness research in Epidemiology [editorial]. *Epidemiology* 2011; **22**: 290-291.

Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15(5):615-25.

Hernán MA, Hernandez-Diaz S, Werler MM, et al, Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155(2):176-84.

Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JAC, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011; **343**: d5928.

Institute of Medicine. *Ethical and Scientific Issues in Studying the Safety of Approved Drugs*. Washington, DC: The National Academies Press, 2012.

Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *International Journal of Epidemiology* 2006; **35**: 337–344.

Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ* 2010; **340**: c365.

Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010; **21**: 383-8.

Magid DJ, Shetterly SM, Margolis KL, Tavel HM, O'Connor PJ, Selby JV, Ho PM. Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blocker as second-line therapy for hypertension. *Circulation: Cardiovascular Quality and Outcomes* 2010; **3**: 453-458.

Majeed AW, Troy G, Nicholl JP, et al, Randomized, prospective, single-blind comparison of laparascopic versus small-incision cholesystectomy. *Lancet* 1996; 347: 989-94.

McMahon AJ, Russell IT, Baxter JN, et al, Laparascopic versus minilaparotomy cholesystectomy: a randomised trial. *Lancet* 1994; **343**: 135-8.

Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *American Journal of Epidemiology* 2003;158(9):915-20.

Schünemann HJ, Tugwell P, Reeves BC, et al, Non-randomized studies as a source of complementary, sequential or replacement evidence for randomized controlled trials in systematic reviews on the effects of interventions. *Research Synthesis Methods* 2013; **4**: 49-62.

Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use*. 3rd ed. New York: Oxford University Press, 2003.

Suissa S, Spitzer WO, Rainville B, et al, Recurrent use of newer oral contraceptives and the risk of venous thromboembolism. *Human Reproduction* 2000;15(4):817-21.

Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167(4):492-9.

Rothman KJ and Greenland S. *Modern Epidemiology*. 2nd ed. Philadelphia: Lippincott Williams & Wilkins 1998.

Velie EM, Shaw GM. Impact of prenatal diagnosis and elective termination on prevalence and risk estimates of neural tube defects in California, 1989-1991. *American Journal of Epidemiology* 1996;144(5):473-9.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011; **155**: 529-36.

The Risk Of Bias In Non-randomized Studies – of Interventions (ROBINS-I) assessment tool

(version for cohort-type studies)

Developed by: Jonathan AC Sterne, Miguel A Hernán, Barnaby C Reeves, Jelena Savović, Nancy D Berkman, Meera Viswanathan, David Henry, Douglas G Altman, Mohammed T Ansari, Isabelle Boutron, James Carpenter, An-Wen Chan, Rachel Churchill, Asbjørn Hróbjartsson, Jamie Kirkham, Peter Jüni, Yoon Loke, Terri Pigott, Craig Ramsay, Deborah Regidor, Hannah Rothstein, Lakhbir Sandhu, Pasqualina Santaguida, Holger J Schünemann, Beverly Shea, Ian Shrier, Peter Tugwell, Lucy Turner, Jeffrey C Valentine, Hugh Waddington, Elizabeth Waters, Penny Whiting and Julian PT Higgins

Version 1 August 2016

ROBINS-I tool (Stage I): At protocol stage

Specify the review question

Participants	
Experimental intervention	
Comparator	
Outcomes	

List the confounding domains relevant to all or most studies

List co-interventions that could be different between intervention groups and that could impact on outcomes

ROBINS-I tool (Stage II): For each study

Specify a target randomized trial specific to the study

Design	Individually randomized / Cluster randomized / Matched (e.g. cross-over)
Participants	
Experimental intervention	
Comparator	

Is your aim for this study…?

□ to assess the effect of *assignment to* intervention

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

Specify the outcome

Specify which outcome is being assessed for risk of bias (typically from among those earmarked for the Summary of Findings table). Specify whether this is a proposed benefit or harm of intervention.

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.

Preliminary consideration of confounders

Complete a row for each important confounding domain (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as potentially important.

"Important" confounding domains are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention. "Validity" refers to whether the confounding variable or variables fully measure the domain, while "reliability" refers to the precision of the measurement (more measurement error means less reliability).

(i) Confounding domains listed in the review protocol				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favour the experimental intervention or the comparator?
			Yes / No / No information	Favour experimental / Favour comparator / No information

(ii) Additional confounding domains relevant to the setting of this particular study, or which the study authors identified as important				
Confounding domain	Measured variable(s)	Is there evidence that controlling for this variable was unnecessary?*	Is the confounding domain measured validly and reliably by this variable (or these variables)?	OPTIONAL: Is failure to adjust for this variable (alone) expected to favour the experimental intervention or the comparator?
			Yes / No / No information	Favour experimental / Favour comparator / No information

* In the context of a particular study, variables can be demonstrated not to be confounders and so not included in the analysis: (a) if they are not predictive of the outcome; (b) if they are not predictive of intervention; or (c) because adjustment makes no or minimal difference to the estimated effect of the primary parameter. Note that "no statistically significant association" is not the same as "not predictive".

Preliminary consideration of co-interventions

Complete a row for each important co-intervention (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as important.

"Important" co-interventions are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention.

(i) Co-interventions listed in the review protocol		
Co-intervention	Is there evidence that controlling for this co-intervention was unnecessary (e.g. because it was not administered)?	Is presence of this co-intervention likely to favour outcomes in the experimental intervention or the comparator
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information

(ii) Additional co-interventions relevant to the setting of this particular study, or which the study authors identified as important		
Co-intervention	Is there evidence that controlling for this co-intervention was unnecessary (e.g. because it was not administered)?	Is presence of this co-intervention likely to favour outcomes in the experimental intervention or the comparator
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information
		Favour experimental / Favour comparator / No information

Risk of bias assessment (cohort-type studies)

Responses underlined in green are potential markers for low risk of bias, and responses in red are potential markers for a risk of bias. Where questions relate only to sign posts to other questions, no formatting is used.

Bias domain	Signalling questions	Elaboration	Response options
Bias due to confounding	1.1 Is there potential for confounding of the effect of intervention in this study? **If N/PN to 1.1:** the study can be considered to be at low risk of bias due to confounding and no further signalling questions need be considered	In rare situations, such as when studying harms that are very unlikely to be related to factors that influence treatment decisions, no confounding is expected and the study can be considered to be at low risk of bias due to confounding, equivalent to a fully randomized trial. There is no NI (No information) option for this signalling question.	Y / PY / PN / N
	If Y/PY to 1.1: determine whether there is a need to assess time-varying confounding:		
	1.2. Was the analysis based on splitting participants' follow up time according to intervention received? **If N/PN**, answer questions relating to baseline confounding (1.4 to 1.6) **If Y/PY**, proceed to question 1.3.	If participants could switch between intervention groups then associations between intervention and outcome may be biased by time-varying confounding. This occurs when prognostic factors influence switches between intended interventions.	NA / Y / PY / PN / N / NI
	1.3. Were intervention discontinuations or switches likely to be related to factors that are prognostic for the outcome? **If N/PN**, answer questions relating to baseline confounding (1.4 to 1.6) **If Y/PY**, answer questions relating to both baseline and time-varying confounding (1.7 and 1.8)	If intervention switches are unrelated to the outcome, for example when the outcome is an unexpected harm, then time-varying confounding will not be present and only control for baseline confounding is required.	NA / Y / PY / PN / N / NI
	Questions relating to baseline confounding only		
	1.4. Did the authors use an appropriate analysis method that controlled for all the important confounding domains?	Appropriate methods to control for measured confounders include stratification, regression, matching, standardization, and inverse probability weighting. They may control for individual variables or for the estimated propensity score. Inverse probability weighting is based on a function of the propensity score. Each method depends on the assumption that there is no unmeasured or residual confounding.	NA / Y / PY / PN / N / NI

	1.5. **If Y/PY to 1.4**: Were confounding domains that were controlled for measured validly and reliably by the variables available in this study?	Appropriate control of confounding requires that the variables adjusted for are valid and reliable measures of the confounding domains. For some topics, a list of valid and reliable measures of confounding domains will be specified in the review protocol but for others such a list may not be available. Study authors may cite references to support the use of a particular measure. If authors control for confounding variables with no indication of their validity or reliability pay attention to the subjectivity of the measure. Subjective measures (e.g. based on self-report) may have lower validity and reliability than objective measures such as lab findings.	NA / Y / PY / PN / N / NI
	1.6. Did the authors control for any post-intervention variables that could have been affected by the intervention?	Controlling for post-intervention variables that are affected by intervention is not appropriate. Controlling for mediating variables estimates the direct effect of intervention and may introduce bias. Controlling for common effects of intervention and outcome introduces bias.	NA / Y / PY / PN / N / NI
	Questions relating to baseline and time-varying confounding		
	1.7. Did the authors use an appropriate analysis method that adjusted for all the important confounding domains and for time-varying confounding?	Adjustment for time-varying confounding is necessary to estimate the effect of starting and adhering to intervention, in both randomized trials and NRSI. Appropriate methods include those based on inverse probability weighting. Standard regression models that include time-updated confounders may be problematic if time-varying confounding is present.	NA / Y / PY / PN / N / NI
	1.8. **If Y/PY to 1.7**: Were confounding domains that were adjusted for measured validly and reliably by the variables available in this study?	See 1.5 above.	NA / Y / PY / PN / N / NI
	Risk of bias judgement	See Table 1.	Low / Moderate / Serious / Critical / NI
	Optional: What is the predicted direction of bias due to confounding?	Can the true effect estimate be predicted to be greater or less than the estimated effect in the study because one or more of the important confounding domains was not controlled for? Answering this question will be based on expert knowledge and results in other studies and therefore can only be completed after all of the studies in the body of evidence have been reviewed. Consider the potential effect of each of the unmeasured domains and whether all important confounding domains not controlled for in the analysis would be likely to change the estimate in the same direction, or if one important confounding domain that was not controlled for in the analysis is likely to have a dominant impact.	Favours experimental / Favours comparator / Unpredictable

Bias in selection of participants into the study	2.1. Was selection of participants into the study (or into the analysis) based on participant characteristics observed after the start of intervention? **If N/PN to 2.1:** go to 2.4	This domain is concerned only with selection into the study based on participant characteristics observed *after* the start of intervention. Selection based on characteristics observed *before* the start of intervention can be addressed by controlling for imbalances between experimental intervention and comparator groups in baseline characteristics that are prognostic for the outcome (baseline confounding).	Y / PY / PN / N / NI
	2.2. **If Y/PY to 2.1**: Were the post-intervention variables that influenced selection likely to be associated with intervention?	Selection bias occurs when selection is related to an effect of either intervention or a cause of intervention **and** an effect of either the outcome or a cause of the outcome. Therefore, the result is at risk of selection bias if selection into the study is related to both the intervention and the outcome.	NA / Y / PY / PN / N / NI
	2.3 **If Y/PY to 2.2**: Were the post-intervention variables that influenced selection likely to be influenced by the outcome or a cause of the outcome?		NA / Y / PY / PN / N / NI
	2.4. Do start of follow-up and start of intervention coincide for most participants?	If participants are not followed from the start of the intervention then a period of follow up has been excluded, and individuals who experienced the outcome soon after intervention will be missing from analyses. This problem may occur when prevalent, rather than new (incident), users of the intervention are included in analyses.	Y / PY / PN / N / NI
	2.5. **If Y/PY to 2.2 and 2.3, or N/PN to 2.4**: Were adjustment techniques used that are likely to correct for the presence of selection biases?	It is in principle possible to correct for selection biases, for example by using inverse probability weights to create a pseudo-population in which the selection bias has been removed, or by modelling the distributions of the missing participants or follow up times and outcome events and including them using missing data methodology. However such methods are rarely used and the answer to this question will usually be “No”.	NA / Y / PY / PN / N / NI
	Risk of bias judgement	See Table 1.	Low / Moderate / Serious / Critical / NI
	Optional: What is the predicted direction of bias due to selection of participants into the study?	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

Bias in classification of interventions	3.1 Were intervention groups clearly defined?	A pre-requisite for an appropriate comparison of interventions is that the interventions are well defined. Ambiguity in the definition may lead to bias in the classification of participants. For individual-level interventions, criteria for considering individuals to have received each intervention should be clear and explicit, covering issues such as type, setting, dose, frequency, intensity and/or timing of intervention. For population-level interventions (e.g. measures to control air pollution), the question relates to whether the population is clearly defined, and the answer is likely to be 'Yes'.	Y / PY / PN / N / NI
	3.2 Was the information used to define intervention groups recorded at the start of the intervention?	In general, if information about interventions received is available from sources that could not have been affected by subsequent outcomes, then differential misclassification of intervention status is unlikely. Collection of the information at the time of the intervention makes it easier to avoid such misclassification. For population-level interventions (e.g. measures to control air pollution), the answer to this question is likely to be 'Yes'.	Y / PY / PN / N / NI
	3.3 Could classification of intervention status have been affected by knowledge of the outcome or risk of the outcome?	Collection of the information at the time of the intervention may not be sufficient to avoid bias. The way in which the data are collected for the purposes of the NRSI should also avoid misclassification.	Y / PY / PN / N / NI
	Risk of bias judgement	See Table 1.	Low / Moderate / Serious / Critical / NI
	Optional: What is the predicted direction of bias due to measurement of outcomes or 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

Bias due to deviations from intended interventions	**If your aim for this study is to assess the effect of assignment to intervention, answer questions 4.1 and 4.2**		
	4.1. Were there deviations from the intended intervention beyond what would be expected in usual practice?	Deviations that happen in usual practice following the intervention (for example, cessation of a drug intervention because of acute toxicity) are part of the intended intervention and therefore do not lead to bias in the effect of assignment to intervention. Deviations may 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). Such deviations are not part of usual practice, so may lead to biased effect estimates. However these are not expected in observational studies of individuals in routine care.	Y / PY / PN / N / NI
	4.2. **If Y/PY to 4.1**: 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
	If your aim for this study is to assess the effect of starting and adhering to intervention, answer questions 4.3 to 4.6		
	4.3. 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.	Y / PY / PN / N / NI
	4.4. Was the intervention implemented successfully for most participants?	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
	4.5. 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	Y / PY / PN / N / NI

		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. We distinguish between analyses where follow-up time after interventions switches (including cessation of intervention) is assigned to (1) the new intervention or (2) the original intervention. (1) is addressed under time-varying confounding, and should not be considered further here.	
	4.6. **If N/PN to 4.3, 4.4 or 4.5**: 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. Specialist advice may be needed to assess studies that used these approaches. If everyone in one group received a co-intervention, adjustments cannot be made to overcome this.	NA / Y / PY / PN / N / NI
	Risk of bias judgement	See Table 2	
	Optional: What is the predicted direction of bias due to deviations from the 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.	

Bias due to missing data	5.1 Were outcome data available for all, or nearly all, participants?	“Nearly all” should be interpreted as “enough to be confident of the findings”, and a suitable proportion depends on the context. In some situations, availability of data from 95% (or possibly 90%) of the participants may be sufficient, providing that events of interest are reasonably common in both intervention groups. One aspect of this is that review authors would ideally try and locate an analysis plan for the study.	Y / PY / PN / N / NI
	5.2 Were participants excluded due to missing data on intervention status?	Missing intervention status may be a problem. This requires that the *intended* study sample is clear, which it may not be in practice.	Y / PY / PN / N / NI
	5.3 Were participants excluded due to missing data on other variables needed for the analysis?	This question relates particularly to participants excluded from the analysis because of missing information on confounders that were controlled for in the analysis.	Y / PY / PN / N / NI
	5.4 **If PN/N to 5.1, or Y/PY to 5.2 or 5.3**: Are the proportion of participants and reasons for missing data similar across interventions?	This aims to elicit whether either (i) differential proportion of missing observations or (ii) differences in reasons for missing observations could substantially impact on our ability to answer the question being addressed. “Similar” includes some minor degree of discrepancy across intervention groups as expected by chance.	NA / Y / PY / PN / N / NI
	5.5 **If PN/N to 5.1, or Y/PY to 5.2 or 5.3**: Is there evidence that results were robust to the presence of missing data?	Evidence for robustness may come from how missing data were handled in the analysis and whether sensitivity analyses were performed by the investigators, or occasionally from additional analyses performed by the systematic reviewers. It is important to assess whether assumptions employed in analyses are clear and plausible. Both content knowledge and statistical expertise will often be required for this. For instance, use of a statistical method such as multiple imputation does not guarantee an appropriate answer. Review authors should seek naïve (complete-case) analyses for comparison, and clear differences between complete-case and multiple imputation-based findings should lead to careful assessment of the validity of the methods used.	NA / Y / PY / PN / N / NI
	Risk of bias judgement	See Table 2	Low / Moderate / Serious / Critical / NI
	Optional: What is the predicted direction of bias due to missing 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

Bias in measurement of outcomes	6.1 Could the outcome measure have been influenced by knowledge of the intervention received?	Some outcome measures involve negligible assessor judgment, e.g. all-cause mortality or non-repeatable automated laboratory assessments. Risk of bias due to measurement of these outcomes would be expected to be low.	Y / PY / PN / N / NI
	6.2 Were outcome assessors aware of the intervention received by study participants?	If outcome assessors were blinded to intervention status, the answer to this question would be 'No'. In other situations, outcome assessors may be unaware of the interventions being received by participants despite there being no active blinding by the study investigators; the answer this question would then also be 'No'. In studies where participants report their outcomes themselves, for example in a questionnaire, the outcome assessor is the study participant. In an observational study, the answer to this question will usually be 'Yes' when the participants report their outcomes themselves.	Y / PY / PN / N / NI
	6.3 Were the methods of outcome assessment comparable across intervention groups?	Comparable assessment methods (i.e. data collection) would involve the same outcome detection methods and thresholds, same time point, same definition, and same measurements.	Y / PY / PN / N / NI
	6.4 Were any systematic errors in measurement of the outcome related to intervention received?	This question refers to differential misclassification of outcomes. Systematic errors in measuring the outcome, if present, could cause bias if they are related to intervention or to a confounder of the intervention-outcome relationship. This will usually be due either to outcome assessors being aware of the intervention received or to non-comparability of outcome assessment methods, but there are examples of differential misclassification arising despite these controls being in place.	Y / PY / PN / N / NI
	Risk of bias judgement	See Table 2	Low / Moderate / Serious / Critical / NI
	Optional: What is the predicted direction of bias due to measurement of outcomes?	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

Bias in selection of the reported result	Is the reported effect estimate likely to be selected, on the basis of the results, from... 7.1. ... multiple outcome *measurements* within the outcome domain?	For a specified outcome domain, it is possible to generate multiple effect estimates for different measurements. If multiple measurements were made, but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
	7.2 ... multiple *analyses* of the intervention-outcome relationship?	Because of the limitations of using data from non-randomized studies for analyses of effectiveness (need to control confounding, substantial missing data, etc), analysts may implement different analytic methods to address these limitations. Examples include unadjusted and adjusted models; use of final value vs change from baseline vs analysis of covariance; different transformations of variables; a continuously scaled outcome converted to categorical data with different cut-points; different sets of covariates used for adjustment; and different analytic strategies for dealing with missing data. Application of such methods generates multiple estimates of the effect of the intervention versus the comparator on the outcome. If the analyst does not pre-specify the methods to be applied, and multiple estimates are generated but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
	7.3 ... different *subgroups*?	Particularly with large cohorts often available from routine data sources, it is possible to generate multiple effect estimates for different subgroups or simply to omit varying proportions of the original cohort. If multiple estimates are generated but only one or a subset is reported, there is a risk of selective reporting on the basis of results.	Y / PY / PN / N / NI
	Risk of bias judgement	See Table 2	Low / Moderate / Serious / Critical / NI
	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

Overall bias	**Risk of bias judgement**	See Table 3.	Low / Moderate / Serious / Critical / NI
	Optional: What is the overall predicted direction of bias for this outcome?		Favours experimental / Favours comparator / Towards null /Away from null / Unpredictable

Table 1. Reaching risk of bias judgements in ROBINS-I: pre-intervention and at-intervention domains

Judgement	Bias due to confounding	Bias in selection of participants into the study	Bias in classification of interventions
Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	No confounding expected.	(i) All participants who would have been eligible for the target trial were included in the study; *and* (ii) For each participant, start of follow up and start of intervention coincided.	(i) Intervention status is well defined; *and* (ii) Intervention definition is based solely on information collected at the time of intervention.
Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial):	(i) Confounding expected, all known important confounding domains appropriately measured and controlled for; *and* (ii) Reliability and validity of measurement of important domains were sufficient, such that we do not expect serious residual confounding.	(i) Selection into the study may have been related to intervention and outcome; *and* The authors used appropriate methods to adjust for the selection bias; *or* (ii) Start of follow up and start of intervention do not coincide for all participants; *and* (a) the proportion of participants for which this was the case was too low to induce important bias; *or* (b) the authors used appropriate methods to adjust for the selection bias; *or* (c) the review authors are confident that the rate (hazard) ratio for the effect of intervention remains constant over time.	(i) Intervention status is well defined; *and* (ii) Some aspects of the assignments of intervention status were determined retrospectively.

Serious risk of bias (the study has some important problems);	(i) At least one known important domain was not appropriately measured, or not controlled for; *or* (ii) Reliability or validity of measurement of an important domain was low enough that we expect serious residual confounding.	(i) Selection into the study was related (but not very strongly) to intervention and outcome; *and* This could not be adjusted for in analyses; *or* (ii) Start of follow up and start of intervention do not coincide; *and* A potentially important amount of follow-up time is missing from analyses; *and* The rate ratio is not constant over time.	(i) Intervention status is not well defined; *or* (ii) Major aspects of the assignments of intervention status were determined in a way that could have been affected by knowledge of the outcome.
Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention);	(i) Confounding inherently not controllable *or* (ii) The use of negative controls strongly suggests unmeasured confounding.	(i) Selection into the study was very strongly related to intervention and outcome; *and* This could not be adjusted for in analyses; *or* (ii) A substantial amount of follow-up time is likely to be missing from analyses; *and* The rate ratio is not constant over time.	(Unusual) An extremely high amount of misclassification of intervention status, e.g. because of unusually strong recall biases.
No information on which to base a judgement about risk of bias for this domain.	No information on whether confounding might be present.	No information is reported about selection of participants into the study or whether start of follow up and start of intervention coincide.	No definition of the intervention or no explanation of the source of information about intervention status is reported.

Table 2. Reaching risk of bias judgements in ROBINS-I: post-intervention domains

Judgement	**Bias due to deviations from intended intervention**	**Bias due to missing data**	**Bias in measurement of outcomes**	**Bias in selection of the reported result**
Low risk of bias (the study is comparable to a well-performed randomized trial with regard to this domain)	**Effect of assignment to intervention:** (i) Any deviations from intended intervention reflected usual practice; *or* (ii) Any deviations from usual practice were unlikely to impact on the outcome. **Effect of starting and adhering to intervention:** 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.	(i) Data were reasonably complete; *or* (ii) Proportions of and reasons for missing participants were similar across intervention groups; *or* (iii) The analysis addressed missing data and is likely to have removed any risk of bias.	(i) The methods of outcome assessment were comparable across intervention groups; *and* (ii) The outcome measure was unlikely to be influenced by knowledge of the intervention received by study participants (i.e. is objective) or the outcome assessors were unaware of the intervention received by study participants; *and* (iii) Any error in measuring the outcome is unrelated to intervention status.	There is clear evidence (usually through examination of a pre-registered protocol or statistical analysis plan) that all reported results correspond to all intended outcomes, analyses and sub-cohorts.

Moderate risk of bias (the study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial):	**Effect of assignment to intervention:** There were deviations from usual practice, but their impact on the outcome is expected to be slight. **Effect of starting and adhering to intervention:** (i) There were deviations from intended intervention, but their impact on the outcome is expected to be slight. *or* (ii) The important co-interventions were not balanced across intervention groups, or there were deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *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.	(i) Proportions of and reasons for missing participants differ slightly across intervention groups; *and* (ii) The analysis is unlikely to have removed the risk of bias arising from the missing data.	(i) The methods of outcome assessment were comparable across intervention groups; *and* (ii) The outcome measure is only minimally influenced by knowledge of the intervention received by study participants; *and* (iii) Any error in measuring the outcome is only minimally related to intervention status.	(i) The outcome measurements and analyses are consistent with an *a priori* plan; or are clearly defined and both internally and externally consistent; *and* (ii) There is no indication of selection of the reported analysis from among multiple analyses; *and* (iii) There is no indication of selection of the cohort or subgroups for analysis and reporting on the basis of the results.

Serious risk of bias (the study has some important problems);	**Effect of assignment to intervention:** There were deviations from usual practice that were unbalanced between the intervention groups and likely to have affected the outcome. **Effect of starting and adhering to intervention:** (i) The important co-interventions were not balanced across intervention groups, or there were deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *and* (ii) The analysis was not 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.	(i) Proportions of missing participants differ substantially across interventions; *or* Reasons for missingness differ substantially across interventions; *and* (ii) The analysis is unlikely to have removed the risk of bias arising from the missing data; *or* Missing data were addressed inappropriately in the analysis; *or* The nature of the missing data means that the risk of bias cannot be removed through appropriate analysis.	(i) The methods of outcome assessment were not comparable across intervention groups; *or* (ii) The outcome measure was subjective (i.e. vulnerable to influence by knowledge of the intervention received by study participants); *and* The outcome was assessed by assessors aware of the intervention received by study participants; *or* (iii) Error in measuring the outcome was related to intervention status.	(i) Outcomes are defined in different ways in the methods and results sections, or in different publications of the study; *or* (ii) There is a high risk of selective reporting from among multiple analyses; *or* (iii) The cohort or subgroup is selected from a larger study for analysis and appears to be reported on the basis of the results.

Critical risk of bias (the study is too problematic to provide any useful evidence on the effects of intervention);	**Effect of assignment to intervention:** There were substantial deviations from usual practice that were unbalanced between the intervention groups and likely to have affected the outcome. **Effect of starting and adhering to intervention:** (i) There were substantial imbalances in important co-interventions across intervention groups, or there were substantial deviations from the intended interventions (in terms of implementation and/or adherence) that were likely to impact on the outcome; *and* (ii) The analysis was not 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.	(i) (Unusual) There were critical differences between interventions in participants with missing data; *and* (ii) Missing data were not, or could not, be addressed through appropriate analysis.	The methods of outcome assessment were so different that they cannot reasonably be compared across intervention groups.	(i) There is evidence or strong suspicion of selective reporting of results; *and* (ii) The unreported results are likely to be substantially different from the reported results.

No information on which to base a judgement about risk of bias for this domain.	No information is reported on whether there is deviation from the intended intervention.	No information is reported about missing data or the potential for data to be missing.	No information is reported about the methods of outcome assessment.	There is too little information to make a judgement (for example, if only an abstract is available for the study).

Table 3. Interpretation of domain-level and overall risk of bias judgements in ROBINS-I

Judgement	**Within each domain**	**Across domains**	**Criterion**
Low risk of bias	The study is comparable to a well-performed randomized trial with regard to this domain	The study is comparable to a well-performed randomized trial	The study is judged to be at **low risk of bias for all domains**.
Moderate risk of bias	The study is sound for a non-randomized study with regard to this domain but cannot be considered comparable to a well-performed randomized trial	The study provides sound evidence for a non-randomized study but cannot be considered comparable to a well-performed randomized trial	The study is judged to be at **low or moderate risk of bias for all domains**.
Serious risk of bias	the study has some important problems in this domain	The study has some important problems	The study is judged to be at **serious risk of bias** in at least one domain, but not at critical risk of bias in any domain.
Critical risk of bias	the study is too problematic in this domain to provide any useful evidence on the effects of intervention	The study is too problematic to provide any useful evidence and should not be included in any synthesis	The study is judged to be at **critical risk of bias in at least one domain.**
No information	No information on which to base a judgement about risk of bias for this domain	No information on which to base a judgement about risk of bias	There is no clear indication that the study is at serious or critical risk of bias *and* there is a lack of information in one or more key domains of bias (*a judgement is required for this*).

Preliminary tool for risk of bias in exposure studies (1): At protocol stage

Specify the research question by defining a generic target experiment

Participants	
Experimental exposure	
Control exposure	

List the confounding domains relevant to all or most studies

List the possible co-exposures that could differ between exposure groups and could have an impact on study outcomes

List the criteria used to determine the accuracy of exposure measurement

Factors to consider when evaluating health outcome assessment

Preliminary tool for risk of bias in exposure studies (2): For each study

Specify a target experiment specific to the study.

□ The protocol-specified target experiment fully applies *OR*

Participant

Experimental exposure

Control exposure

Specify the outcome

Specify which outcome is being assessed for risk of bias (typically from among those earmarked for the Summary of Findings table). Specify whether this is a proposed benefit or harm of exposure.

Is your aim for this study…?

□ to assess the effect of initiating intervention (as in an intention-to-treat analysis)

□ to assess the effect of initiating and adhering to intervention (as in a per-protocol analysis)

□ other (specify)

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.

Preliminary consideration of confounders

Complete a row for each important confounding area (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as potentially important.

"Important" confounding areas are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the exposure. "Validity" refers to whether the confounding variable or variables fully measure the area, while "reliability" refers to the precision of the measurement (more measurement error means less reliability).

<table>
<tr><td colspan="5"><b>(i) Confounding areas listed in the review protocol</b></td></tr>
<tr><td>Confounding area</td><td>Measured variable(s)</td><td>Is there evidence that controlling for this variable was unnecessary?*</td><td>Is the confounding area measured validly and reliably by this variable (or these variables)?</td><td>OPTIONAL: Is adjusting for this variable (alone) expected to move the effect estimate up or down?</td></tr>
<tr><td rowspan="2"></td><td></td><td></td><td rowspan="2">Yes / No / No information</td><td>Favor intervention / Favor control / No information</td></tr>
<tr><td></td><td></td><td></td></tr>
<tr><td rowspan="2"></td><td></td><td></td><td rowspan="2"></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

<table>
<tr><td colspan="5"><b>(ii) Additional confounding areas relevant to the setting of this particular study, or which the study authors identified as important</b></td></tr>
<tr><td>Confounding area</td><td>Measured variable(s)</td><td>Is there evidence that controlling for this variable was unnecessary?*</td><td>Is the confounding area measured validly and reliably by this variable (or these variables)?</td><td>OPTIONAL: Is adjusting for this variable (alone) expected to move the effect estimate up or down?</td></tr>
<tr><td></td><td></td><td></td><td>Yes / No / No information</td><td>Favor intervention / Favor control / No information</td></tr>
</table>

* In the context of a particular study, variables can be demonstrated not to be confounders and so not included in the analysis: (a) if they are not predictive of the outcome; (b) if they are not predictive of exposure; or (c) because adjustment makes no or minimal difference to the estimated effect of the primary parameter. Note that “no statistically significant association” is not the same as “not predictive”.

Preliminary consideration of criteria used to determine the accuracy of measurement of exposure and outcome

Complete a row for each measure listed in the study for the (i) exposure and (ii) outcome. Of the measures listed in the protocol, consider the sensitivity, specificity, and confidence in the methods used in the study.

(i) Exposure measurement method listed in the study		
Method of measurement	Measured exposure	Is the exposure measured validly and reliably by this method (or these methods)?
		Yes / No / No information

(ii) Outcome measurement method listed in the study		
Method of measurement	Measured outcome	Is the outcome measured validly and reliably by this method (or these methods)?
		Yes / No / No information

Preliminary consideration of co-exposures

Complete a row for each important co-intervention (i) listed in the review protocol; and (ii) relevant to the setting of this particular study, or which the study authors identified as important.
"Important" co-interventions are those for which, in the context of this study, adjustment is expected to lead to a clinically important change in the estimated effect of the intervention.

(i) Co-exposures listed in the review protocol		
Co-exposure	Is there evidence that controlling for this co-exposure was unnecessary (e.g., because it was not administered)?	Is presence of this co-exposure likely to favor outcomes in the experimental or the control group
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information

(ii) Additional co-exposures relevant to the setting of this particular study, or which the study authors identified as important		
Co-exposure	Is there evidence that controlling for this co-exposure was unnecessary (e.g., because it was not administered)?	Is presence of this co-exposure likely to favor outcomes in the experimental or the control group
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information
		Favor experimental / Favor comparator / No information

Risk of bias assessment (cohort-type studies)

Bias due to confounding	1.1 **Is there potential for confounding of the effect of exposure in this study? If N or PN to 1.1:** the study can be considered to be at low risk of bias due to confounding and no further signaling questions need be considered	Y / PY / PN / N	[Description]
	If Y/PY to 1.1, answer 2.1 and 1.3 to determine whether there is a need to assess time-varying confounding:		
	1.2. **If Y or PY to 1.1:** Was the analysis based on splitting follow up time according to exposure received? **If N or PN to 1.2**, answer questions 1.4 to 1.6, which relate to baseline confounding	NA / Y / PY / PN / N / NI	[Description]
	1.3. **If Y or PY to 1.2**: Were exposure discontinuations or switches likely to be related to factors that are prognostic for the outcome?	NA / Y / PY / PN / N / NI	[Description]
	If N or PN to 1.3, answer questions 1.4 to 1.6, which relate to baseline confounding		
	1.4. Did the authors use an appropriate analysis method that adjusted for all the critically important confounding areas?	NA / Y / PY / PN / N / NI	[Description]
	1.5. **If Y or PY to 1.4**: Were confounding areas that were adjusted for measured validly and reliably by the variables available in this study?	NA / Y / PY / PN / N / NI	[Description]
	1.6. Did the authors avoid adjusting for post-exposure variables?	NA / Y / PY / PN / N / NI	[Description]
	If Y or PY to 1.3, answer questions 1.7 and 1.8, which relate to time-varying confounding		

	1.7. Did the authors use an appropriate analysis method that adjusted for all the critically important confounding areas and for time-varying confounding?	NA / Y / PY / PN / N / NI	[Description]
	1.8. **If Y or PY to 1.7**: Were confounding areas that were adjusted for measured validly and reliably by the variables available in this study?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to confounding?	Favors experimental / Favors comparator / Unpredictable	[Rationale]
Bias in selection of participants into the study	2.1. Was selection of participants into the study (or into the analysis) based on variables measured after the start of the exposure? **<u>If N or PN to 2.1 go to 2.4</u>**	Y / PY / PN / N / NI	[Description]
	2.2. **<u>If Y/PY to 2.1:</u>** Were the post-exposure variables that influenced selection associated with exposure?	Y / PY / PN / N / NI	[Description]
	2.3. **<u>If Y/PY to 2.2:</u>** Were the post-exposure variables that influenced eligibility selection influenced by the outcome or a cause of the outcome?	NA / Y / PY / PN / N / NI	[Description]
	2.4 Do start of follow-up and start of exposure coincide for most participants?	NA / Y / PY / PN / N / NI	[Description]
	2.5 If Y/PY to 2.2 and 2.3, or N/PN to 2.4: Were adjustment techniques used that are likely to correct for the presence of selection biases?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to selection of participants into the study?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]

Bias in classification of exposures	3.1 Is exposure status well defined?	Y / PY / PN / N / NI	[Description]
	3.2 Did entry into the study begin with start of the exposure?	Y / PY / PN / N / NI	[Description]
	3.3 Was information used to define exposure status recorded prior to outcome assessment?	Y / PY / PN / N / NI	[Description]
	3.4 Could classification of exposure status have been affected by knowledge of the outcome or risk of the outcome?	Y / PY / PN / N / NI	[Description]
	3.5 Were exposure assessment methods robust (including methods used to input data)?	Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to measurement of outcomes or exposures?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]
Bias due to departures from intended exposures	4.1. Is there concern that changes in exposure status occurred among participants? **If your aim for this study is to assess the effect of initiating and adhering to an exposure (as in a per-protocol analysis), answer questions 4.2 and 4.3, otherwise continue to 4.4 if Y or PY to 4.1.**	Y / PY / PN / N / NI	[Description]
	4.2. Did many participants switch to other exposures?	Y / PY / PN / N / NI	[Description]
	4.3. Were the critical co-exposures balanced across exposure groups?	Y / PY / PN / N / NI	[Description]
	4.4. **If NY/PN PY to 4.1, or Y/PY to 4.2, or 4.3:** Were adjustment techniques used that are likely to correct for these issues?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to departures from the intended exposures?	Favors experimental / Favors comparator / Towards null	[Rationale]

		/Away from null / Unpredictable	
Bias due to missing data	5.1 Were there missing outcome data?	Y / PY / PN / N / NI	[Description]
	5.2 Were participants excluded due to missing data on exposure status?	Y / PY / PN / N / NI	[Description]
	5.3 Were participants excluded due to missing data on other variables needed for the analysis?	Y / PY / PN / N / NI	[Description]
	5.4 <u>**If Y/PY to 5.1, 5.2 or 5.3:**</u> Are the proportion of participants and reasons for missing data similar across exposures?	NA / Y / PY / PN / N / NI	[Description]
	5.5 <u>**If Y/PY to 5.1, 5.2 or 5.3:**</u> Were appropriate statistical methods used to account for missing data?	NA / Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to missing data?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]
Bias in measurement of outcomes	6.1 Could the outcome measure have been influenced by knowledge of the exposure received?	Y / PY / PN / N / NI	[Description]
	6.2 Was the outcome measure sensitive?	Y / PY / PN / N / NI	[Description]
	6.3 Were outcome assessors unaware of the exposure received by study participants?	Y / PY / PN / N / NI	[Description]
	6.4 Were the methods of outcome assessment comparable across exposure groups?	Y / PY / PN / N / NI	[Description]
	6.5 Were any systematic errors in measurement of the outcome unrelated to exposure received?	Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]

	Optional: What is the predicted direction of bias due to measurement of outcomes?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]
Bias in selection of	Is the reported effect estimate likely to be selected, on the basis of the results, from…?		
the reported result	7.1. … multiple outcome *measurements* within the outcome domain?	Y / PY / PN / N / NI	[Description]
	7.2 … multiple *analyses* of the exposure-outcome relationship?	Y / PY / PN / N / NI	[Description]
	7.3 … different *subgroups*?	Y / PY / PN / N / NI	[Description]
	Risk of bias judgement	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the predicted direction of bias due to selection of the reported result?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]
Overall bias	**Risk of bias judgement**	Low / Moderate / Serious / Critical / NI	[Support for judgement]
	Optional: What is the overall predicted direction of bias for this outcome?	Favors experimental / Favors comparator / Towards null /Away from null / Unpredictable	[Rationale]

OHAT Risk of Bias Rating Tool
for Human and Animal Studies

INTRODUCTION

This document is written to outline a tool for evaluating individual study risk of bias or internal validity – the assessment of whether the design and conduct of a study compromised the credibility of the link between exposure and outcome (Higgins and Green 2011, IOM 2011, Viswanathan *et al.* 2012). The risk-of-bias rating tool presents a parallel approach to evaluating risk of bias in human and non-human animal studies to facilitate consideration of risk of bias across elements and across evidence streams with common terms and categories.

This tool was developed based on the most recent guidance from the Agency for Healthcare Research and Quality (Viswanathan *et al.* 2012, 2013), the Cochrane risk-of-bias tool for non-randomized studies of interventions (Sterne *et al.* 2014), Cochrane Handbook (Higgins and Green 2011), CLARITY Group at McMaster University (2013), SYRCLE's risk-of-bias tool for animal studies (Hooijmans *et al.* 2014), the Navigation Guide (Johnson *et al.* 2013, Koustas *et al.* 2013, Johnson *et al.* 2014, Koustas *et al.* 2014, Woodruff and Sutton 2014), comments from the public and technical advisors on draft methods and risk-of-bias instructions (NTP 2013d, c, b, a), staff at other federal agencies, and other sources (Downs and Black 1998, Genaidy *et al.* 2007, Dwan *et al.* 2010, Shamliyan *et al.* 2010, Shamliyan *et al.* 2011, Krauth *et al.* 2013, Wells *et al.* 2014).

For each study, risk of bias is assessed at the outcome level because certain aspects of study design and conduct may increase risk of bias for some outcomes and not others within the same study.

Organization of This Document

The majority of this document is devoted to providing detailed instructions for rating risk of bias of individual studies. Potential sources of bias are assessed with a set of 10 questions or "domains" and an additional category to consider "other potential threats to internal validity." Study design determines which questions apply [e.g., questions #1, 2, 5, 6, 7, 8, 9, 10 and 11 (or "other") apply to experimental animal studies with a different set for case-control human studies]. Detailed criteria are provided under each question that are specific for each study design. The instructions outline criteria by which individual studies are assessed and define aspects of study design, conduct, and reporting that are used to assign a risk-of-bias rating for each question.

The introduction section includes clarification of risk of bias relative to indirectness and other factors that are not considered within the OHAT risk-of-bias framework. It also provides suggestions for customizing the risk-of-bias criteria for a specific research question.

Indirectness, Timing, and Other Factors Related to Risk of Bias

Risk of bias vs indirectness:

This risk-of-bias tool evaluates internal validity – the assessment of whether the design and conduct of the study compromised the credibility of the link between exposure and outcome (Higgins and Green 2011, IOM 2011, Viswanathan *et al.* 2012). There are other aspects of a study that will impact its utility for addressing the research question such as external validity – indirectness or applicability, which are addressed elsewhere in the OHAT Approach. In other words, risk of bias addresses the question "Are the results of the study credible?" Whereas indirectness addresses the question "Did the study design address the topic of the evaluation?"

It is useful to note that some study features may be relevant to risk of bias and indirectness (Viswanathan *et al.* 2012). In particular, there are several aspects of a study relating to time, that need to be considered in both risk of bias and indirectness. For example, if there are differences in the duration of follow up across study groups, this would be a source of bias considered under detection bias "Can we be confident in the outcome assessment?" That same duration of follow up is also relevant to the indirectness or applicability of a study. If the duration of follow up was not sufficient for the development of the outcome of interest (e.g., a 6-week study of cancer endpoints), then an otherwise well-designed and well-conducted study may suffer from indirectness despite having low risk of bias.

These interrelated factors regarding timing of exposure and outcome can be considered at multiple places during an evaluation.

Time-related factors are considered at 4 points in the OHAT Approach:

- Eligibility criteria for selecting studies in Step 2 can exclude studies *a priori* where the timing of the exposure or outcome assessment are clearly inappropriate for consideration in an evaluation (e.g., chronic endpoints assessed in an acute exposure study).
- A risk-of-bias question under detection bias "Can we be confident in the exposure assessment?" considers if the exposure was assessed at a consistent time point across study groups.
- A risk-of-bias question under detection bias "Can we be confident in the outcome assessment?" considers if the outcome was assessed at a consistent time point across study groups.
- And under indirectness and applicability in Step 5 considers if the timing of exposure and outcome is acceptable for the evaluation.

The following questions are addressed in rating confidence in the body of evidence (Step 5 of the OHAT Approach), not in the risk-of-bias assessment:

- Did exposure assessment represent exposures that occurred prior to the development of the outcome? This is considered as a key feature of study design for the initial confidence rating.
- Was the exposure in the appropriate biological window to affect the outcome? This is considered under indirectness.
- Was the outcome assessed at an adequate amount of time after the exposure for the development of the outcome? This is considered under indirectness.
- Does the timing of exposure or outcome assessment impact the consistency of results? If the appropriate biological window is unclear for an outcome of interest, differences in timing of exposure or outcome assessment could be used to stratify results when considering unexplained inconsistency.

Customizing Risk-of-bias Criteria During Protocol Development

The risk-of-bias criteria and rating instructions provided in this document can be applied to many research questions, but in all cases they should be tailored to the specific research question for a given systematic review. While the criteria for most of the risk-of-bias questions will be largely similar across different reviews, the criteria for three questions should be explicitly customized for each evaluation: 1) consideration of potential confounders, 2) confidence in the exposure characterization, and 3) confidence in the outcome assessment.

Systematic review authorities recommended that subject-matter experts with knowledge of the literature participate in drafting a list of potential confounders when a review protocol is developed (Viswanathan *et al.* 2013, Sterne *et al.* 2014). Expertise and knowledge of both the exposure and outcomes of interest is required for identifying potential confounders. We recommended that experts with knowledge of the literature (including both exposure and outcome) participate in drafting the risk-of-bias criteria for potential confounders, exposure characterization, and outcome assessment when a review protocol is developed. It may be helpful to draft an analytic framework to show potential confounders that could affect the relationship between exposure and outcomes of interest. Even with early expert consultation, questions may arise when the actual studies are assessed. Additional consultation and modifications to the risk-of-bias criteria for confounders, exposure, and outcomes may be necessary. When changes are made, they should be documented along with the date on which modifications were made and the logic or justification for the changes.

Direction of Bias

Empirical evidence about the direction of bias is discussed for each of the risk-of-bias questions. Users of this document are encouraged to judge the direction of bias when possible. For some questions, the evidence will be easier to evaluate as toward or away from the null. For example, non-differential unintended co-exposure to high background phytoestrogen content in the diet will bias experimental studies of low-dose estrogenic effects toward the null. However, if there is no clear rationale for judging the likely direction of bias, review authors should simply outline the evidence and not attempt to guess the direction of evidence (Sterne *et al.* 2014).

General Instruction Format

How this tool is structured:	**Study Type Abbreviations:**
• 11 Risk-of-bias questions or domains • Each question is applicable to 1 to 6 study design types • Questions are rated by selecting among 4 possible answers (see below) • Questions are grouped under 6 types of bias (selection, confounding, performance, attrition/exclusion, detection, and selective reporting) • In practice, we will use web-based forms and reviewers will only see questions and instructions that are relevant to the study under review (i.e., text related to human studies will not appear during the evaluation of an animal study)	**EA:** Experimental Animal **HCT:** Human Controlled Trial[1] **Co:** Cohort **CaCo:** Case-Control **CrSe:** Cross-sectional **CaS:** Case Series/Case report

[1] Human controlled trial study design used here refers to studies in humans with a controlled exposure including randomized controlled trials and non-randomized experimental studies

Question Format:

- Background
  - Definition of the general category of bias
  - Clarifying text to explain what study aspects are relevant
  - Available empirical information about the direction and magnitude of the bias
  - Information about other internal validity assessment tools that consider this element
- Specific risk-of-bias rating instructions customized to each study type
  - Detailed criteria are outlined that define aspects of the study design, conduct, and reporting required to reach each risk-of-bias rating
  - The criteria are focused on distinguishing among the 4 risk-of-bias answers or ratings (e.g., outlining factors that separate "definitely low" from "probably low" risk of bias

Answer Format:

++ ***Definitely Low*** *risk of bias*:

There is direct evidence of low risk-of-bias practices
(May include specific examples of relevant low risk-of-bias practices)

+ ***Probably Low*** *risk of bias*:

There is indirect evidence of low risk-of-bias practices **OR** it is deemed that deviations from low risk-of-bias practices for these criteria during the study would not appreciably bias results, <u>including consideration of direction and magnitude of bias</u>.

– NR ***Probably High*** *risk of bias*:

There is indirect evidence of high risk-of-bias practices **OR** there is insufficient information (e.g., not reported or "NR") provided about relevant risk-of-bias practices

–– ***Definitely High*** *risk of bias*:

There is direct evidence of high risk-of-bias practices
(May include specific examples of relevant high risk-of-bias practices)

The system for answering each risk-of-bias question requires reviewers to choose between low and high risk-of-bias options. This 4-point scale is based on the approach taken by the Clarity Group at McMaster University without an answer for mixed or unclear evidence (2013). A conservative approach is taken wherein insufficient information to clearly judge the risk of bias for an individual question results in an answer rating of "Probably High" risk of bias. To clearly identify answers that were reached due to insufficient information, there are two separate symbols for "Probably High" risk of bias: 1) "-" for indirect evidence of high risk-of-bias practices, and 2) "NR" or not reported when there is insufficient information. The general answer format was adapted from (Koustas *et al.* 2013).

RISK OF BIAS RATING INSTRUCTIONS

Selection Bias

Selection bias refers to systematic differences between baseline characteristics of the groups that are compared (Higgins and Green 2011).

1. Was administered dose or exposure level adequately randomized?

Randomization of exposure or sequence generation (along with allocation concealment in question #2) helps to assure that treatment is not given selectively based on potential differences in human subjects or non-human experimental animals (e.g., randomization by animal body weight avoids potential selection bias introduced by assigning all of the smallest animals to the high-dose exposure group). Randomization requires that each human subject or animal had an equal chance of being assigned to any study group including controls (e.g., use of random number table or computer generated randomization). This applies to a concurrent negative control group (i.e., a group for which exposure is to vehicle or media alone or un-treated) which must be included in the study to address randomization as well as any positive control group that may be part of the study. For some experimental designs, the analyses are performed relative to basal levels and therefore a human subject or animal may serve as its own control.

A lack of randomization can bias results away from the null towards larger effect sizes. This effect has been empirically assessed in both controlled human trials (reviewed in Higgins and Green 2011) and experimental animals (reviewed in Krauth *et al.* 2013).

This element is widely recommended to assess risk of bias for controlled human trials (Guyatt *et al.* 2011, Higgins and Green 2011, IOM 2011, Viswanathan *et al.* 2012) and is included in most risk-of-bias instruments for animal studies (reviewed in Krauth *et al.* 2013, Hooijmans *et al.* 2014).

We recognize that given reporting practices for experimental animal studies it is unlikely that the allocation method will be explicitly reported in most studies. Thus, in cases where randomization is reported but the method is unknown (i.e., not reported and cannot be obtained through author query), we will classify studies as “probably low risk of bias”. In cases where randomization is not reported, we will assume that randomization was not undertaken and classify such studies as “probably high risk of bias”.

Note: normalization is discussed in a separate risk-of-bias question under confounding bias: Did the study design or analysis account for important confounding or modifying variables?

Applies to: HCT, EA

Definitely Low risk of bias:

HCT: There is direct evidence that subjects were allocated to any study group including controls using a method with a random component. Acceptable methods of randomization include: referring to a random number table, using a computer random number generator, coin tossing, shuffling cards or envelopes, throwing dice, or drawing of lots (Higgins and Green 2011). Restricted randomization (e.g., blocked randomization) to ensure particular allocation ratios will be considered low risk of bias. Similarly, stratified randomization and minimization approaches that attempt to minimize imbalance between groups on important prognostic factors (e.g., body weight) will be considered acceptable.

EA: There is direct evidence that animals were allocated to any study group including controls using a method with a random component,

AND there is direct evidence that the study used a concurrent control group as an indication that randomization covered all study groups.

Note: Acceptable methods of randomization include: referring to a random number table, using a computer random number generator, coin tossing, shuffling cards or envelopes, throwing dice, or drawing of lots (Higgins and Green 2011). Restricted randomization (e.g., blocked randomization) to ensure particular allocation ratios will be considered low risk of bias. Similarly, stratified randomization and minimization approaches that attempt to minimize imbalance between groups on important prognostic factors (e.g., body weight) will be considered acceptable. This type of approach is used by NTP, i.e., random number generator with body weight as a covariate.

Note: Investigator-selection of animals from a cage is not considered random allocation because animals may not have an equal chance of being selected, e.g., investigator selecting animals with this method may inadvertently choose healthier, easier to catch, or less aggressive animals.

Probably Low risk of bias:

HCT: There is indirect evidence that subjects were allocated to study groups using a method with a random component (i.e., authors state that allocation was random, without description of the method used),

OR it is deemed that allocation without a clearly random component during the study would not appreciably bias results. For example, approaches such as biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization may require consultation with a statistician to determine risk-of-bias rating (Higgins and Green 2011).

EA: There is indirect evidence that animals were allocated to any study group including controls using a method with a random component (i.e., authors state that allocation was random, without description of the method used),

AND there is direct or indirect evidence that the study used a concurrent control group as an indication that randomization covered all study groups,

OR it is deemed that allocation without a clearly random component during the study would not appreciably bias results. For example, approaches such as biased coin or urn randomization, replacement randomization, mixed randomization, and maximal randomization may require consultation with a statistician to determine risk-of-bias rating (Higgins and Green 2011).

Probably High risk of bias:

HCT: There is indirect evidence that subjects were allocated to study groups using a method with a non-random component,

OR there is insufficient information provided about how subjects were allocated to study groups (record “NR” as basis for answer).

Note: Non-random allocation methods may be systematic, but have the potential to allow participants or researchers to anticipate the allocation to study groups. Such “quasi-random” methods include alternation, assignment based on date of birth, case record number, or date of presentation to study (Higgins and Green 2011).

EA: There is indirect evidence that animals were allocated to study groups using a method with a non-random component,

OR there is indirect evidence that there was a lack of a concurrent control group,

OR there is insufficient information provided about how subjects were allocated to study groups (record "NR" as basis for answer).

Note: Non-random allocation methods may be systematic, but have the potential to allow researchers to anticipate the allocation of animals to study groups (Higgins and Green 2011). Such "quasi-random" methods include investigator-selection of animals from a cage, alternation, assignment based on shipment receipt date, date of birth, or animal number.

Definitely High risk of bias:

HCT: There is direct evidence that subjects were allocated to study groups using a non-random method including judgment of the clinician, preference of the participant, the results of a laboratory test or a series of tests, or availability of the intervention (Higgins and Green 2011).

EA: There is direct evidence that animals were allocated to study groups using a non-random method including judgment of the investigator, the results of a laboratory test or a series of tests (Higgins and Green 2011),

OR there is direct evidence that there was a lack of a concurrent control group, indicating that randomization did not cover all study groups.

2. Was allocation to study groups adequately concealed?

Allocation concealment prior to assigning the exposure level or treatment group (along with randomization in question #1) helps to assure that treatment is not given selectively based on potential differences in human subjects or non-human experimental animals.

Allocation concealment requires that research personnel allocating subjects or animals to treatment groups (including the control group) could not foresee which administered dose or exposure level is going to be assigned at the start of a study. Human studies also require that allocation be concealed from human subjects prior to entering the study.

A lack of allocation concealment can bias results away from the null towards larger effect sizes. This effect has been empirically assessed in both controlled human trials [(Schulz *et al.* 1995, Schulz *et al.* 2002, Pildal *et al.* 2007); see also studies reviewed in (Higgins and Green 2011)] and in animal studies [(Macleod *et al.* 2008) ; see also studies reviewed in (Krauth *et al.* 2013)].

This element is widely recommended to assess risk of bias for controlled human trials (Guyatt *et al.* 2011, Higgins and Green 2011, IOM 2011, Viswanathan *et al.* 2012) and included in some risk-of-bias instruments for animal studies (reviewed in Krauth *et al.* 2013).

Note: there are separate risk-of-bias issues affected by concealment or blinding that are important for selection, performance and detection bias: 1) a question under performance bias addresses blinding of research personnel and human subjects to study groups during the study; and 2) a question under detection bias addresses blinding during outcome assessment.

Applies to: HCT, EA

Definitely Low risk of bias:

HCT: There is direct evidence that at the time of recruitment the research personnel and subjects did not know what study group subjects were allocated to, and it is unlikely that they could have broken the blinding of allocation until after recruitment was complete and irrevocable. Acceptable methods used to ensure allocation concealment include central allocation (including telephone, web-based and pharmacy-controlled randomization); sequentially numbered drug containers of identical appearance; sequentially numbered, opaque, sealed envelopes; or equivalent methods.

EA: There is direct evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to, and it is unlikely that they could have broken the blinding of allocation until after assignment was complete and irrevocable. Acceptable methods used to ensure allocation concealment include sequentially numbered treatment containers of identical appearance or equivalent methods.

Probably Low risk of bias:

HCT: There is indirect evidence that the research personnel and subjects did not know what study group subjects were allocated to and it is unlikely that they could have broken the blinding of allocation until after recruitment was complete and irrevocable,

OR it is deemed that lack of adequate allocation concealment would not appreciably bias results.

EA: There is indirect evidence that at the time of assigning study groups the research personnel did not know what group animals were allocated to and it is unlikely that they could have broken the blinding of allocation until after assignment was complete and irrevocable,

OR it is deemed that lack of adequate allocation concealment would not appreciably bias results.

Probably High risk of bias:

HCT: There is indirect evidence that at the time of recruitment it was possible for the research personnel and subjects to know what study group subjects were allocated to, or it is likely that they could have broken the blinding of allocation before recruitment was complete and irrevocable,

OR there is insufficient information provided about allocation to study groups (record "NR" as basis for answer).

Note: Inadequate methods include using an open random allocation schedule (e.g., a list of random numbers); assignment envelopes used without appropriate safeguards (e.g., if envelopes were unsealed or non-opaque or not sequentially numbered); alternation or rotation; date of birth; case record number; or any other explicitly unconcealed procedure. For example, if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.

EA: There is indirect evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable,

OR there is insufficient information provided about allocation to study groups (record "NR" as basis for answer).

Definitely High risk of bias:

HCT: There is direct evidence that at the time of recruitment it was possible for the research personnel and subjects to know what study group subjects were allocated to, or it is likely that they could have broken the blinding of allocation before recruitment was complete and irrevocable.

EA: There is direct evidence that at the time of assigning study groups it was possible for the research personnel to know what group animals were allocated to, or it is likely that they could have broken the blinding of allocation before assignment was complete and irrevocable.

3. Did selection of study participants result in appropriate comparison groups?

Comparison group appropriateness refers to having similar baseline characteristics of factors related to the outcome measures of interest between groups aside from the exposures (and outcomes for case-control studies).

Assessment of appropriate selection of comparison groups is a widely used element of tools to assess study quality for observational human studies (Downs and Black 1998, Shamliyan *et al.* 2010, Viswanathan *et al.* 2012, CLARITY Group at McMaster University 2013, Sterne *et al.* 2014, Wells *et al.* 2014). This question addresses whether exposed and unexposed subjects were recruited from the same populations in cohort or cross-sectional studies and consideration of appropriate selection of cases and controls in case-control studies.

The direction of the bias (towards or away from the null) will differ based on the nature of differences between comparison groups and may be difficult to predict.

For example, in occupational cohorts, it is common for workers to have lower rates of disease and mortality than the general population – the healthy worker effect – because the severely ill and chronically disabled are commonly excluded from employment (Gerstman 2013). Therefore, comparing workers to an inherently less healthy group (general population or workers with less physically demanding work) can bias the estimate of disease risk towards the null (Rothman *et al.* 2012). Conversely, if cases of disease identified from a screening program were compared to controls from the general population, the effect estimate could be overestimated as those being screened may inherently have a higher risk (e.g., family history) so the better comparison group would be subjects screened as not having disease (Szklo and Nieto 2007).

For controlled exposure studies (i.e., experimental human or animal studies), the potential for imbalance of baseline characteristics is controlled for through randomization and allocation concealment. Imbalance can arise from chance alone, but baseline characteristics should be similar for truly randomized human controlled trials (Higgins and Green 2011) or other experimental studies. The majority of study quality tools for experimental animals do not have a separate question on baseline characteristics (Krauth *et al.* 2013, Koustas *et al.* 2014); although the SYRCLE tool asks whether groups were "similar at baseline or were they adjusted for confounders in the analysis" (Hooijmans *et al.* 2014). The Cochrane risk-of-bias tool for randomized controlled trials does not include a routine question on baseline characteristics, and instead suggests that reviewers consider "inexplicable baseline imbalance" under other potential threats to internal validity (Higgins *et al.* 2011). This tool takes the same approach for all experimental studies and addresses baseline imbalance for these studies only where it is strongly suspected with a question at the end of the risk-of-bias-tool under other potential threats to internal validity.

Applies to: Co, CaCo, CrSe [HCT and EA see other potential threats to internal validity]

Definitely Low risk of bias:

Co, CrSe: There is direct evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), recruited within the same time frame, and had the similar participation/response rates.

CaCo: There is direct evidence that cases and controls were similar (e.g., recruited from the same eligible population including being of similar age, gender, ethnicity, and eligibility criteria other than outcome of interest as appropriate), recruited within the same time frame, and controls are described as having no history of the outcome.

Note: A study will be considered low risk of bias if baseline characteristics of groups differed but these differences were considered as potential confounding or stratification variables (see question #4).

Probably Low risk of bias:

Co, CrSe: There is indirect evidence that subjects (both exposed and non-exposed) were similar (e.g., recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age and health status), recruited within the same time frame, and had the similar participation/response rates,
OR differences between groups would not appreciably bias results.

CaCo: There is indirect evidence that cases and controls were similar (e.g., recruited from the same eligible population, recruited with the same method of ascertainment using the same inclusion and exclusion criteria, and were of similar age), recruited within the same time frame, and controls are described as having no history of the outcome,
OR differences between cases and controls would not appreciably bias results.

Probably High risk of bias:

Co, CrSe: There is indirect evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had the very different participation/response rates,
OR there is insufficient information provided about the comparison group including a different rate of non-response without an explanation (record “NR” as basis for answer).

CaCo: There is direct evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames,
OR there is insufficient information provided about the appropriateness of controls including rate of response reported for cases only (record “NR” as basis for answer).

Definitely High risk of bias:

Co, CrSe: There is direct evidence that subjects (both exposed and non-exposed) were not similar, recruited within very different time frames, or had the very different participation/response rates.

CaCo: There is direct evidence that controls were drawn from a very dissimilar population than cases or recruited within very different time frames.

Confounding Bias

Bias relating to confounding and co-exposures is addressed under selection bias and performance in study quality tools such as Cochrane, AHRQ, and SYRCLE (Higgins and Green 2011, Higgins *et al.* 2011, Viswanathan *et al.* 2012, Hooijmans *et al.* 2014). The grouping of these related factors under "confounding bias" does not change the questions or the evaluation of bias, but rather is done for clarity in communicating bias related to confounding, modifying variables, and other exposures that are anticipated to bias results.

4. Did the study design or analysis account for important confounding and modifying variables?

Interpretation of study findings may be distorted by failure to consider the extent to which systematic differences in baseline characteristics risk factors, prognostic variables[2], or co-occurring exposures among comparison groups may reduce or increase the observed effect (IOM 2011). Confounding variables or confounders include any factor that is: 1) associated with the exposure, 2) an independent risk factor for a given outcome, and 3) unequally distributed between study groups (Gerstman 2013). The potential confounder cannot be an intermediate effect on the causal pathway between exposure and the outcome (Gerstman 2013, Sterne *et al.* 2014). Appropriate methods to account for these differences would include multivariable analysis, stratification, matching of cases and controls, or other approaches.

Adjusting or controlling for confounding is dependent on valid, reliable, and sensitive methods for assessing the confounding or modifying variables applied consistently across study groups. The requirement for assessing the confounding variables with valid and reliable measures is directly linked to the relative importance of the confounding variable considered under selection bias (i.e., if a confounder needed to be accounted for in design or analyses, then measurement of that variable had to be reliable).

This element is included in this current risk-of-bias tool because it is widely recommended in tools used to assess the quality of observational human studies (Downs and Black 1998, Shamliyan *et al.* 2010, Viswanathan *et al.* 2012, CLARITY Group at McMaster University 2013, Viswanathan *et al.* 2013, Sterne *et al.* 2014). The direction of the bias (towards or away from the null) will differ based on the nature of differences between comparison groups. Generally, confounding results in effect sizes that are

[2] "Risk" factors are those which as associated with causing a condition (like smoking for lung cancer or being born premature for chronic lung disease). 'Prognostic' factors are those which, in people who have the condition, influence the outcome (like resectability of tumor for lung cancer or duration of intubation for CLD). Risk factors are determined by looking at things that influence new cases ('incident' ones), while prognostic factors can only be determined by following up people who already have the disease (http://blogs.bmj.com/adc-archimedes/2009/03/09/risk-vs-prognostic-factors/) .

overestimated. However, confounding factors can lead to an underestimation of the effect of a treatment or exposure, particularly in observational studies. In other words, if the confounding variables were not present, the measured effect would have been even larger (IOM 2011).

Unintended co-exposures may represent a confounding factor if associated with exposure and the outcome of interest, or a modifying factor if they are independent of exposure, but associated with outcome. When an unintended exposure is an effect modifier, its level will alter the magnitude of the effect of the primary outcome. The direction of the bias (towards or away from the null) will differ based on the nature of unintended exposure and whether or not it is associated with the primary exposure. For example, an exposed group in a human study living at a Superfund site may also be exposed to high levels of other environmental contaminants; if these co-exposures are not accounted for in the analyses, they may bias results away from the null (towards larger effects sizes). Alternately, a co-exposure that is non-differentially distributed among both the exposed and control groups will usually bias the results toward the null by lowering precision and therefore reducing the ability to distinguish potential effects between groups based on the primary exposure.

It is understood in environmental health that people are exposed to complex mixtures of environmental contaminants and other types of exposures that make it difficult to establish chemical-specific associations. Thus, in most cases we will not penalize studies if other exposures or potential exposures are not adjusted or controlled for in the analyses of a target exposure. For some projects, exceptions may include studies where levels of other chemicals aside from the chemical of interest are likely to be high, such as in occupational cohorts or contaminated regions (e.g., Superfund sites). For some health outcomes, consideration of additional therapies, including medications, may also be appropriate.

By definition, confounders are specific for the outcome and the exposure. Therefore, the list of potential confounders has to be developed specifically for each evaluation and will require subject-matter expertise on both the outcome and exposure of interest. Systematic review authorities recommended that subject-matter experts with some knowledge of the literature participate in drafting a list of potential confounders when a review protocol is developed (Viswanathan *et al.* 2013, Sterne *et al.* 2014). It may be helpful to draft an analytic framework that shows potential confounders that could affect the relationship between exposure and outcomes of interest. Even when a list of potential confounders is developed when drafting the protocol, it is likely that new confounders will be identified when actually assessing the risk of bias of studies.

Although confounding is a much greater concern for observational studies, experimental studies are not entirely free of these issues. Controlled exposure studies (i.e., experimental human or animal studies) can address confounding and selection bias through study design features such as randomization and allocation concealment. Confounding by chance (i.e., confounding that is unknown, unmeasured, or poorly measured) is expected to be equally distributed between groups under true randomization; however, experimental studies may not always successfully randomize potential confounders (Viswanathan *et al.* 2013). Recognizing this, the SYRCLE risk-of-bias tool for experimental animal studies asks whether groups were "similar at baseline or were they adjusted for confounders in the analysis" (Hooijmans *et al.* 2014). The 2012 risk-of-bias guidance from AHRQ recommends consideration of confounding for randomized clinical trials largely because studies may fail to randomize confounders. However, the Cochrane risk-of-bias tool for randomized controlled trials does not include a question for confounding, nor do the majority of study quality tools for experimental animals (Krauth *et al.* 2013, Koustas *et al.* 2014).

For this tool, we have not included a separate question for confounding in experimental human or experimental animal studies because randomization and allocation concealment should address the issue of confounding. Therefore, the issue of confounding overlaps with randomization and allocation concealment, and multiple questions would address the same issue. We recognize that in some cases confounding or effect modification may be a potential risk of bias despite procedures to address randomization. For example, confounding would be a concern if there were differential distribution of baseline characteristics such as body weight or BMI in a study of obesity, despite adequate procedures for randomization and allocation concealment. In another example, effect modification and bias toward the null would be of concern in an experimental study designed to test reproductive effects of estrogenic chemicals with non-differential co-exposures to high levels of phytoestrogens through the diet. For experimental studies where confounding is strongly suspected, randomization and allocation concealment should be addressed first. If these questions are rated "probably low" or "definitely low risk of bias," then confounding may be addressed under "other potential threats to internal validity."

Note: in the current OHAT tool, assessment of confounding requires consideration of whether or not 1) the design or analysis accounted for confounding and modifying variables, 2) the confounding variables were measured reliably and consistently, and 3) there were other exposures anticipated to bias results in reaching a single risk-of-bias rating on confounding. Previous versions of the OHAT tool used three separate questions for these factors (Did the study design or analysis account for important confounding and modifying variables?" "Were confounding variables assessed consistently across groups using valid and reliable measures" and "Did researchers adjust or control for other exposures that are anticipated to bias results?" The current tool considers these factors together because they are interrelated and recent guidance has taken a similar approach (e.g., Sterne et al. 2014).

Previous versions of the OHAT risk-of-bias tool applied the question on confounding to experimental study designs. As described above, this tool does not routinely apply this question to experimental studies because the issue of confounding overlaps with randomization and allocation concealment, and multiple questions would address the same issue. However, for review questions or individual experimental studies where confounding is strongly suspected despite adequate control for randomization and allocation concealment, confounding may be addressed under "other potential threats to internal validity."

Applies to: Co, CaCo, CrSe, CaS [HCT and EA see other potential threats to internal validity]

Definitely Low risk of bias:

Co, CrSe, CaS: There is direct evidence that appropriate adjustments or explicit considerations were made for primary covariates and confounders in the final analyses through the use of statistical models to reduce research-specific bias including standardization, matching, adjustment in multivariate model, stratification, propensity scoring, or other methods that were appropriately justified. Acceptable consideration of appropriate adjustment factors includes cases when the factor is not included in the final adjustment model because the author conducted analyses that indicated it did not need to be included,
AND there is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements,
AND there is direct evidence that other exposures anticipated to bias results were not present or were appropriately measured and adjusted for. In occupational studies or studies of contaminated sites, other chemical exposures known to be associated with those settings were appropriately considered.

CaCo: There is direct evidence that appropriate adjustments were made for primary covariates and confounders in the final analyses through the use of statistical models to reduce research-specific bias including standardization, matching of cases and controls, adjustment in

multivariate model, stratification, propensity scoring, or other methods were appropriately justified,
AND there is direct evidence that primary covariates and confounders were assessed using valid and reliable measurements,
AND there is direct evidence that other exposures anticipated to bias results were not present or were appropriately measured and adjusted for.

Probably Low risk of bias:

Co, CaCo, CrSe, CaS: There is indirect evidence that appropriate adjustments were made,
OR it is deemed that not considering or only considering a partial list of covariates or confounders in the final analyses would not appreciably bias results.
AND there is evidence (direct or indirect) that primary covariates and confounders were assessed using valid and reliable measurements,
OR it is deemed that the measures used would not appreciably bias results (i.e., the authors justified the validity of the measures from previously published research),
AND there is evidence (direct or indirect) that other co-exposures anticipated to bias results were not present or were appropriately adjusted for,
OR it is deemed that co-exposures present would not appreciably bias results.
Note: As discussed above, this includes insufficient information provided on co-exposures in general population studies.

Probably High risk of bias:

Co, CrSe, CaS: There is indirect evidence that the distribution of primary covariates and known confounders differed between the groups and was not appropriately adjusted for in the final analyses,
OR there is insufficient information provided about the distribution of known confounders (record “NR” as basis for answer),
OR there is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity,
OR there is insufficient information provided about the measurement techniques used to assess primary covariates and confounders (record “NR” as basis for answer),
OR there is indirect evidence that there was an unbalanced provision of additional co-exposures across the primary study groups, which were not appropriately adjusted for,
OR there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated (record “NR” as basis for answer).

CaCo: There is indirect evidence that the distribution of primary covariates and known confounders differed between cases and controls and was not investigated further,
OR there is insufficient information provided about the distribution of known confounders in cases and controls (record “NR” as basis for answer),
OR there is indirect evidence that primary covariates and confounders were assessed using measurements of unknown validity,
OR there is insufficient information provided about the measurement techniques used (record “NR” as basis for answer),
OR there is indirect evidence that there was an unbalanced provision of additional co-exposures across cases and controls, which were not appropriately adjusted for,

OR there is insufficient information provided about co-exposures in occupational studies or studies of contaminated sites where high exposures to other chemical exposures would have been reasonably anticipated (record "NR" as basis for answer).

Definitely High risk of bias:

Co, CrSe, CaS: There is direct evidence that the distribution of primary covariates and known confounders differed between the groups, confounding was demonstrated, and was not appropriately adjusted for in the final analyses,
OR there is direct evidence that primary covariates and confounders were assessed using non valid measurements,
OR there is direct evidence that there was an unbalanced provision of additional co-exposures across the primary study groups, which were not appropriately adjusted for.

CaCo: There is direct evidence that the distribution of primary covariates and known confounders differed between cases and controls, confounding was demonstrated, but was not appropriately adjusted for in the final analyses,
OR there is direct evidence that primary covariates and confounders were assessed using non valid measurements,
OR there is direct evidence that there was an unbalanced provision of additional co-exposures across cases and controls, which were not appropriately adjusted for.

Performance Bias

Performance bias refers to systematic differences in the care provided to human participants or experimental animals by study groups. Examples include contamination of the control group with the exposure or intervention, unbalanced provision of additional interventions or co-interventions, difference in co-interventions, inadequate blinding of providers and participants in human studies (Viswanathan *et al.* 2012), and inadequate blinding of research personnel to the animal's study group (Sena *et al.* 2007).

5. Were experimental conditions identical across study groups?

Housing conditions and husbandry practices should be identical across control and experimental groups because these variables may impact the outcome of interest (Duke *et al.* 2001, Gerdin *et al.* 2012). Identical conditions include use of the same vehicle in control and experimental animals. This risk-of-bias element is included in some tools used to assess animal studies (Krauth *et al.* 2013).

We recognize that given reporting practices it is unlikely that similarity of conditions will be explicitly reported in most animal studies. Thus, we will assume unless stated otherwise that experimental conditions (other than use of appropriate vehicle for control animals) were identical across groups which will result in most studies considered "probably low risk of bias". Thus in this tool, the rating for this risk-of-bias element will depend largely on the consistent use vehicle across treatment groups. This risk-of-bias element is unlikely to be informative for the purposes of discriminating between studies based on housing conditions or husbandry practices. However, in the long-term, especially if reporting standards improve, collecting this information may generate data that will allow us to empirically assess evidence of bias based on housing conditions or husbandry practices or to remove those features from consideration.

Applies to: EA

Definitely Low risk of bias:

EA: There is direct evidence that same vehicle was used in control and experimental animals,
AND there is direct evidence that non-treatment-related experimental conditions were identical across study groups (i.e., the study report explicitly provides this level of detail).

Probably Low risk of bias:

EA: There is indirect evidence that the same vehicle was used in control and experimental animals,
OR it is deemed that the vehicle used would not appreciably bias results.
AND as described above, identical non-treatment-related experimental conditions are assumed if authors did not report differences in housing or husbandry.

Probably High risk of bias:

EA: There is indirect evidence that the vehicle differed between control and experimental animals,
OR authors did not report the vehicle used (record "NR" as basis for answer),
OR there is indirect evidence that non-treatment-related experimental conditions were not comparable between study groups.

Definitely High risk of bias:

EA: There is direct evidence from the study report that control animals were untreated, or treated with a different vehicle than experimental animals,
OR there is direct evidence that non-treatment-related experimental conditions were not comparable between study groups.

6. Were the research personnel and human subjects blinded to the study group during the study?

Blinding requires that research personnel do not know which administered dose or exposure level the human subject or animal is being given (i.e., study group). Human studies also require blinding of the human subjects when possible.

Human introductory text: If research personnel or human subjects are not blinded to the study groups it could affect the actual outcomes of the participants due to differential behaviors across intervention groups. During the course of a study blinding of participants and research personnel is a recommended risk-of-bias element in the most recent Cochrane guidance for assessing randomized clinical trials (Higgins and Green 2011).

No empirical evidence of bias due to failure to blind during the course of a study is currently available. However, 'blind' or 'double-blind' study descriptions usually include blinding of research personnel, human subjects, or both. Without distinguishing between the different stages of blinding during the conduct of a study, lack of blinding in randomized trials has been empirically shown to be associated with larger estimations of intervention effects (on average a 9% increase in an odds ratio) (Pildal *et al.*

2007). Schulz *et al*. (1995) analyzed 250 controlled trials and found that studies that were not double-blinded had a 17% larger estimation of treatment effect, on average. If additional investigations or co-interventions occur differentially across intervention groups, bias can also be introduced by not blinding research personnel or human subjects.

For some exposures, it is not possible to entirely blind research personnel and subjects during the course of the study (an exercise intervention or patients receiving surgery). However, adherence to a strict study protocol to minimize differential behaviors by research personnel and human subjects can reduce the risk of bias. In practice, successful blinding cannot be ensured, as it can be compromised for most interventions. In some case the treatment may have side effects possibly allowing the participant to detect which intervention they received, unless the study compares interventions with similar side effects or uses an active placebo (Boutron *et al.* 2006).

Animal introductory text: Lack of blinding of research personnel could bias the results by affecting the actual outcomes of the animals in the study. This may be due to differences in handling of animals (e.g., stress-related effects) or monitoring for health outcomes. For example, an investigator may be more likely to take measures to ensure that animals in experimental groups receive the appropriate dose volume compared to animals in the control group. Lack of blinding might also lead to bias caused by additional investigations or co-interventions regardless of the type of outcomes, if these occur differentially across intervention groups (Higgins and Green 2011).

This element is recommended to assess performance bias for controlled human trials (Higgins and Green 2011) and animal studies (reviewed in Krauth *et al.* 2013), although empirical evidence of bias due to lack of blinding of research personnel during the course of the study is not currently available. Rosenthal and Lawson (1964) reported that rats that experimenters had been told were “bright” performed better than rats labeled “dull” in Skinner box learning tests, despite the fact that they were the same rats. The study design did not allow clear separation between experimenter bias introduced during handling or training from bias at outcome assessment. As discussed under detection bias, lack of blinding at outcome assessment in animal studies is associated with larger measures of the effect (Bebarta *et al.* 2003, Sena *et al.* 2007, Vesterinen *et al.* 2010).

In animal studies, blinding of study group during the course of the study is often not possible for animal welfare considerations and the need to determine if treated animals are affected relative to controls in a treatment or dose-dependent manner (examples include clinical observations and histopathologic assessment of non-neoplastic lesions). Knowledge and tracking of higher exposed animals may also be part of animal welfare practices designed to avoid suffering associated with overtly toxic treatment doses. Under some conditions it is unlikely that blinding of research personnel during the course of a study can be fully achieved. However, animal studies are in general more tightly controlled than human studies and additional measures may be taken to reduce the risk of bias, such as the generation and use of standard operating procedures, training, and randomized husbandry or handling practices (e.g., placement in the animal room, necropsy order, etc.).

Note: there are separate risk-of-bias issues affected by concealment or blinding that are important for selection, performance and detection bias: 1) a question under selection bias addresses allocation concealment of research personnel and human subjects; and 2) a question under detection bias addresses blinding during outcome assessment.

Applies to: HCT, EA

Definitely Low risk of bias:

HCT: There is direct evidence that the subjects and research personnel were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study. Methods used to ensure blinding include central allocation; sequentially numbered drug containers of identical appearance; sequentially numbered, opaque, sealed envelopes; or equivalent methods.

EA: There is direct evidence that the research personnel were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study. Methods used to ensure blinding include central allocation; sequentially numbered treatment containers of identical appearance; sequentially numbered animal cages; or equivalent methods.

Probably Low risk of bias:

HCT: There is indirect evidence that the research personnel and subjects were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study,

OR it is deemed that lack of adequate blinding during the study would not appreciably bias results.

EA: There is indirect evidence that the research personnel were adequately blinded to study group, and it is unlikely that they could have broken the blinding during the study,

OR it is deemed that lack of adequate blinding during the study would not appreciably bias results. This would include cases where blinding was not possible but research personnel took steps to minimize potential bias, such as restricting the knowledge of study group to veterinary or supervisory personnel monitoring for overt toxicity, or randomized husbandry or handling practices (e.g., placement in the animal room, necropsy order, etc.).

Probably High risk of bias:

HCT: There is indirect evidence that it was possible for research personnel or subjects to infer the study group,

OR there is insufficient information provided about blinding to study group during the study (record "NR" as basis for answer).

Note: Inadequate methods include using an open random allocation schedule (e.g., a list of random numbers), assignment envelopes used without appropriate safeguards (e.g., if envelopes were unsealed or non-opaque or not sequentially numbered), alternation or rotation; date of birth; case record number; or any other explicitly unconcealed procedure. For example, if the use of assignment envelopes is described, but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.

EA: There is indirect evidence that the research personnel were not adequately blinded to study group,

OR there is insufficient information provided about blinding to study group during the study (record "NR" as basis for answer).

Definitely High risk of bias:

HCT: There is direct evidence for lack of adequate blinding of the study group including no blinding or incomplete blinding of research personnel and subjects. For some treatments, such as behavioral interventions, allocation to study groups cannot be concealed.

EA: There is direct evidence that the research personnel were not adequately blinded to study group.

Attrition/Exclusion Bias

Attrition or exclusion bias refers to systematic differences in the loss or exclusion from analyses of participants or animals from the study and how they were accounted for in the results (Viswanathan *et al.* 2012).

7. Were outcome data complete without attrition or exclusion from analysis?

Incomplete outcome data includes loss due to attrition (nonresponse, dropout, or loss to follow-up) or exclusion from analyses. The degree of bias resulting from incomplete outcome data depends on the reasons that outcomes are missing, the amount and distribution of missing data across groups, and the potential association between outcome values and likelihood of missing data (Higgins and Green 2011). The risk of bias from incomplete outcome data can be reduced if study authors address the problem in their analyses (e.g., intention to treat analysis and imputation). Exclusion of individuals or animals from analyses should be clearly reported and outliers identified with appropriate statistical procedures.

Human introductory text: Differential or overall attrition because of nonresponse, dropping out, loss to follow-up, and exclusion of participants can introduce bias when missing outcome data are related to both exposure/treatment and outcome. Those who drop out of the study or who are lost to follow-up may be systematically different from those who remain in the study. Attrition or exclusion bias can potentially change the collective (group) characteristics of the relevant groups and their observed outcomes in ways that affect study results by confounding and spurious associations (Viswanathan *et al.* 2012). This risk-of-bias element is recommended to assess controlled human trials (Higgins and Green 2011), observational human studies (Viswanathan *et al.* 2012, Sterne *et al.* 2014) and animal studies (Krauth *et al.* 2013). However, concern over bias from incomplete outcome data is mainly theoretical and most studies that have looked at whether aspects of missing data are associated with magnitude of effect estimates have not found clear evidence of bias (reviewed in Higgins and Green 2011).

Animal introductory text: Attrition or exclusion because of illness, death, or other reasons can introduce bias when missing outcome data are related to both exposure and outcome. Attrition bias can potentially change the collective (group) characteristics of the relevant groups and their observed outcomes in ways that affect study results by confounding and spurious associations (Viswanathan *et al.* 2012). This risk-of-bias element is recommended to assess controlled human trials (Higgins and Green 2011), observational human studies (Viswanathan *et al.* 2012, Sterne *et al.* 2014) and animal studies (Krauth *et al.* 2013). However, concern over bias from incomplete outcome data is mainly theoretical and most studies that have looked at whether aspects of missing data are associated with magnitude of effect estimates have not found clear evidence of bias (reviewed in Higgins and Green 2011).

Applies to: HCT, EA, Co, CaCo, CrSe

Definitely Low risk of bias:

HCT: There is direct evidence that there was no loss of subjects during the study and outcome data were complete,

OR loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study or analyses. Review authors should be confident that the participants included in the analysis are exactly those who were randomized into the trial. Acceptable handling of subject attrition includes: very little missing outcome data (less than 10% in each group (Genaidy *et al.* 2007)); reasons for missing subjects unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups,

OR analyses (such as intention-to-treat analysis) in which missing data have been imputed using appropriate methods(insuring that the characteristics of subjects lost to follow up or with unavailable records are described in identical way and are not significantly different from those of the study participants).

Note: Participants randomized but subsequently found not to be eligible need not always be considered as having missing outcome data (Higgins and Green 2011).

EA: There is direct evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study. Acceptable handling of attrition includes: very little missing outcome data; reasons for missing animals unlikely to be related to outcome (or for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups; missing outcomes is not enough to impact the effect estimate,

OR missing data have been imputed using appropriate methods (insuring that characteristics of animals are not significantly different from animals retained in the analysis).

Co: There is direct evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study. Acceptable handling of subject attrition includes: very little missing outcome data; reasons for missing subjects unlikely to be related to outcome (for survival data, censoring unlikely to be introducing bias); missing outcome data balanced in numbers across study groups, with similar reasons for missing data across groups,

OR missing data have been imputed using appropriate methods and characteristics of subjects lost to follow up or with unavailable records are described in identical way and are not significantly different from those of the study participants.

CaCo, CrSe: There is direct evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.

Probably Low risk of bias:

HCT: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study,

OR it is deemed that the proportion lost to follow-up would not appreciably bias results (less than 20% in each group (Genaidy *et al.* 2007)). This would include reports of no statistical differences in characteristics of subjects lost to follow up or with unavailable records from those of the study participants. Generally, the higher the ratio of participants with missing data to participants with events, the greater potential there is for bias. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable.

EA: There is indirect evidence that loss of animals was adequately addressed and reasons were documented when animals were removed from a study,

OR it is deemed that the proportion lost would not appreciably bias results. This would include reports of no statistical differences in characteristics of animals removed from the study from those remaining in the study.

Co: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was adequately addressed and reasons were documented when human subjects were removed from a study,

OR it is deemed that the proportion lost to follow-up would not appreciably bias results. This would include reports of no statistical differences in characteristics of subjects lost to follow up or with unavailable records from those of the study participants. Generally, the higher the ratio of participants with missing data to participants with events, the greater potential there is for bias. For studies with a long duration of follow-up, some withdrawals for such reasons are inevitable.

CaCo, CrSe: There is indirect evidence that exclusion of subjects from analyses was adequately addressed, and reasons were documented when subjects were removed from the study or excluded from analyses.

Probably High risk of bias:

HCT: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large (greater than 20% in each group (Genaidy *et al.* 2007)) and not adequately addressed,

OR there is insufficient information provided about numbers of subjects lost to follow-up (record “NR” as basis for answer).

EA: There is indirect evidence that loss of animals was unacceptably large and not adequately addressed,

OR there is insufficient information provided about loss of animals (record “NR” as basis for answer).

Co: There is indirect evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large and not adequately addressed,

OR there is insufficient information provided about numbers of subjects lost to follow-up (record “NR” as basis for answer).

CaCo, CrSe: There is indirect evidence that exclusion of subjects from analyses was not adequately addressed,

OR there is insufficient information provided about why subjects were removed from the study or excluded from analyses (record “NR” as basis for answer).

Definitely High risk of bias:

HCT, Co: There is direct evidence that loss of subjects (i.e., incomplete outcome data) was unacceptably large and not adequately addressed. Unacceptable handling of subject attrition includes: reason for missing outcome data likely to be related to true outcome, with either imbalance in numbers or reasons for missing data across study groups; or potentially inappropriate application of imputation.

EA: There is direct evidence that loss of animals was unacceptably large and not adequately addressed. Unacceptable handling of attrition or exclusion includes: reason for loss is likely to be related to true outcome, with either imbalance in numbers or reasons for loss across study groups.

CaCo, CrSe: There is direct evidence that exclusion of subjects from analyses was not adequately addressed. Unacceptable handling of subject exclusion from analyses includes: reason for exclusion likely to be related to true outcome, with either imbalance in numbers or reasons for exclusion across study groups.

Detection Bias

Detection bias refers to systematic differences between experimental and control groups with regards to how outcomes and exposures are assessed (Higgins and Green 2011) and also considers validity and reliability of methods used to assess outcomes and exposures (Viswanathan *et al.* 2012).

8. Can we be confident in the exposure characterization?

Confidence in the exposure requires valid, reliable, and sensitive methods to measure exposure applied consistently across groups. Exposure misclassification or measurement error may be independent of the outcomes (non-differential) or related to the outcome of interest (differential). Non-differential measurement error of exposures will usually bias the results toward the null by lowering precision and therefore reducing the ability to distinguish potential effects between exposure levels. Therefore, this tool considers the accuracy of the exposure characterization, including both purity and stability for controlled exposure studies, as part of the risk-of-bias rating for exposure. Differential measurement error of exposures can bias the exposure-outcome relationship and result in detection bias.

Detection bias can be minimized by using valid and reliable exposure measures applied consistently across groups (i.e., under the same method and time-frame). Studies that directly measure exposure in subjects (e.g., measurement of the chemical in blood, plasma, urine, etc.) are likely to have less measurement error and less risk of bias for exposure than studies relying on indirect measures (e.g., predictions from activity patterns and microenvironment concentrations). Exposure information obtained by self-report depends on the recall of participants and differential errors in recall can attenuate, strengthen, or even invert the true relationship (White 2003). Self-reporting of exposures for case-control studies are frequently cited as leading to differential measurement errors because cases often remember past exposures better than controls (i.e., recall bias) (e.g., see Rothman *et al.* 2012). Differential measurement error could also be introduced if the exposure data for different groups come from different sources for observational studies or are taken at different time points for experimental studies.

Acceptable methods for measuring exposure will be highly exposure dependent and therefore a specific list of acceptable, inaccurate, or potentially biased methods should be developed for each evaluation and will require subject-matter expertise. It is recommended that experts with some knowledge of the literature (including exposure and outcomes) participate in drafting the risk-of-bias criteria for exposure characterization when a review protocol is developed. Even with early expert consultation and planning, exposure questions may arise when the actual studies are assessed. Additional consultation and modifications to the exposure risk-of-bias criteria may be necessary. When changes are made, they should be documented along with the date on which modifications were made and the logic for the changes.

For controlled exposure studies (i.e., experimental human or animal studies), the use of reliable methods to measure exposure depends primarily on ensuring the purity and stability of the treatment compound. Independent verification of purity would be considered best practice because the identity and purity as listed on the bottle can be inaccurate. In NTP's experience, about 3% of chemicals purchased are the wrong chemical and the inaccuracy rate of chemical labelling rises to 10% if you include inaccurate reporting of purity (unpublished, personal communication Brad Collins, NTP chemist). It is also possible that impurities may be more toxic than the compound of interest. This occurred during an NTP study of PCB 118 where analysis revealed the presence of 0.622% of the much more potent PCB 126, resulting in the study being continued as a mixture study [(NTP 2006), see page 13]. The directions below takes a conservative approach in requiring independent verification of ≥99% purity for a single substance for "definitely low" risk of bias. However, the risk of bias associated with exposure to impurities depends on the identity of the impurities and the sensitivity of the outcome of interest which could result in potential effects of those impurities on the outcome of interest. The threshold for these values should be developed for specific research questions and reflect empirical data for the substance and outcome under consideration when possible. Therefore, for some chemicals like PCBs, ≥99% purity may not be sufficient for "definitely low" risk of bias and for others the appropriate purity value may be lower.

Exposure characterization should also include verification of the compound over the course of the test period. This is particularly important if the compound is volatile or instable. For example, daily preparation of treatment solutions may be required for unstable compounds (e.g., half-lives on the order of days). Special apparatus such as flow-through systems are needed to ensure exposure to volatile compounds. For example, Durda and Preziosi (2000) suggest the use of flow-through systems in aquatic exposures to volatile compounds (e.g., those with Henry's Law values in the range of 10^{-5} atm-m^3/mol or greater).

Human introductory text: Assessment of exposure is a widely used element of tools to assess study quality for observational human studies (Downs and Black 1998, Shamliyan *et al.* 2010, Viswanathan *et al.* 2012, CLARITY Group at McMaster University 2013, Wells *et al.* 2014). Exposure is much more difficult to measure and to accurately ensure for observational studies than for controlled exposure studies. Therefore, exposure measurement error and misclassification are more likely to contribute to risk of bias for observational studies.

The direction of the bias (towards or away from the null) will differ based on the nature of differences between comparison groups and may be difficult to predict. Non-differential misclassification of exposure will generally bias results towards the null, but differential misclassification can bias towards or away from the null, making it difficult to predict the direction of effect (Szklo and Nieto 2007). For controlled exposure studies, noncompliance with the allocated treatment could introduce differential misclassification if compliance was unequal across study groups. Adherence to a strict study protocol that includes measures to assure or assess compliance can reduce the risk of bias.

Animal introductory text: For laboratory or experimental animal studies, exposure assessment has only been included in a few (e.g., Durda and Preziosi 2000) study quality or risk-of-bias tools (reviewed in Krauth *et al.* 2013). However, as described above, this tool considers the accuracy of the exposure characterization as part of the risk-of-bias rating because non-differential exposure misclassification tends to bias the results toward the null. Wildlife or environmental-exposure animal studies are analogous to human observational studies and therefore inclusion of this element would be expected based on guidance for human studies.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, EA: There is direct evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was independently characterized and purity confirmed generally as ≥99%[3] for single substance or non-mixture evaluations (see NTP 2006 for example of study effects attributable to impurities of approximately 1%),

AND that exposure was consistently administered (i.e., with the same method and time-frame) across treatment groups.

Co, CaCo, CrSe, CaS: There is direct evidence that exposure was consistently assessed (i.e., under the same method and time-frame) using well-established methods that directly measure exposure (e.g., measurement of the chemical in air or measurement of the chemical in blood, plasma, urine, etc.),

OR exposure was assessed using less-established methods that directly measure exposure and are validated against well-established methods.

Probably Low risk of bias:

HCT, EA: There is indirect evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was independently characterized and purity confirmed generally as ≥99%[3] (i.e., the supplier of the chemical provides documentation of the purity of the chemical),

OR direct evidence that purity was independently confirmed as ≥98%[3] it is deemed that impurities of up to 2% would not appreciably bias results,

AND there is indirect evidence that exposure was consistently administered (i.e., with the same method and time-frame) across treatment groups.

Co, CaCo, CrSe, CaS: There is indirect evidence that the exposure was consistently assessed using well-established methods that directly measure exposure,

OR exposure was assessed using indirect measures (e.g., questionnaire or occupational exposure assessment by a certified industrial hygienist) that have been validated or empirically shown to be consistent with methods that directly measure exposure (i.e., inter-methods validation: one method vs. another).

Probably High risk of bias:

HCT, EA: There is indirect evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was assessed using poorly validated methods,

OR there is insufficient information provided about the validity of the exposure assessment method, but no evidence for concern (record "NR" as basis for answer).

[3] Note purity thresholds should be developed for specific research questions and reflect empirical data for the substance and outcome under consideration when possible. Therefore, the appropriate cut-off purity value may be lower or higher than the values listed below for ≥99% defining the difference between "definitely low" and "probably low" or ≥98% defining the difference between "probably low" and "probably high" risk of bias.

Co, CaCo, CrSe, CaS: There is indirect evidence that the exposure was assessed using poorly validated methods that directly measure exposure,

OR there is direct evidence that the exposure was assessed using indirect measures that have not been validated or empirically shown to be consistent with methods that directly measure exposure (e.g., a job-exposure matrix or self-report without validation) (record "NR" as basis for answer),

OR there is insufficient information provided about the exposure assessment, including validity and reliability, but no evidence for concern about the method used (record "NR" as basis for answer).

Definitely High risk of bias:

HCT, EA: There is direct evidence that the exposure (including purity and stability of the test substance and compliance with the treatment, if applicable) was assessed using poorly validated methods.

Co, CaCo, CrSe, CaS: There is direct evidence that the exposure was assessed using methods with poor validity,

OR evidence of exposure misclassification (e.g., differential recall of self-reported exposure).

9. Can we be confident in the outcome assessment?

Confidence in the outcome requires valid, reliable, and sensitive methods to assess the outcome applied consistently across groups. Outcome misclassification or measurement error may be unrelated to the exposure (non-differential) or related to the exposure (differential). Non-differential measurement error of outcomes will usually bias the results toward the null by lowering precision and therefore reducing the ability to distinguish potential effects on exposure between exposure levels. Differential measurement error of outcomes can bias the exposure-outcome relationship and result in detection bias. There are three important factors for assessing bias in the outcome assessment: 1) the objectivity of the outcome assessment, 2) consistency in measurement of outcomes, and 3) blinding of the outcome assessors (for knowledge of the exposure).

Detection bias can be minimized by using valid and reliable methods to assess the outcome applied consistently across groups (i.e., under the same method and time-frame). Objectivity of the outcome assessment and the need for blinding are two sides of the same issue. Blinding requires that outcome assessors do not know the study group or exposure level of the human subject or animal when the outcome was assessed. The objectivity of procedures used for measuring and reporting an outcome will impact the degree to which outcome assessors could bias the reported results. For example, a behavioral outcome rated by a researcher (i.e., direct observation of behaviors) relies on subjective judgment and therefore may be impacted by potential bias of the outcome assessor to a greater degree than outcomes that are measured by machines (e.g., automated red blood cell counts). Similarly, studies relying on self-report of outcome may be rated as having a higher risk of bias than studies with clinically observed outcomes (Viswanathan *et al.* 2012). Although objective measures are less prone to bias by researchers than subjective measures, bias could be introduced during sample preparation or handling and therefore blinding still has a role in controlling for potential bias unless sample preparation and outcome measurement are accomplished with automated procedures. For example, the potential for outcome assessors to introduce bias would be minimized for *ex vivo* studies where samples are collected and outcomes are assessed automatically within an apparatus.

Acceptable methods for measuring the outcomes of interest will be highly dependent on the outcome and therefore a specific list of acceptable, inaccurate, or potentially biased methods should be developed for each evaluation and will require subject-matter expertise. It is recommended that experts with some knowledge of the literature (including both exposure and outcome) participate in drafting the risk-of-bias criteria for outcome assessment when a review protocol is developed. Even with early expert consultation and planning, outcome questions may arise when the actual studies are assessed because of non-traditional methods, application to non-traditional species, or endpoints that are indirectly related to the outcome of interest. Additional consultation and modifications to the outcome risk-of-bias criteria may be necessary. When changes are made, they should be documented along with the date on which modifications were made and the logic for the changes.

Human introductory text: Differential methods used in the assessment of outcomes is a source of bias and this is a widely used risk-of-bias element in tools for observational human studies (Downs and Black 1998, Genaidy *et al.* 2007, Shamliyan *et al.* 2010, Viswanathan *et al.* 2012, Sterne *et al.* 2014). The recent guidance for non-randomized studies of interventions suggests considering the objectivity of the outcome assessment when evaluating bias in the outcome assessment (Sterne *et al.* 2014) and we have included consideration of the objectivity in this document for evaluating the potential impact of blinding practices. Blinding of outcome assessors is a widely recommended risk-of-bias element for controlled trials and observational studies (Higgins and Green 2011, Viswanathan *et al.* 2012, Sterne *et al.* 2014). For human studies blinding of the subject to exposure levels should also be considered. For example, a subject's knowledge of their own exposure levels would represent an increased risk of bias for self-reported outcomes relative to clinically measured outcomes.

Without distinguishing between the different stages of blinding during the conduct of a study, lack of blinding in randomized trials has been empirically shown to be associated with larger estimations of intervention effects (on average a 9% increase in an odds ratio) (Pildal *et al.* 2007). Schulz et al. (1995) analyzed 250 controlled trials and found that studies that were not double-blinded had a 17% larger estimation of treatment effect, on average. In trials with more subjective outcomes, more bias has been observed with lack of blinding (Wood *et al.* 2008), indicating that blinding outcome assessors could be more important for these effects.

For some exposures, it is not possible to entirely blind outcome assessors, particularly if subjects are self-reporting outcomes. In practice, successful blinding cannot always be ensured, as it can be compromised for most interventions. In some cases the treatment may have side effects possibly allowing the participant to detect which intervention they received, unless the study compares interventions with similar side effects or uses an active placebo (Boutron *et al.* 2006).

Animal introductory text: Blinding of outcome assessors is a widely recommended risk-of-bias element for animal studies (reviewed in Krauth *et al.* 2013). This tool assesses blinding and also considers differential methods, procedures, or time points for measuring outcomes to be a source of bias based on the common use of this element in study quality tools for human controlled trials and observational studies (Higgins and Green 2011, Viswanathan *et al.* 2012, Sterne *et al.* 2014).

There is empirical evidence that lack of blinding at outcome assessment in animal studies is associated with larger measures of the effect (Bebarta *et al.* 2003, Sena *et al.* 2007, Vesterinen *et al.* 2010). In animal studies, blinding of study group at outcome assessment may not be possible because of the need to determine if treated animals are affected relative to controls in a treatment or dose-dependent manner (examples include clinical observations and histopathological assessment of non-neoplastic lesions). However, animal studies are in general more tightly controlled than human studies and additional measures may be taken to reduce the risk of bias.

Note: for case-control studies, confirmation that the control subjects are free of the outcome is considered under as separate risk-of-bias question, "Did selection of study participants result in appropriate comparison groups?"

There are separate risk-of-bias issues affected by concealment or blinding that are important for selection, performance and detection bias: 1) a question under selection bias addresses allocation concealment of research personnel and human subjects; and 2) a question under performance bias addresses blinding of research personnel and human subjects to the study group during the study.

Previous versions of the OHAT risk-of-bias tool had separate questions for "Can we be confident in the outcome assessment" and "Were the outcome assessors blinded to study group or exposure level?" These two questions are interrelated and therefore have been combined as factors to consider for a single risk-of-bias rating on confidence in the outcome assessment. Recent guidance, such as the Cochrane risk-of-bias tool for non-randomized studies of intervention (Sterne et al. 2014) has taken a similar approach for addressing blinding and outcome measurement in a single question on outcome assessment.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, Co: There is direct evidence that the outcome was assessed using well-established methods (e.g., the "gold standard" with validity and reliability >0.70 Genaidy *et al.* 2007),

AND subjects had been followed for the same length of time in all study groups. Acceptable assessment methods will depend on the outcome, but examples of such methods may include: objectively measured with diagnostic methods, measured by trained interviewers, obtained from registries (Shamliyan *et al.* 2010),

AND there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes.

EA: There is direct evidence that the outcome was assessed using well-established methods (the gold standard),

AND assessed at the same length of time after initial exposure in all study groups,

AND there is direct evidence that the outcome assessors were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes.

CaCo: There is direct evidence that the outcome was assessed in cases (i.e., case definition) and controls using well-established methods (the gold standard),

AND subjects had been followed for the same length of time in all study groups,

AND there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level when outcome was assessed in cases (i.e., case definition) and controls.

CrSe, CaS: There is direct evidence that the outcome was assessed using well-established methods (the gold standard),

AND there is direct evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes.

Probably Low risk of bias:

HCT, Co: There is indirect evidence that the outcome was assessed using acceptable methods (i.e., deemed valid and reliable but not the gold standard) (e.g., validity and reliability ≥0.40 Genaidy *et al.* 2007),

AND subjects had been followed for the same length of time in all study groups [Acceptable, but not ideal assessment methods will depend on the outcome, but examples of such methods may include proxy reporting of outcomes and mining of data collected for other purposes],
OR it is deemed that the outcome assessment methods used would not appreciably bias results,
AND there is indirect evidence that the outcome assessors (including study subjects, if outcomes were self-reported) were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes,
OR it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results, which is more likely to apply to objective outcome measures.

EA: There is indirect evidence that the outcome was assessed using acceptable methods (i.e., deemed valid and reliable but not the gold standard),
AND assessed at the same length of time after initial exposure in all study groups,
OR it is deemed that the outcome assessment methods used would not appreciably bias results,
AND there is indirect evidence that the outcome assessors were adequately blinded to the study group, and it is unlikely that they could have broken the blinding prior to reporting outcomes,
OR it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results, which is more likely to apply to objective outcome measures. For some outcomes, particularly histopathology assessment, outcome assessors are not blind to study group as they require comparison to the control to appropriately judge the outcome, but additional measures such as multiple levels of independent review by trained pathologists can minimize this potential bias.

CaCo: There is indirect evidence that the outcome was assessed in cases (i.e., case definition) and controls using acceptable methods,
AND subjects had been followed for the same length of time in all study groups,
OR it is deemed that the outcome assessment methods used would not appreciably bias results,
AND there is direct evidence that the outcome assessors were adequately blinded to the exposure level when reporting outcomes,
OR it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results (including that subjects self-reporting outcomes were likely not aware of reported links between the exposure and outcome or lack of blinding is unlikely to bias a particular outcome).

CrSe, CaS: There is indirect evidence that the outcome was assessed using acceptable methods,
OR it is deemed that the outcome assessment methods used would not appreciably bias results,
AND there is indirect evidence that the outcome assessors were adequately blinded to the exposure level, and it is unlikely that they could have broken the blinding prior to reporting outcomes,
OR it is deemed that lack of adequate blinding of outcome assessors would not appreciably bias results (including that subjects self-reporting outcomes were likely not aware of reported links between the exposure and outcome lack of blinding is unlikely to bias a particular outcome).

Probably High risk of bias:

HCT, Co: There is indirect evidence that the outcome assessment method is an insensitive instrument (e.g., a questionnaire used to assess outcomes with no information on validation),

OR the length of follow up differed by study group,
OR there is indirect evidence that it was possible for outcome assessors (including study subjects if outcomes were self-reported) to infer the study group prior to reporting outcomes,
OR there is insufficient information provided about blinding of outcome assessors (record “NR” as basis for answer).

EA: There is indirect evidence that the outcome assessment method is an insensitive instrument,

OR the length of time after initial exposure differed by study group,

OR there is indirect evidence that it was possible for outcome assessors to infer the study group prior to reporting outcomes without sufficient quality control measures,

OR there is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).

CaCo: There is indirect evidence that the outcome was assessed in cases (i.e., case definition) using an insensitive instrument,

OR there is insufficient information provided about how cases were identified (record "NR" as basis for answer),

OR there is indirect evidence that it was possible for outcome assessors to infer the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were likely aware of reported links between the exposure and outcome),

OR there is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).

CrSe, CaS: There is indirect evidence that the outcome assessment method is an insensitive instrument,

OR there is indirect evidence that it was possible for outcome assessors to infer the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were likely aware of reported links between the exposure and outcome),

OR there is insufficient information provided about blinding of outcome assessors (record "NR" as basis for answer).

Definitely High risk of bias:

HCT, Co: There is direct evidence that the outcome assessment method is an insensitive instrument,

OR the length of follow up differed by study group,

OR there is direct evidence for lack of adequate blinding of outcome assessors (including study subjects if outcomes were self-reported), including no blinding or incomplete blinding.

EA: There is direct evidence that the outcome assessment method is an insensitive instrument,

OR the length of time after initial exposure differed by study group,

OR there is direct evidence for lack of adequate blinding of outcome assessors, including no blinding or incomplete blinding without quality control measures.

CaCo: There is direct evidence that the outcome was assessed in cases (i.e., case definition) using an insensitive instrument,

OR there is direct evidence that outcome assessors were aware of the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were aware of reported links between the exposure and outcome).

CrSe, CaS: There is direct evidence that the outcome assessment method is an insensitive instrument,

OR there is direct evidence that outcome assessors were aware of the exposure level prior to reporting outcomes (including that subjects self-reporting outcomes were aware of reported links between the exposure and outcome).

Selective Reporting Bias

Selective reporting bias refers to selective inclusion of outcomes in the publication of the study on the basis of the results (Hutton and Williamson 2000, Higgins and Green 2011).

10. Were all measured outcomes reported?

Selective reporting of results is a recommended element of assessing risk of bias (Guyatt *et al.* 2011, Higgins *et al.* 2011, IOM 2011, Viswanathan *et al.* 2012). Selective reporting is present if pre-specified outcomes are not reported or incompletely reported. It is likely widespread and difficult to assess with confidence for most studies unless the study protocol is available. Selective reporting bias can be assessed by comparing the “methods” and “results” section of the paper, and by considering outcomes measured in the context of knowledge in the field. Abstracts of presentations relating to the study may contain information about outcomes not subsequently mentioned in publications. Selective reporting bias should be suspected if the study does not report outcomes in the results section that would have been expected based on the methods, or if a composite score is present without the individual component outcomes (Guyatt *et al.* 2011). It may be useful to pay attention to author affiliations and funding source which can contribute to selective outcome reporting when results are not consistent with expectations or value to the research objectives.

Applies to: HCT, EA, Co, CaCo, CrSe, CaS

Definitely Low risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is direct evidence that all of the study’s measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported. This would include outcomes reported with sufficient detail to be included in meta-analysis or fully tabulated during data extraction and analyses had been planned in advance.

Probably Low risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is indirect evidence that all of the study’s measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported,

OR analyses that had not been planned in advance (i.e., retrospective unplanned subgroup analyses) are clearly indicated as such and it is deemed that the unplanned analyses were appropriate and selective reporting would not appreciably bias results (e.g., appropriate analyses of an unexpected effect). This would include outcomes reported with insufficient detail such as only reporting that results were statistically significant (or not).

Probably High risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is indirect evidence that all of the study’s measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have been reported,

OR and there is indirect evidence that unplanned analyses were included that may appreciably bias results,

OR there is insufficient information provided about selective outcome reporting (record "NR" as basis for answer).

Definitely High risk of bias:

HCT, EA, Co, CaCo, CrSe, CaS: There is direct evidence that all of the study's measured outcomes (primary and secondary) outlined in the protocol, methods, abstract, and/or introduction (that are relevant for the evaluation) have not been reported. In addition to not reporting outcomes, this would include reporting outcomes based on composite score without individual outcome components or outcomes reported using measurements, analysis methods or subsets of the data (e.g., subscales) that were not pre-specified or reporting outcomes not pre-specified, or that unplanned analyses were included that would appreciably bias results.

Other Bias

11. Were there no other potential threats to internal validity (e.g., statistical methods were appropriate and researchers adhered to the study protocol)?

On a project specific basis, additional questions for other potential threats to internal validity can be added and applied to study designs as appropriate.

Were statistical methods appropriate?

Some of the more extensive quality tools have a separate question for appropriateness of the statistical methods (e.g., 1 of the 25 elements in the Downs and Black 1998 tool addresses the statistics); however most do not include a separate question. The OHAT risk-of-bias tool suggests consideration of statistical methods with the other potential threats to internal validity. One of the common statistical issues identified has been reporting of statistical tests that require normally distributed data (e.g., t-test or ANOVA) without reporting that the homogeneity of variance was tested or confirmed.

It is recommended that experts with some knowledge of statistical methods used in the literature participate in drafting the risk-of-bias criteria for identifying inappropriate statistical methods when a review protocol is developed. Even with early expert consultation and planning, statistical methods questions may arise when the actual studies are assessed. Additional consultation and modifications to the statistical methods risk-of-bias criteria may be necessary. When changes are made, they should be documented along with the date on which modifications were made and the logic for the changes.

Did researchers adhere to the study protocol?

Failure of the study to maintain fidelity to the protocol is recommended as an important consideration when assessing performance bias (IOM 2011, Viswanathan *et al.* 2012). However, it will likely be difficult to assess with confidence for most studies, particularly when the methods section of a publication is all that is available. In some instances the protocol is meant to be "fluid" and the protocol explicitly allows for modification based on need; such fluidity does not mean the interventions are implemented incorrectly. The deviation may not result in a risk of bias, or if it does the direction of the bias (towards or away from the null) will differ based on the deviation from the protocol.

We recognize that given reporting practices it is unlikely that deviations from the protocol will be explicitly reported in most studies. Thus, we will assume unless stated otherwise that no deviations occurred which will result in most studies considered “probably low risk of bias”. In the short-term, this risk-of-bias element is unlikely to be informative for the purposes of discriminating between studies of higher quality and studies of lower quality. However, in the long-term, especially if reporting standards improve, collecting this information may generate data that will allow us to empirically assess evidence of bias or to remove this risk-of-bias question from consideration.

Animal introductory text: One of the more common deviations from protocol that can occur in toxicity studies is when a dose level is decreased based on evidence of mortality or severe toxicity. However, depending upon how the author addresses this change it may or may not impact results. For example, when this occurs in NTP studies, the usual analysis would be conducted on the dose groups remaining after the toxic dose level is dropped. A similar situation arises when a dose group has to be euthanized due to overt toxicity.

Previous versions of the OHAT risk-of-bias tool had a separate question for “Did researchers adhere to the study protocol?” The overwhelming majority of studies examined during case study evaluations were not reported in sufficient detail to permit a meaningful answer to whether or not the study adhered to a study protocol. Therefore, we will continue to collect data on this element, but have moved it to be considered under other potential threats to internal validity.

Did the study design or analysis account for important confounding and modifying variables (including unintended co-exposures) in experimental studies?

There is a separate risk-of-bias question to address confounding and modifying variables (including co-exposures) for observational studies because confounding is a much greater concern for observational studies. Controlled exposure studies (i.e., experimental human or animal studies) can address confounding through study design features such as randomization and allocation concealment. Therefore, most study quality tools for experimental studies do not include questions for confounding (Higgins *et al.* 2011, Krauth *et al.* 2013, Koustas *et al.* 2014). Confounding by chance (i.e., confounding that is unknown, unmeasured, or poorly measured) is expected to be equally distributed between groups under true randomization; however, experimental studies may not always successfully randomize potential confounders (Viswanathan *et al.* 2013). Recognizing this, the SYRCLE risk-of-bias tool for experimental animal studies asks whether groups were “similar at baseline or were they adjusted for confounders in the analysis”(Hooijmans *et al.* 2014). In the context of an animal study, this element would include consideration of covariates such as body weight, litter size, or other outcome-specific covariates. Similarly, the 2012 risk-of-bias guidance from AHRQ recommends consideration of confounding for randomized clinical trials. For this tool, we have only included the consideration of confounding in controlled exposure studies (i.e., experimental human or animal studies) under “other potential threats to internal validity” for cases where it is strongly suspected because randomization and allocation concealment should address the issue of confounding. The issue of confounding overlaps with randomization and allocation concealment, and multiple questions would address the same issue.

Animal introductory text: Randomization and allocation concealment in experimental studies address should result in non-differential distribution of potential confounders or co-exposures. Experimental study design generally reduces or eliminates co-exposures; however the impact of inadvertent chemical or biological co-exposures should be considered. For example, if the experimental exposure is to bisphenol A or other chemical with estrogenic properties, husbandry practices that raise the background level of estrogenicity across all study groups (e.g., a diet high in phytoestrogens) may make the model system less sensitive to detect low-dose effects of BPA (Thigpen *et al.* 2007, Muhlhauser *et al.* 2009). In this case, the direction of the bias would be towards the null (towards smaller effect sizes). Infectious agents and non-treatment related co-morbidity should also be monitored as potential sources of bias.

The direction of the bias will depend on the nature of co-exposure and whether or not there are differences between study groups. For example, certain types of infections may be related to outcomes of interest (reviewed in NRC 1991, Baker 1998, GV-SOLAS 1999). *Helicobacter pylori* is a bacterial carcinogen and may cause may cause chronic active hepatitis, hepatic tumors, and proliferative typhlocolitis in rodents (Kusters *et al.* 2006). If the infection occurs in control animals or across all study groups, then the bias for an effect on the liver may be towards the null (smaller effect size). If the infection occurs only in treated animals, then the bias for an effect on the liver may be away from the null (larger effect size).

Examples:

- **Statistics:** Failure to statistically or experimentally adjust for litter in an animal study with a developmental outcome. The direction of the bias is away from the null towards a larger effect size (Haseman *et al.* 2001).
- **Deviations from the protocol:** Evidence of deviations in the protocol should be noted as direct (definitely high risk of bias) or indirect (probably high risk of bias). Given reporting practices it is unlikely that deviations from the protocol will be explicitly reported in most studies and therefore the bias is very difficult to assess. Caution should be taken so that studies that do provide a protocol and report deviations are not ”punished” for having better reporting practices.
- **Unintended co-exposures for experimental studies**: Evidence of other exposures that are anticipated to bias results should be noted as direct (definitely high risk of bias) or indirect (probably high risk of bias) evidence of other exposures anticipated to bias results, if present and not appropriately adjusted for. Non-differential co-exposures that are likely to bias the results toward the null should be considered in the context of the study findings.

REFERENCES

Baker DG. 1998. Natural pathogens of laboratory mice, rats, and rabbits and their effects on research. *Clin Microbiol Rev* 11(2): 231-266.

Bebarta V, Luyten D, Heard K. 2003. Emergency medicine animal research: does use of randomization and blinding affect the results? *Acad Emerg Med* 10(6): 684-687.

Boutron I, Estellat C, Guittet L, Dechartres A, Sackett DL, Hrobjartsson A, Ravaud P. 2006. Methods of blinding in reports of randomized controlled trials assessing pharmacologic treatments: a systematic review. *PLoS Med* 3(10): e425.

CLARITY Group at McMaster University. 2013. *Tools to assess risk of bias in cohort studies, case control studies, randomized controlled trials, and longitudinal symptom research studies aimed at the general population*. Available: http://www.evidencepartners.com/resources/ [accessed 15 January 2013].

Downs SH, Black N. 1998. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 52(6): 377-384.

Duke JL, Zammit TG, Lawson DM. 2001. The effects of routine cage-changing on cardiovascular and behavioral parameters in male Sprague-Dawley rats. *Contemp Top Lab Anim Sci* 40(1): 17-20.

Durda J, Preziosi D. 2000. Data quality evaluation of toxicological studies used to derive ecotoxicological benchmarks. *Human and Ecological Risk Assessment: An International Journal* 6(5): 747-765.

Dwan K, Gamble C, Kolamunnage-Dona R, Mohammed S, Powell C, Williamson PR. 2010. Assessing the potential for outcome reporting bias in a review: a tutorial. *Trials* 11: 52.

Genaidy AM, Lemasters GK, Lockey J, Succop P, Deddens J, Sobeih T, Dunning K. 2007. An epidemiological appraisal instrument - a tool for evaluation of epidemiological studies. *Ergonomics* 50(6): 920-960.

Gerdin AK, Igosheva N, Roberson LA, Ismail O, Karp N, Sanderson M, Cambridge E, Shannon C, Sunter D, Ramirez-Solis R, Bussell J, White JK. 2012. Experimental and husbandry procedures as potential modifiers of the results of phenotyping tests. *Physiol Behav* 106(5): 602-611.

Gerstman BB. 2013. *Epidemiology kept simple* 3rd ed., New York, NY: Wiley-Blackwell.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW, Jr., Atkins D, Meerpohl J, Schunemann HJ. 2011. GRADE guidelines: 4. Rating the quality of evidence-study limitations (risk of bias). *J Clin Epidemiol* 64(4): 407-415.

GV-SOLAS. 1999. Implications of infectious agents on results of animal experiments. Report of the Working Group on Hygiene of the Gesellschaft fur Versuchstierkunde--Society for Laboratory Animal Science (GV-SOLAS). *Lab Anim* 33 Suppl 1: S39-87.

Haseman JK, Bailer AJ, Kodell RL, Morris R, Portier K. 2001. Statistical issues in the analysis of low-dose endocrine disruptor data. *Toxicol Sci* 61(2): 201-210.

Higgins J, Green S. 2011. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration. Available: www.cochrane-handbook.org [accessed 3 February 2013].

Higgins JP, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Br Med J* 343: d5928.

Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. 2014. SYRCLE's risk of bias tool for animal studies. *BMC medical research methodology* 14(1): 43.

Hutton JL, Williamson PR. 2000. Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society Series C-Applied Statistics* 49: 359-370.

IOM (Institute of Medicine). 2011. *Finding what works in health care: Standards for systematic reviews*. Washington, DC, The National Academies Press: 318. Available: http://www.nap.edu/openbook.php?record_id=13059 [accessed 3 May 2013].

Johnson P, Sutton P, Atchley D, Koustas E, Lam J, Robinson K, Sen S, Axelrad D, Woodruff T. 2013. *Applying the Navigation Guide: Case study #1. The impact of developmental exposure to perfluorooctanoic acid (PFOA) on fetal growth. A systematic review of the human evidence - Protocol*. Available: http://prhe.ucsf.edu/prhe/pdfs/PFOA%20Human%20Protocol.pdf.

Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014. The Navigation Guide - evidence-based medicine meets environmental health: systematic review of human evidence for PFOA effects on fetal growth. *Environ Health Perspect* 122(10): 1028-1039.

Koustas E, Lam J, Sutton P, Johnson P, Atchley D, Robinson K, Sen S, Axelrad D, Woodruff T. 2013. *Applying the Navigation Guide: Case Study #1. The Impact of Developmental Exposure to Perfluorooctanoic Acid (PFOA) On Fetal Growth. A Systematic Review of the Non-Human Evidence - Protocol*. Available: http://prhe.ucsf.edu/prhe/pdfs/PFOA%20NON-HUMAN%20PROTOCOL.pdf.

Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, Robinson KA, Axelrad DA, Woodruff TJ. 2014. The Navigation Guide - evidence-based medicine meets environmental health: systematic review of nonhuman evidence for PFOA effects on fetal growth. *Environ Health Perspect* 122(10): 1015-1027.

Krauth D, Woodrull T, Bero L. 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: a systematic review. *Environ Health Perspect* 121(9): 985-992.

Kusters JG, van Vliet AH, Kuipers EJ. 2006. Pathogenesis of Helicobacter pylori infection. *Clin Microbiol Rev* 19(3): 449-490.

Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. 2008. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke* 39(10): 2824-2829.

Muhlhauser A, Susiarjo M, Rubio C, Griswold J, Gorence G, Hassold T, Hunt PA. 2009. Bisphenol A effects on the growing mouse oocyte are influenced by diet. *Biol Reprod* 80(5): 1066-1071.

NRC (National Research Council,). 1991. *Infectious diseases of mice and rats*. 9780309063326. Washington, DC, Press TNA: 397. Available: http://www.nap.edu/openbook.php?record_id=1429.

NTP (National Toxicology Program,). 2006. *Toxicology and carcinogenesis studies of a binary mixture of 3,3',4,4',5-Pentachlorobiphenyl (PCB 126) (CAS No. 57465-28-8) and 2,3',4,4',5-Pentachlorobiphenyl (PCB 118) (CAS No. 31508-00-6) in female Harlan Sprague-Dawley rats (Gavage Studies)*. Available: http://ntp.niehs.nih.gov/?objectid=D16D6C59-F1F6-975E-7D23D1519B8CD7A5 [accessed 28 January 2013].

NTP (National Toxicology Program). 2013a. *Informational meeting on the draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments. April 23, 2013*. Available: http://ntp.niehs.nih.gov/go/38751 [accessed 28 January 2014].

NTP (National Toxicology Program). 2013b. *Draft Protocol for Systematic Review to Evaluate the Evidence for an Association Between Bisphenol A (BPA) and Obesity*. RTP, NC: Office of Health Assessment and Translation. Available: http://ntp.niehs.nih.gov/go/38673 [accessed 9 April 2013].

NTP (National Toxicology Program). 2013c. *Draft Protocol for Systematic Review to Evaluate the Evidence for an Association Between Perfluorooctanoic Acid (PFOA) or Perfluorooctane Sulfonate (PFOS) Exposure and Immunotoxicity*. RTP, NC: Office of Health Assessment and Translation. Available: http://ntp.niehs.nih.gov/go/38673 [accessed 9 April 2013].

NTP (National Toxicology Program). 2013d. *Draft OHAT Approach for Systematic Review and Evidence Integration for Literature-based Health Assessments – February 2013*. RTP, NC: Office of Health Assessment and Translation. Available: http://ntp.niehs.nih.gov/go/38138 [accessed 10 March 2013].

Pildal J, Hrobjartsson A, Jorgensen KJ, Hilden J, Altman DG, Gotzsche PC. 2007. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int J Epidemiol* 36(4): 847-857.

Rosenthal R, Lawson R. 1964. A Longitudinal Study of the Effects of Experimenter Bias on the Operant Learning of Laboratory Rats. *Journal of psychiatric research* 69: 61-72.

Rothman KJ, Greenland S, Lash TL. 2012. *Modern Epidemiology* 3rd ed., Boston, MA: Lippincott, Williams & Wilkins.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *J Am Med Assoc* 273(5): 408-412.

Schulz KF, Altman DG, Moher D. 2002. Allocation concealment in clinical trials. *J Am Med Assoc* 288(19): 2406-2407; author reply 2408-2409.

Sena E, van der Worp HB, Howells D, Macleod M. 2007. How can we improve the pre-clinical development of drugs for stroke? *TRENDS Neurosci* 30(9): 433-439.

Shamliyan T, Kane RL, Dickinson S. 2010. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *J Clin Epidemiol* 63(10): 1061-1070.

Shamliyan TA, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, Janes G, Maglione M, Moher D, Nasser M, Robinson KA, Segal JB, Tsouros S. 2011. Development of quality criteria to evaluate nontherapeutic studies of incidence, prevalence or risk factors of chronic diseases: Pilot study of new checklists. Available at http://www.ncbi.nlm.nih.gov/books/NBK53272/ [accessed March 6, 2012]. *Rockville (MD): Agency for Healthcare Research and Quality (US); 2011 Jan. Report No.: 11-EHC008-EF. AHRQ Methods for Effective Health Care.*

Sterne J, Higgins J, Reeves B, on behalf of the development group for ACROBAT-NRSI. 2014. *A Cochrane Risk Of Bias Assessment Tool: for Non-Randomized Studies of Interventions (ACROBAT-NRSI), Version 1.0.0*. Available: http://www.riskofbias.info [accessed 28 September 2014].

Szklo M, Nieto FJ. 2007. *Epidemiology: Beyond the Basics* 2nd ed., Sudbury, MA: Jones and Bartlett Publishers.

Thigpen JE, Setchell KD, Padilla-Banks E, Haseman JK, Saunders HE, Caviness GF, Kissling GE, Grant MG, Forsythe DB. 2007. Variations in phytoestrogen content between different mill dates of the same diet produces significant differences in the time of vaginal opening in CD-1 mice and F344 rats but not in CD Sprague-Dawley rats. *Environ Health Perspect* 115(12): 1717-1726.

Vesterinen HM, Sena ES, ffrench-Constant C, Williams A, Chandran S, Macleod MR. 2010. Improving the translational hit of experimental treatments in multiple sclerosis. *Mult Scler* 16(9): 1044-1055.

Viswanathan M, Ansari M, Berkman ND, Chang S, Hartling L, McPheeters LM, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. 2012. *Assessing the risk of bias of individual studies when comparing medical interventions*. Publication No. 12-EHC047-EF. Rockville, MD. Agency for Healthcare Research and Quality (AHRQ). Available:

http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=998 [accessed 3 January 2013].

Viswanathan M, Ansari M, Berkman ND, Chang S, Hartling L, McPheeters LM, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. 2013. *Assessing risk of bias and confounding in observational studies of interventions or exposures: Further development of the RTI item bank*. Publication No. 13-EHC106-EF. Rockville, MD. Agency for Healthcare Research and Quality (AHRQ). Available: http://www.effectivehealthcare.ahrq.gov/ehc/products/414/1612/RTI-item-bank-bias-precision-130805.pdf [accessed 11 January 2014].

Wells G, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. 2014. *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses*. Ottawa Hospital Research Institute. Available: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp [accessed 20 February 2014].

White E. 2003. Design and interpretation of studies of differential exposure measurement error. *Am J Epidemiol* 157(5): 380-387.

Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, Gluud C, Martin RM, Wood AJ, Sterne JA. 2008. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *Br Med J* 336(7644): 601-605.

Woodruff TJ, Sutton P. 2014. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. *Environ Health Perspect* 122(10): 1007-1014.